RESEARCH IN ESTONIA

PRESENT AND FUTURE

TALLINN 2011

RESEARCH IN ESTONIA

PRESENT AND FUTURE

J. Engelbrecht (editor-in-chief)
G. Varlamova

ISBN

In this book the views on cutting edge research in Estonia, especially on its future perspectives are collected, authored by the eminent scientists and scholars of Estonia. The overviews bear the imprint of authors because no strict rules and formats for papers have been introduced.

The book is meant for everyone who is interested in science and scholarship in Estonia, let them be researchers themselves or people dealing with science policy and forward looks, let them reside in Estonia or abroad. We hope that the Estonian representatives in other countries find also this book useful for representing ideas of scientific community in Estonia as a part of the endless enlargement of knowledge in the world.

The idea of the publication belongs to the Estonian Academy of Sciences and its publication is funded by the Estonian Ministry of Education and Research. Sincere thanks to all the authors who contributed to this book.

CONTENTS

III Biology, Geology and Chemistry

IV Humanities and Social Sciences

INTRODUCTION: TOWARDS A KNOWLEDGE-BASED SOCIETY

J. Engelbrecht

Activity is the only road to knowledge.
George Bernard Shaw

Science and scholarship are the drivers for society. Since Estonia is not rich in natural resources, its future depends primarily upon the knowledge of its people. Knowledge and research are inseparably linked and it is clear that in contemporary societies, research matters. Research not only contributes to innovation and to economic development, it is about man, society and the world, about culture and human perception, about inquiry into phenomena, it is a response to societal problems, to natural hazards and to climate change, a way to improving health and education and so on.

In this book the views on cutting edge research in Estonia, especially on its future perspectives are collected, authored by the eminent scientists and scholars of Estonia.

The Introduction starts with general thoughts on how the knowledge in the world is developed. Then, moving on to Estonia, first a short historical overview on research in Estonia is presented in order to explain where we come from. After that the road through the recent structural changes over last 20 years is described for explaining the present situation. This description is followed by a rather general overview on recent activities and on building a knowledge based society. Then the way is paved for the main part of this book – selected chapters in perspective fields of research.

THE WORLD OF KNOWLEDGE

We live in a dynamic world where all the constituents are linked with each other, where changes occur at various time-scales and where nature, social life and man-made systems altogether may be referred to as a complex system. To understand the world with all its changes and predict the future changes is not easy but it is a must for man and society; that is why science is needed. However, "the world is far too rich to be expressed in a single language," argued I. Prigogine (1980). We may also remind the Declaration of the

World Congress of Science in 1999, "...science is a powerful resource for understanding natural and social phenomena, and its role promises to be even greater in the future as the growing complexity between society and the environment is better understood" (Declaration 1999). Indeed, it is not only the need to understand better the matter and life for enlarging our knowledge; it is also the need to face the Great Challenges for mankind on health, poverty, natural hazards, climate influence, technology, etc. And even more, now we understand that the role of social sciences and humanities is growing because society must cope with all of those complex problems.

It is no surprise that complexity studies are developing fast. Based on nonlinear dynamics, self-organisation, hierarchical structures, etc. (see, for example, Érdi 2008; Nicolis, Nicolis 2007) the complexity studies involve physics and chemistry, biosystems and computer science, social systems, economics and financial sphere and so on. Have we only now understood the importance of such an approach? The answer is no – old Greeks knew it and many thinkers over later centuries indicated the need to tackle the problems in a holistic way (see the collection of essays edited by Juarrero and Rubino (2008)). Many of those ideas were ahead of time and only now we understand how right were Immanuel Kant, John Stuart Mill, Charles Sanders Peirce, Henri Poincaré and others describing the complexity of world. However, it took a long time before our approaches in research were changed. The reason was best explained by A.Toffler (1984), "One of the most highly developed skills in contemporary Western civilization is dissection: the split-up of problems into their smallest components. We are good at it. So good, we often forget to put the pieces back together again." Now there is a clear understanding that pieces must be put together.

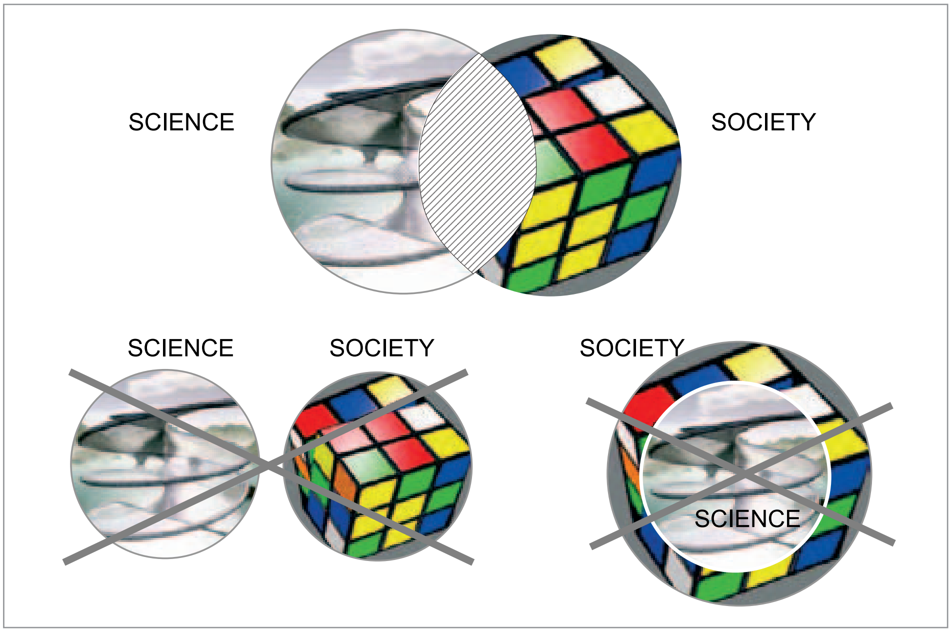
Another important issue in research is that given the scale of problems, the needs of society and limits of funding, strategies of research play essential role. This is first of all for general interest of governments and society because the tax-payers money must be spent transparently. But a strategy may envisage only general fields of science and this is a role of scientists to look forward in their respective fields. The serendipity may certainly help (like the story of penicillin more than half a century ago or recently the story of graphene) but experts could foresee the trends and envisage possible progress. The European Science Foundation has during the last decade published several such forward-looks (www.esf.org) like on nanomedicine and nanoscience, industrial mathematics, systems biology, global change, transcultural identities, etc. These forward-looks are all characterised by deep analysis of future perspectives. Certainly it is impossible to predict discoveries but the fields where research might move the frontiers of knowledge are marked. The Standing

Committees of the ESF have summarised trends in natural and life sciences as well as in humanities and social sciences.

Scientists are aware that too complicated strategies with many deliverables and indicators may not drive actual research. Take for example this famous 3% R&D funding target for the EU which was fixed in Lisbon. Announced with a good faith, the actions did not follow and the average funding in the EU in 2010 is still below 2%. Now in 2011 the main attention in the EU is on innovation. Again, the idea is good but without actions and changes in rules (for example in patents) the goals might not be achieved. These new goals are described in the recent EC Green Paper (2011) on a Common Strategic Framework which emphasises very much innovation leaving aside social sciences and humanities. European Academies (ALLEA) proposed to elaborate also these fields because without social innovation, values, humanities and cultural transformation there will be no success. In addition to that, the most important asset in building the world of knowledge is the education starting from the children (science education) until higher education and PhD studies. Many countries in Europe and worldwide have understood this need and act accordingly.

In organising research there are two main issues. The first is excellence in research which is an inevitable component in every field and branch of research. There is no need to retell the stories of the Nobel awardees or the famous universities over the entire world. In Europe the European Research Council (ERC) has from its start fixed the main requirement for grantees – excellence and applications are reviewed from the viewpoint of ambitions. The results of the ERC grantees are brilliant. But the best minds should also have good conditions for their activities and this makes the second main issue – the research structures must be well organised and the infrastructure in general must be on the level. This need is also understood by policy-makers and there are good examples both internationally and nationally. Many international centres of research like CERN, ESO EMBO, etc. have already a long history and excellent results. The programmes like ESFRI have been launched to build up solid base of infrastructures in new international centres. However, there is another figurative aspect in research structures related to the complexity ideas – namely the self-organisation and fractality. Indeed, instead of a kind of pyramids with best at the top or matrix systems with every element interacting with others, we might think of fractals as irregular networks of elements with nodes and links that interact sometimes hierarchically, sometimes concurrently (see Engelbrecht 2006). This enables the various centres to have interactions of a different nature (interdisciplinary, interregional, etc.) and intensity (some attractive nodes interact more). And as with fractals, it is the repetition of the same rule(s) which makes structures developing, may be more complex but stable. For research systems these rules are really simple: support quality, support young people. Applying these rules steadfastly, the world of knowledge is built.

One important aspect of research beside direct results is related to society, or in other words, to science-society relations. Following J. Lotman (2001), the manifolds of society and science are partly overlapping (see the figure below).



Overlapping (of signs, systems, ...) is trivial, links between non-overlapping areas are important (J. Lotman).

The extreme situations – manifolds are separate or the manifold of science is fully within the manifold of society – are not acceptable. In the first case the science is in an ivory tower and not sustainable in the long run, in the second case science fulfils only the wishes of society like a design bureau. The sustainable situation is that a part of science functions following the internal logic of science for moving the frontiers of knowledge according to the wisdom of scientists who know more than society at large; another part of science works with problems needed in society just in the present situation. And according to Lotman (2001), the most important problem is the communication between the non-overlapping parts. The society must understand the aspiration of science and have a trust in scientists. The balance between the parts is complicated, it is time-dependent, it is related to funding possibilities and maybe in the first place, it depends on the education of society. To reach a balance, the efforts from both sides are needed.

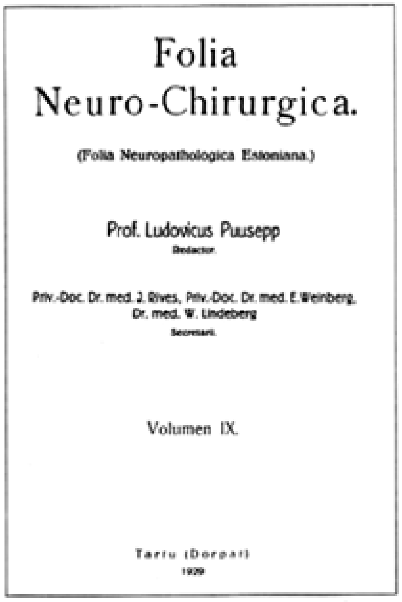
To sum up, the world should move towards the knowledge-based society. This is the goal of policies, programmes, structural changes, etc. of world organisations together with national governments and organisations. The interpretation of the 'knowledge-based society' varies because of its wide context and fast-changing world. In a policy paper of ALLEA on the European Research Area (Challenges 2007), the main characteristics of the knowledge-based society are given: (i) knowledge is a prerequisite for the quality of life and welfare of society; (ii) knowledge is based on good education and well-organised research structures; (iii) knowledge is disseminated fast and there are equal possibilities for everyone to obtain information; (iv) links between academia, society, industry and government are well-organised; (v) a knowledge-based economy uses all the potential of scientists and scholars, engineers and other specialists; (vi) innovation is encouraged at every level including industry-academia collaboration, social welfare, fiscal incentives, etc.; (vii) knowledge is a basis for policy decisions in society; (viii) dialogue between science and society is promoted. This is a challenge for all the nations to move towards a knowledge-based society. Further a brief overview on research policy and research structures is presented from a viewpoint of a small country – Estonia.

BRIEF HISTORICAL OVERVIEW ON RESEARCH IN ESTONIA

Formal scientific activities in Estonia began with the establishment of the University of Tartu by the King of Sweden, Gustavus II Adolphus, in 1632. After hectic changes in the 18th century due to several wars which passed over the territory of contemporary Estonia, the University of Tartu gained an international reputation in the 19th century. Astronomer Wilhelm von Struve, the embryologist Karl Ernst von Baer, chemist Wilhelm Friedrich Ostwald and others who worked at the University are known for fundamental contributions in their fields. Learned societies, the forerunners of the present Academy of Sciences, were formed during this period, as they were throughout Europe. In Estonia, these included the Estonian Learned Society (1838), the Literary Society of Estonia (1872) and Estonian Naturalist's Society (1853). The earliest scientific periodical published in Estonia was "Astronomische Beyträge" (1806–1807). By the end of the 19th century, the importance of knowledge and schooling was widely accepted in Estonia.

In 1919, after Estonia became independent, professors of the Tartu University started teaching in Estonian. Scientific terminology in Estonian and the education of the Estonian people in their native language was developed. At the same time, scientific and scholarly research prospered in several fields.

In the 1920s and 1930s Estonian research in astronomy, medicine, geobotany and oil shale chemistry gained a worldwide recognition. Astronomers



Folia
Neuro-Chirurgica.

(Folia Neuropathologica Estoniana.)

Prof. Ludovicus Puusepp

Priv.-Doc. Dr. med. J. Rives, Priv.-Doc. Dr. med. E. Weinberg.
Dr. med. W. Lindeberg

Volumen IX.

Tartu (Dorpat)
1929

in Tartu, led by Ernst Öpik, developed a complex theory of the evolution of the galaxy, involving several pioneering ideas. The Estonian school of neuropathologists and neurosurgeons, led by Ludwig Puusepp, was recognised for its achievements all over the world. The periodical published by L. Puusepp "Folia Neuropathologica Estoniana" (1923-1939) was an excellent scientific publication of that time. The botanist Teodor Lippmaa made significant advances in the field of phytocoenosis and ecology. Considerable progress in the study of oil shale was made by Paul Kogerman.

Systematic studies in linguistics (by Johannes Voldemar Veski and Johannes Aavik) and humanities in general had an enormous impact on Estonian culture. The eight-volume Estonian Encyclopaedia was published in the 1930s. In 1936, Tallinn Technical University was established and became the centre of higher technical education and technical sciences in Estonia. It must be stressed that the technical education in Estonia started already in 1918 at Tallinn Technical College, the forerunner of Tallinn University of Technology. In 1938 the Estonian Academy of Sciences was founded as a body of prominent Estonian scientists and scholars involving also the scientific societies and institutions.

By 1939, prior to the outbreak of the World War II, Estonia had set up the basic research institutions needed for national development. All that abruptly changed by the Soviet annexation, and the World War II.

Under Soviet rule, Estonian science organisations were centralised and guided by the socialist ideology. Attention to national culture and heritage was minimal, some branches of research, such as marine research, were under a special control, whereas the social sciences were given no freedom at all. Despite the heavy pressure, science in Estonia, especially physical and natural sciences, continued to develop. The achievements of scientists in linguistics, semiotics, archaeology, psychology, and ethnology were remarkable, since they developed their fields despite the ideological pressure.

YEARS OF CHANGES AND RESTRUCTURING

The restructuring of the Soviet-style research and education system started as early as in 1988 after the declaration of sovereignty was approved by the highest authority of that time – the Estonian Supreme Soviet. Two evaluations greatly influenced the restructuring of research in general. In 1991–1992, the Royal Swedish Academy of Sciences and the Royal Swedish Academy of Engineering carried out an evaluation of science in Estonia. The Evaluation Reports (see Evaluation 1992, as an example) pointed out several strong fields but also indicated needs for changes. Among the recommendations were: restructuring the policies for decision-making, integrating research and education and paying more attention to the PhD students. The next evaluation took place in 1994 and was carried out by local panels because the aim was to propose a more effective system of research establishments. Changes started everywhere, as said before, already after 1988, but in R&D after first changes, more steps were taken based on these two evaluations. The main aims in research policy were centred on the following principles:

- developing strong partnerships between science and society;
- ensuring the efficiency and effectiveness of the public research;
- promoting education;
- contributing to international research.

The restructuring certainly took some time. Altogether, the main changes were the following:

- Estonian Science Foundation was established (1990);
- a new system of academic degrees was introduced (1990);
- Estonian Science Council as the main decision-making body in R&D policy was established (1991) and reorganised into the Research and Development Council (1994);
- the Law on R&D was passed in the Parliament (1994);
- the Law on Universities was passed in the Parliament (1994) and some years later – the Law on the Academy of Sciences (1997).

This laid a legal basis for the contemporary system of R&D. What was important – the funding of research was linked to quality. The Science Competence Council (established in 1997) under the Ministry of Education and Research was (and is) responsible for long-term funding of research themes (as enacted by the Law) over five-six year periods with accompanying funding of the infrastructure and now also for a certain institutional funding. Estonian Science Foundation was (and is) responsible for research grants. The grants for post-docs were given by the Science Competence Council, now this obligation is vested in the Estonian Science Foundation. The duties of the Ministry of Education and Research also included coordination and funding of international research cooperation on state level.

Looking back now, it is clear that the new system of funding including long-term part for funding themes which guaranteed continuity of research plus grant-type projects, both based on quality requirements, paved the road to progress of the research. It cut off the low-level research which was inherited from the old system and the mindset of the research community was changed – excellence in research counts! The system to put more emphasis on the top-level research took more time to build up but in 2001 the first National Programme for Centres of Excellence started with centres launched in 2002 (Excellence 2002; Centres of Excellence 2004). More on restructuring and the public funding can be found in (Engelbrecht 1998; Masso, Ukrainski 2009).

HOW DOES IT WORK?

Beside the legal acts and funding schemes, the targets must be set up for successful research. Research must have freedom and nobody can predict the results across the frontier of research – society-driven and industry-driven research should be targeted according to the needs. In addition, scientific community must be developed together with education and science-society links. It has become clear world-wide, especially after the WWII that very general forward-looks called research strategies help set up the goals for science and society. It is important for all countries, large or small (Engelbrecht 2002). The first R&D Strategy "Knowledge-based Estonia" was formulated for 2002–2006. Not all the goals were reached, mostly due to insufficient funding because of other societal needs. The next Strategy called The Estonian Research and Development and Innovation Strategy "Knowledge-based Estonia II" (2007–2013) (2007) makes an attempt to link research and innovation. The targets are educated society, excellent scientists and innovative leaders in economy. It also envisages international cooperation as one of the drivers for research in all the fields. This is quite clear because the problems of critical mass, typical for small countries, can be solved only through cooperation. On the one hand, the cooperation opportunities offered by Framework Programmes and by international research organisations have been widely used and on the other hand, the mobility schemes for Estonian researchers, especially for younger researchers have been supported.

Box 1. Estonia, some data.

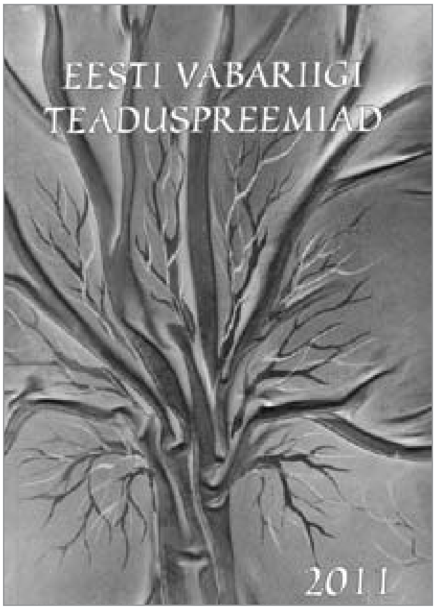
Population of Estonia as of January 1, 2011		1.34 mln
Gross Domestic Product (GDP)	in 2010	14.500 mln euros
	in 2009	13.900 mln euros
Number of researchers incl. engineers		7,453
Number of researchers incl. engineers in full-time equivalents (fte)		4,314
R&D expenditures in 2009		197 mln euros
R&D expenditure % of GDP	in 2009	1.42

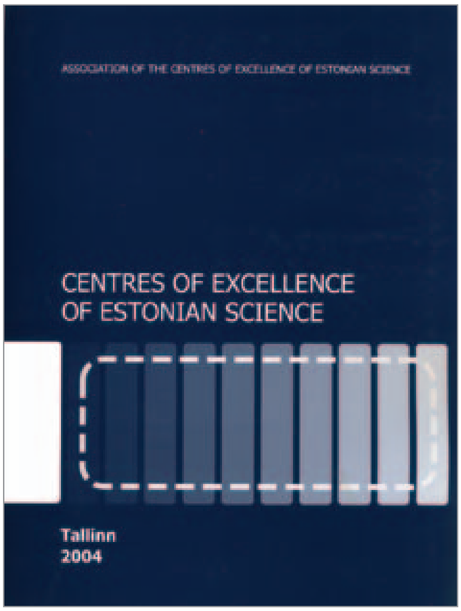
One of the important problems in Estonia is to improve conditions of research infrastructures. During the last couple of years, the EU structural funds have been used for that purpose which have already improved the situation and will certainly have an impact in the future. Another significant step taken recently was related to the Organisation of Research and Development Act. Amendments to the Act were passed in the Parliament (2011) which is going to regulate the positions of young researchers and to bring most of the public research funding under the aegis of one Agency.

WHERE DO WE STAND?

The best results in research starting from 1997 are described in the series "Estonian Research Awards" published by the Estonian Academy of Sciences (Research Awards 1997–2011). Another series reflecting the more general overviews is "Scientific Thought in Estonia". These publications started in 2002 and the volumes up to now cover technical sciences (2002, 2007), exact sciences (2006), medical science (2005), humanities (2009), life and earth sciences (2011), and marine science (in preparation). These both series are meant for scientific community and public in Estonia and published in Estonian.

The world of knowledge must open to everyone and nowadays the research results are collected in international databases which beside their main value – archives of knowledge – allow also to monitor and to compare the progress in various fields, as well as in various countries. Dynamics of

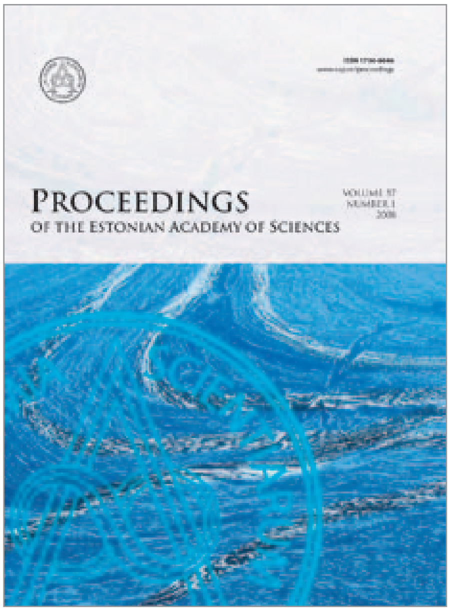




Publications on excellence in research.

scientific publications in Estonia shows a clear tendency of growth in the world scale. Included into the present volume are also the recent bibliometric data (Allik 2008).

One should also mention here that 7 journals published by the Estonian Academy of Sciences belong to the list of the ISI Web of Science, 8 – to Scopus, and 3 – to the database ERIH, not speaking about the specialised databases. In addition, the electronic journal "Folklore", published by the Estonian Literary Museum also is indexed by the ISI WoS. Speaking about rankings, the University of Tartu belongs (2011) into the list of 550 best universities of the world according to the QS World University rankings and Tallinn University of Technology also appears (2011) in the list. According to ISI Web of Science (Essential Science Indicators), the University of Tartu belongs into the list of 1% best research centres of the world in 9 fields of science and scholarship: chemistry, clinical medicine, animal and plant sciences, environmental sciences and ecology, general social sciences, materials science, geosciences, biology, biochemistry. There are various ways to reach excellence. In some cases, excellence in research has a long history like in astronomy and evolutionary biology (see above) and we can witness the leading role of Estonian researchers in cutting-edge studies on dark matter physics and large-scale structure of the universe, population genetics, etc.



Some research journals.

In other cases, however, like in materials science and informatics, the studies have progressed over recent years to be recognised in the world. One is clear – it needs dedicated scientists and scholars. However, the working conditions, i.e. environment and the infrastructures should also be on the level.

As said before, the National Programme for Centres of Excellence started in 2002 funded by the general R&D budget. The overviews on Centres of that time and their results are described in two publications (Excellence 2002; Centres of Excellence 2004). In 2008, the second call for Centres of Excellence was launched with the support of the EU structural funds (funding for 2007–2013). According to this call the following Centres exist: Frontiers in Biodiversity Research; Genomics; Computer Science; Integrated Electronic Systems and Biomedical Engineering; Chemical Biology; Cultural Theory; Translational Medicine. The next call in 2011 added some more: Environmental Changes; Dark Matter in Particle Physics and Cosmology; Nonlinear Studies; High-Tech Materials; Mesosystems (funding for 2011–2015). All these 12 Centres passed mandatorily a strict international evaluation and their studies are at the frontier of science and scholarship.

Estonia is one of the smallest states in the EU and also one of the smallest language communities. That means a responsibility to study and develop the Estonian language which belongs to the Finno-Ugric group of languages. This

University of Tartu, the main building.

Tallinn University of Technology, the main building.

Tartu Observatory (Tõravere).

is why studies in linguistics, culture and archaeology are important and they all are embedded into a wider context of Northern Europe. One of the Centres of Excellence (Cultural Studies) and several National Programmes (Language and cultural memory, Speech technology) indicate the aspirations in this direction. Beside linguistic studies, the Estonian Text-to-Speech synthesiser is an excellent example of recent results.

Contemporary research does not know borders and international cooperation is essential, especially for smaller countries like Estonia. The list of activities of Estonian institutions and scientists in the international scenery is long. Certainly in Europe the Framework Programmes are essential. Estonia could already take part in the FP3 starting 1993. The FP5 was fully accessible and since then the FPs have been an important part of our research. The succeeding rate of applications has been high and if we consider relative ratings, Estonia is among three most successful new member states in the FPs. The return of funding from the FPs is bigger than the needed input into the common pot. For example, the return is 0.25% from the FP7 Budget while Estonian GDP is 0.1% from the EU Budget. But certainly the FPs are not the only tools. Estonia has signed the agreements with CERN and ESA, has a membership in EMBC, takes part in the COST and EUREKA projects, etc. The ERA-NET and ERA-NET plus programmes have also given new impetus to research: BONUS in marine sciences, BIODIVERSA in biodiversity studies, HERA in humanities, PRIOMED-CHILD in children medicine, Complexity-NET in complexity studies, etc.

The Estonian participation in the ESFRI objects is planned in several areas: in studies of Estonian language resources, in genomics, and in structural biology; participation in the European Social Survey and the European Spallation Source is also important. These activities are listed in the Estonian Research Infrastructures Roadmap (2010) which beside the already mentioned activities shows also other fields where infrastructure is important for future studies. In linguistics, Estonia is a hub for studying Uralic languages with many contacts with smaller languages of this group. Estonian younger scientists have been successful in winning very competitive Wellcome Trust and Howard Hughes grants and the Estonian Humboldt Club has many members. The EU schemes like Marie Curie and ERASMUS are actively used.

The relations between academia and industry are important for every country. In Estonia these relations are steadily improving. More specific data on the present stage of innovation and on prospects of its development will be given in a separate paper in this volume.

ARE WE ON THE RIGHT TRACK?

The brief description on developments of research and R&D policy in Estonia must be reflected in the general framework of the knowledge-based society (see above).

Much is done and the main contributions of scientists and scholars to society are without any doubt related to excellent research results in many fields. There are many (but not enough!) young researchers, infrastructures have been improved, some state programmes have been launched, etc. From the side of applications, IT should be noted. Estonia is known by effective implementation of information technologies: e-banking, e-voting, e-governance, e-taxation system, etc. Skype, by the way was invented by young Estonian engineers. This invention might even be considered as a symbolic process of internationalisation – the more we speak to each other, the more we understand each other.

It is tempting, however, to turn here also to complexity studies which are developing fast in the entire world (see above). It is noteworthy that complexity studies are carried on in several research centres in Estonia, like the Centre for Nonlinear Studies (CENS) of the Institute of Cybernetics at Tallinn UT and several institutes of the University of Tartu. The studies in CENS, for example, involve nonlinear dynamics, systems biology, wave engineering, nonlinear control, soft matter physics and fractality, nonlinear optics, etc. Closely related to CENS is the Laboratory of the Proactive Technologies in Tallinn UT. The bioinformatics is a topic in the Institute of Computer Science of University of Tartu; the biosemiotics is studied in the Institute of Philosophy and Semiotics of University of Tartu. The links with the EC Complexity-NET and possible co-operation with one of the EU FET Flagship Programme called FuturICT place these studies into the international scenery. The results in analysis of fiscal time-series are applied in the Swedbank, the proactive technologies are linked to many applications (see this volume), the wave engineering has fundamental results explaining the wave climate in the Baltic sea (see this volume), the bioinformatics research in University of Tartu focuses on the analysis of gene expression data, etc. Complexity studies are explained at various meetings in the Estonian Academy of Sciences (including a seminar on the complexity of state) and at the meeting of the Academic Council of the Estonian President. The knowledge on the importance of complexity is growing.

As explained above, the goals for R&D&I are envisaged in the corresponding strategy (Knowledge-based Estonia 2007) and several programmes are launched by the Ministry of Education and Research (on centres of excellence, on infrastructures, on internationalisation, etc). For that purpose the EC structural funds are used. Excellence is a leading criterion for funding decisions, although the process of restructuring of funding schemes is going on and not everything is clear. Attention is accorded to the doctoral studies and here again the situation is not uniform because beside excellent graduate schools there are also those who should raise their level. The number of PhD graduates is still smaller than expected. The international cooperation is encouraged, there are several schemes to get travel grants and invite PhD students and post-docs as well as experienced researchers from abroad to study or work in Estonia. The science-society relations could be better but there are several schemes to engage media for explaining research for a larger public. The Science Centres in Tallinn and Tartu have fascinating displays and events for every age; in 2011 the Centre in Tartu has opened a brand-new house with new exhibitions including the 4D cinema. And the Physics Bus run by the enthusiasts from the Institute of Physics of University of Tartu won a Descartes Prize for their educational work.

Are we on the right track? In 2010, T. Maimets, the Head of the Estonian Science Foundation said in a special Section of Nature, "The history of science in Estonia spans centuries; however, never has its scale been so comprehensive and its targets so high" (Research in Estonia 2010). The progress in research is very much supported by the funding system which is based on quality criteria concerning the fundamental targeted financing as well as grants, baseline funding and infrastructure support. But funding without talented researchers is not sustainable. Although the critical mass for small countries is always a problem, the tools to keep young scientists and scholars in Estonian research centres seem to work.

FINALLY, THIS BOOK

The cornerstone for a knowledge-based society is research. Collected in this volume are overviews on many successful fields of research in Estonia which will give a picture on the level of research and aspirations for the future. Certainly the presentation of the research scenery in Estonia within these covers is not fully complete and does not pretend being a full foresight. The emphasis is put on fast progressing fields.

We let all the Divisions of the Estonian Academy of Sciences nominate the most relevant and well progressing topics and find the authors. In this way representative overviews presented in this book are divided into 4 Chapters corresponding to the Divisions of the Academy: Astronomy and physics, Informatics and engineering, Biology, geology and chemistry, Humanities and social sciences. Under these traditional names of Divisions, the Chapter 3 includes also medicine. Quite obviously, not only the research in Centres of Excellence mentioned above is reflected in the book but much more. The overviews bear the imprint of authors because no strict rules and formats for papers have been introduced. Finally, after the overviews, one paper describes the bibliometric indicators in order to position the research in Estonia against the larger world scale. To sum up, the idea of this book is to demonstrate that research matters in Estonia and Estonian researchers are a part of international scientific community.

Many questions are not answered. Although the forward-looks are described in many overviews, we do not know today in which areas one could expect fast breakthroughs. What are the fields for inter- and trans-disciplinary studies? How should one develop the better relationships between academia and society? Is the present research scenery for a small country inhabited well enough or is there something essential absent? Clearly such a list of questions could be prolonged.

The book is meant for everyone who is interested in science and scholarship in Estonia, let them be researchers themselves or people dealing with science policy and forward looks, let them reside in Estonia or abroad.

We hope that the Estonian representatives in other countries also find this book useful for representing ideas of scientific community in Estonia as a part of the endless enlargement of knowledge in the world.

And it will be readers who decide whether researchers in Estonia see the world as a whole and are able to put the pieces back into the general picture of knowledge.

REFERENCES

Allik, J. 2008. Quality of Estonian science estimated through bibliometric indicators (1997-2007). Proceedings of the Estonian Academy of Sciences, 57, 4, 255-264.

Centres of Excellence of Estonian Science. 2004. Association of the Centres of Excellence, Tallinn.

Challenges of the Future: Reflections of ALLEA on ERA. 2007. ALLEA, Amsterdam.

Declaration on Science and the Use of Scientific Knowledge. 1999. World Congress on Science, UNESCO and ICSU. Science International, ICSU, 2-6.

Engelbrecht, J. 1998. Scientific development in a small country. Estonia – Business Handbook 1998-1999. Tallinn, 175-185.

Engelbrecht, J. (ed.) 2002. National Strategies of Research in Smaller European Countries. ALLEA and Estonian Academy of Sciences, Amsterdam.

Engelbrecht, J. 2006. Fractals for the European Research Area. EC RDT info, Aug., 46, 14-15.

Érdi, P. 2008. Complexity Explained. Springer, Berlin et al.

Estonian Research Infrastructures Roadmap 2010. Estonian Ministry of Education and Research, Tartu.

Evaluation of Estonian Research in Natural Sciences. 1992. Swedish National Research Council, Stockholm.

Excellence in Research 2001-2002, Estonia. 2002. Estonian Ministry of Education, Tartu.

Juarrero, A., Rubino, C. A. (eds.) 2008. Emergence, Complexity, and Self-Organization: Precursors and Prototypes. ISCE Publ., Goodyear, Arizona.

Knowledge-based Estonia. Estonian Research and Development and Innovation strategy, 2007-2013. 2007. Estonian Ministry of Education and Research, Tartu.

Lotman J. 2001. Culture and Explosion. Varrak, Tallinn (in Estonian).

Masso, J., Ukrainski, K. 2009. Competition for public project funding in a small research system: the case of Estonia. Science and Public Policy, 36, 9, 683-695.

Nicolis, G., Nicolis, C. 2007. Foundations of Complex Systems. World Scientific, New Jersey et al.

Prigogine, I. 1980. From Being to Becoming: Time and Complexity in the Physical Sciences. Freeman, San Francisco.

Research Awards in Estonia. 1997-2011. Estonian Academy of Sciences, Tallinn (in Estonian).

Research in Estonia: race for quality and competitiveness. 2010. Section of Nature, Nov 25.

Scientific Thought in Estonia: technical sciences (2002, 2007), medicine (2005), exact sciences (2006), humanities (2009), life and earth sciences (2011), marine sciences (in preparation). Estonian Academy of Sciences, Tallinn (in Estonian).

Toffler, A. 1984. Science and change, preface. Prigogine, I., Stengers, I. Order Out of Chaos. Heinemann, London, xi-xxvi.

COSMOLOGY NOW AND IN THE NEAR FUTURE

E. Saar

Cosmology means different things to different persons. In this chapter I will talk about physical cosmology – the scientific picture of our Universe, its past and its future. I will describe the recent cosmological research, its present state and possible future trends. As proper for the present book, you will notice an Estonian accent.

COSMOLOGY IN RECENT YEARS

When I started my career in cosmology, in late sixties, cosmology was a branch of theoretical physics. Cosmologists had mainly a general-relativistic background, built pretty idealised (smooth, homogeneous, isotropic) mathematical models of the Universe, and discussed if the Universe was really expanding or if it only seemed that it was. There were a few astronomers that tried to check the current cosmological models and to determine their parameters from observations, but this was not fashionable.

Our present knowledge about the Universe can be illustrated by a few iconic figures – these appear usually in almost every review article or review talk on cosmology. These figures are perfect for demonstrating how much we have learned in the last half a century.

The first figure illustrates the overall evolution of the Universe and the formation of the structure we observe now. In early 1960s, we could describe only the overall outline of expansion, and even that only partially. All the labels in the figure belong to the new cosmology. Inflation and formation of matter from quantum fluctuations was proposed in 1980's, the afterglow light (CMB, the cosmic microwave background) was discovered in 1964, the term 'dark ages' is less than 20 years old, and although there were hypotheses and theories of formation of galaxies and stars, these were pretty ad hoc and were isolated from the overall evolution of the Universe. The possibility of accelerated expansion existed only in theory, and was not taken seriously.

As usual, experiments (observations in astronomy) were the drivers for the change. Gradually, we were able to observe more galaxies, and to start to know the real Universe. But probably the most important change of the cosmological paradigm, and, maybe, of particle physics, came before that. That was the discovery (or pre-discovery) of dark matter.

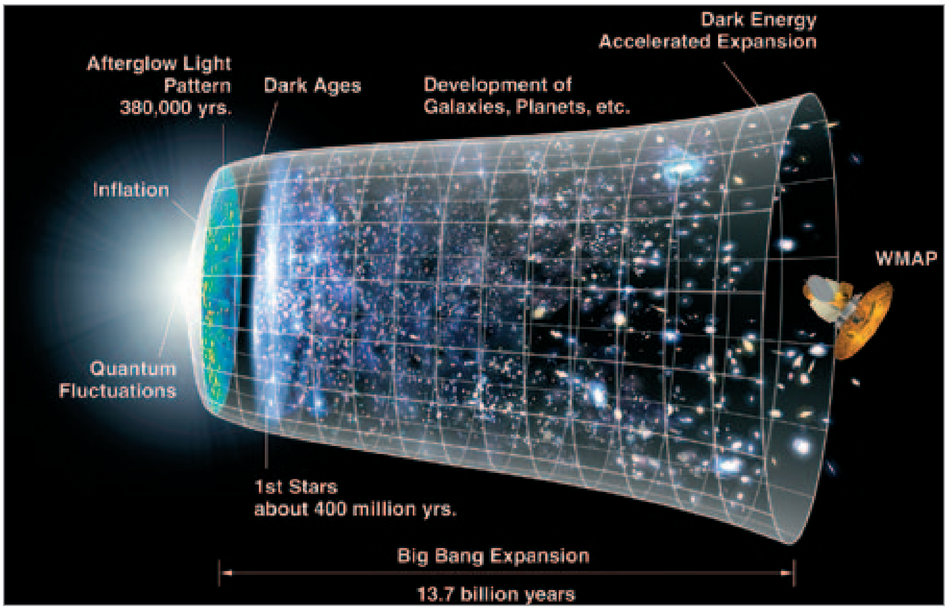


Figure 1.
The evolution of the Universe from its birth to the present day – the cosmic timeline. Credits: the NASA/WMAP team.

Dark matter

A Swiss-American astronomer Fred Zwicky proposed already soon after the second world war that galaxies in galaxy clusters, places where thousands of galaxies live together, had too high velocities to be gravitationally bound there. The only force that can keep galaxies together is gravity, and when Zwicky summed the masses of all the observed galaxies in a cluster, he found that their total gravity was clearly too weak to stop the galaxies leaving the cluster. Nobody could explain that fact, and observational cosmology was not fashionable in the astronomical community in those times, so Zwicky's troubles were forgotten.

The explanation came from a different branch of astronomy – stellar astronomy – that studied motions of stars in galaxies (and especially in our Galaxy) and tried to build exact models of star-composed galaxies. It was a popular branch of astronomy in sixties and seventies; there were many strong groups in the world, several in the Soviet Union, and one in Estonia, Tartu Observatory, led by Grigori Kusmin. One of his pupils, Jaan Einasto, embarked on a program of constructing realistic models of galaxies, and was soon surprised to find that the models did not work – the stellar velocities measured in galaxies were frequently too high to keep the stars inside (gravity was weak, again). His first idea was to assume that there were heavy bodies in the centres of galaxies, but then he realised that the additional mass should be spread throughout all Galaxy. Einasto and his two young colleagues, Ants Kaasik and myself, collected data on the velocities of smaller companions of several giant galaxies (as our Galaxy), stacked all these galaxies together (in a mind experiment, of course), and showed that the additional mass was located mostly around the galaxies. They called these mass balls 'dark coronas' (nobody had yet seen that material), and published a paper about that in "Nature" (Einasto et al. 1974). This last step was much more difficult to carry out than to formulate the idea – at that time, in 1974, it was an unheard of move in the Soviet Union, to publish abroad. And for every scientific paper sent abroad, the authors had to sign a document asserting that this paper contained nothing new and significant. Somehow we managed it, and our paper was published a month and a half before a similar paper was published in the "Astrophysical Journal Letters" by a group of well-known astronomers from Princeton (Ostriker et al. 1974).

Almost all astronomers thought that the idea was preposterous. We and the Princeton group asserted that there was about 10–20 times more dark matter than stars; astronomy was hundreds of years old, and nobody had seen that matter. Many arguments were proposed to explain the observations without the need for the additional gravity and additional matter; we fought back. The astronomical community started to accept the idea only after Vera Rubin's careful observations of outer regions of galaxies – the rotation velocities were too large there, and they did not change with the distance from the centre of the galaxy (Rubin et al. 1980). These facts demanded the presence of an additional extended massive component.

Nowadays, dark matter is a universally accepted part of our (theoretical) Universe. It explains naturally many astronomical observations, it is studied by several different observational methods, and, finally, particle physicists have started to take the idea seriously. We know now that dark matter must consist of a special elementary particle (or maybe several of those) that almost do not interact with each other and with other known particles, but which are predicted by most of the current physical theories. There are several costly experiments dedicated to the search for dark matter, one of them running on the GHC (Great Hadron Collider). But although we know that dark matter is out there, it has not been discovered (seen, observed, touched) yet.

Traces of structure from the past

While the discovery of dark matter was an unexpected observational result, the quest I will describe below had a strong theoretical motivation. We knew that the early Universe was probably pretty uniform (the same matter density everywhere), and that small possible deviations from that uniformity would have been amplified by gravity to form the present objects – galaxies, stars, and planets. General relativity predicted how fast this structure should form, and predicted how strong these small deviations should have been. We also knew that the Universe had started with a Big Bang, was very hot for a long time, and expansion caused it to cool down at a certain moment. This was the time when radiation separated from matter, and could propagate towards us, through the whole Universe. This 'microwave background radiation' ('microwave' because the expansion of the universe shifts optical radiation into radio frequencies), was observed first by Arno Penzias and Robert Wilson (Penzias, Wilson 1965). They were awarded the Nobel Prize for that, but much later. As matter was tightly tied together with radiation before it left towards us, all the information about matter, and about its density structure, had to be recorded in this radiation background, and we should be able to observe it.

The microwave background is weak, but the bigger the radio telescope, the smaller the signals it can detect. We knew how large the signals should be (we ourselves and galaxies have to exist to observe the Universe), calculated how large the telescopes should be, and started the observing programs to find the early traces of the present-day galaxies. Both the U.S. and Soviet radio telescopes were searching for that; I know better the Soviet story.

There was a special large radio telescope, RATAN-600, in the Northern Caucasus, and it was selected for the problem. It consisted of a 600-metre long circle of antennas that were several metres high, and its collecting area was perfect to search for signals from the distant past (about 10 billions of years before). Observations started, and no signal was seen. The detectors were improved, and yet no signals were seen. Theory asserted categorically that the traces of the present-day structures in the CMB should have the amplitude of 1/1000 of the observed mean amplitude; the detectors had better sensitivity, but there were no signals. Since we were there to observe the CMB, could the theory be wrong? When the director of RATAN-600, and the leader of the project, Yuri Pariiski, attended conferences, everybody expected to hear the positive news, but these never came. Those were frustrating times.

The long-awaited traces were found only in 1990, by an US satellite COBE (Cosmic Background Explorer; see Boggess et al. 1992). Their intensity was about 100 times smaller than predicted by the theory, and then, finally, astronomers understood why the prediction was wrong – it did not take into account dark matter. As dark matter lives by itself, the seeds of structure could grow in it, unhindered by radiation, and when the ordinary (we say 'baryonic') matter separated from radiation, it collected in the previously prepared dark matter 'galaxies'. The predicted intensity for the seeds of structure concerned dark matter, not baryonic matter, and it did not leave traces in the radiation background. This way, the COBE results confirmed indirectly the existence of dark matter. By the way, we realised later that the first detection of possible very low-intensity signals from the CMB was done by an experiment aboard a Soviet meteorological satellite, but it was pretty uncertain and did not attract attention.

Nowadays, measuring the seeds of structure in the CMB is a well-established method of study of the early Universe. The COBE results were recently confirmed and much improved by the NASA satellite WMAP (Jarosik et al. 2011). It observed the microwave sky during seven years (Fig. 2). An ESA mission, Planck, started collecting data in 2009, and will continue until 2012 (its detectors are cooled by liquid helium, and it will evaporate by then). The sensitivity and the angular resolution of Planck are several times better than those of the previous missions, and many interesting results will certainly be found. Planck is an international project (Europe, USA, etc.), and Estonians participate in it, also, via a joint project with Tuorla Observatory, Finland.

Large-scale structure of the Universe

This is, finally, the topic that was really driven by observations. In 1970s, any study that determined the redshifts in the spectra of about ten galaxies (their distances) merited a paper in a solid astronomical journal. The position of a galaxy in the sky is easy to find, and if we know its distance, we have a signpost in the Universe. For the geography of the Earth, such signposts were essential to produce maps. For the Universe, the maps have to be 3-dimensional, and many more signposts are needed. The volumes that have to be mapped are immensely larger than those we have encountered on the Earth.

Cosmologists are brave people. Lev Landau, one of the most respected physicists, said once "Cosmologists are often in error, but never in doubt". A good example for that is an early use of a 3-D galaxy map towards the Coma cluster (it was called 'The Slice of the Universe' and contained a whopping number of 200 galaxies). Brave cosmologists cut out a cube of a side of 50 Megaparsec (one Megaparces, Mpc, it is a typical distance between galaxies) from there (it contained 126 galaxies, about 5 galaxies per cube side). They

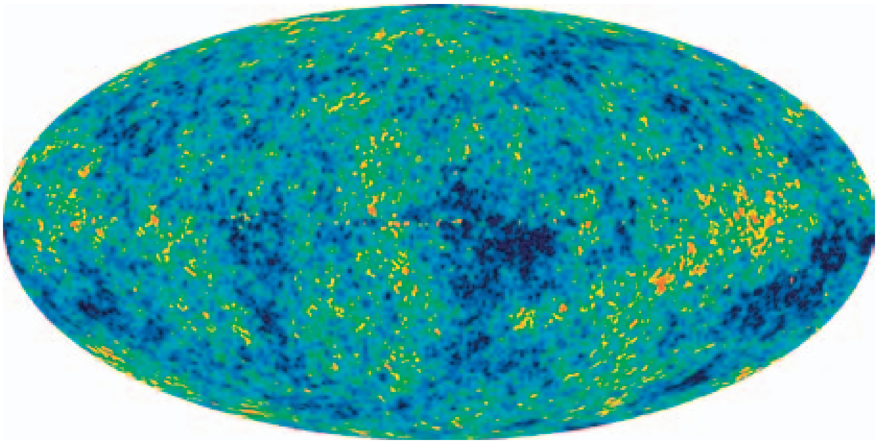


Figure 2.
The full-sky map of the small-scale temperature fluctuations of the cosmic microwave background, from seven years of WMAP satellite observations. Red denotes the highest temperatures, blue – the lowest temperatures; these change from 20 micro-Kelvins below the mean to a similar height above the mean. Galaxies will form where the temperature is high. Credits: the NASA/WMAP team.

continued to calculate the galaxy density in this cube, determined its topology and proudly declared that the topology of the matter distribution in the Universe is Gaussian.

Fortunately, these times are in the deep past and we do not have to be that brave any more. The art of acquiring galaxy redshifts has taken an enormous leap, and the two largest galaxy redshift catalogues (3-D galaxy maps) include about 200,000 galaxies (2dF GRS, the Australian Two-degree Field Galaxy Redshift Survey), and close to 600,000 galaxies (SDSS, the Sloan Digital Sky Survey). The volumes that these surveys span are about one Gpc (Gigaparsec, 1000 Mpc) cubed, meaning that their size in every direction is about one Gpc.

The maps of the Universe are pretty peculiar, and have looked that way from the very beginning. The maps built in Tartu were among the first (both two-dimensional maps of spatial slices, for printing, and a three-dimensional map of coloured plastic balls, for understanding); we coined the term 'cellular structure of the Universe' and held the first international conference (sponsored by the International Astronomical Union) on the large-scale structure of the Universe, in Tallinn, 1977. Compared with the common picture of single or multiple Suns, planetary systems around them, clusters of millions of Suns, and galaxies of thousands of star clusters, the large-scale structure is radically different. The compact ('island') systems here are huge voids empty of galaxies, sparsely populated walls of galaxies that delineate the voids, galaxy filaments that border the walls and connect galaxy clusters that serve as nodes in the network of filaments and walls. As the filaments and clusters are the most visible details of this arrangement, this structure is called nowadays 'the Cosmic Web'. We saw that web in our first maps and have actively studied it for more than 30 years. I include the galaxy map of the 2dF GRS in Fig. 3, where you can see all the details described above.

Although we have not been instrumental in the observing and data processing of the new huge maps, we have been active in interpreting them, designing new statistical tests and explaining the results. The early galaxy maps were produced by Mihkel Jõeveer and Jaan Einasto (Einasto et al. 1980).

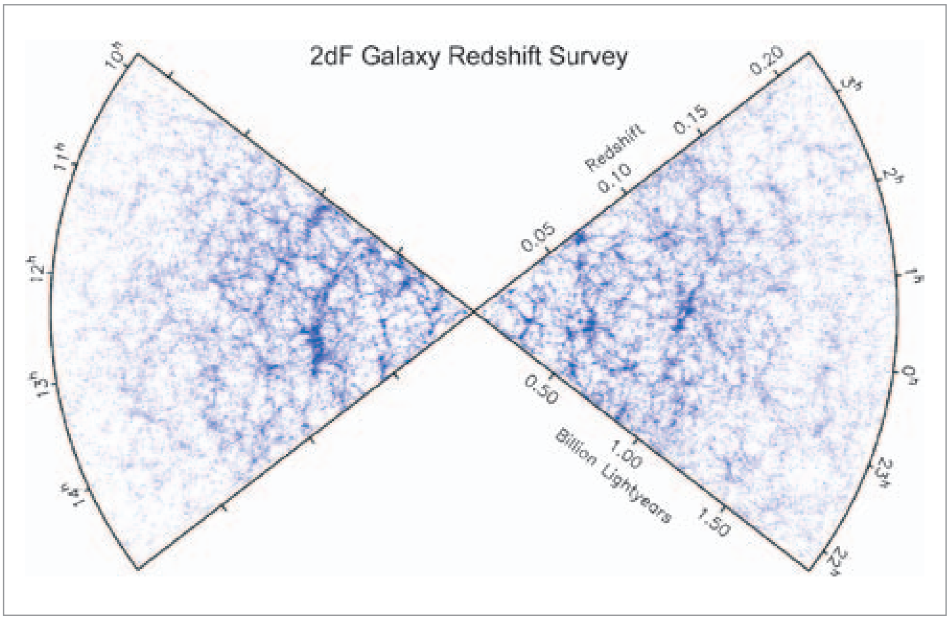


Figure 3.
The map of galaxies in two thin slices (6 degrees thick) of the 2dF GRS, the Australian Two-degree Field Galaxy Redshift Survey (Colless et al. 2001). Galaxies are shown as tiny dots. The cosmic web of huge voids and filaments is well seen. The streaks of galaxies pointing toward the origin (us) are called the 'fingers-of-god') and are caused by fast motion of galaxies in rich clusters. The farther we go, the fewer galaxies we see – they become too dim and we cannot measure their distances any more. Credits: the 2dFGRS team.

Erik Tago has constructed the most reliable catalogues of galaxy groups in the world (see, e.g., Tago et al. 2010), Maret Einasto has used them to construct the catalogues of the largest elements of the cosmic web, superclusters of galaxies (Einasto et al. 1997).

There is one more aspect to the large-scale structure – numerical modelling of the formation and growth of such structure in the Universe. As the problem is 3-dimensional, this work demands state-of-the-art computers and methods. The best models are computed using the fastest computers on the planet and their data is shared among all astronomers. For questions of principle, smaller computers, such as we have, are good enough, also. Our Mirt Gramann was the first to model the growth of structure in a dark-energy dominated Universe (Gramann 1988); it was considered freaky in 1987, but is now the only accepted model. Her student Ivan Suhhonenko is our main expert in numerical models nowadays, and together with Jaan Einasto he is studying how different scales of the structure develop and interact with each other. Ivan Suhhonenko and our colleagues from Tuorla Observatory (Finland) built first detailed numerical models of light-cones. All our observations describe the light-cone – the further we observe a galaxy, the more it is in the past.

Baryonic oscillations

Studies of the statistical properties of the galaxy arrangement in the Universe, the cosmic web, led in recent years to an unexpected result. Looking for the distribution of distances of galaxy pairs, cosmologists found that there was an overabundance of galaxy pairs that were located at about 150 Mpc from each other (Eisenstein et al. 2005). As the typical distances between galaxies are around 1 Mpc, and the largest elements of structure, huge voids, measure about 40-60 Mpc, 150 Mpc is a very large distance for cosmology. This time we knew our theory, and explained this feature as a remnant of the motions of the matter at the end of the cosmic fireball. As at that time those motions were governed by the pressure of the radiation-baryon plasma, the effect was called 'baryonic oscillations' (BAO), the cosmic sound. The scale of BAO provides us with a standard rule to measure and compare the rate of expansion of the Universe. As this rate is a direct probe for the physics and energy content of the Universe, studies of BAO have become very popular, and will continue to be so. The details of BAO and possible deviations from the present simplest theory may give us a handle to learn about the mysterious inflation (extremely rapid expansion) period that is supposed to precede the much quieter Big Bang.

BAO has an Estonian part, too. Gert Hütsi was among the first to find the traces of BAO in the power spectra (a frequency modulation representation of the sound) of the observed galaxy distribution (Hütsi 2006), and I am a member of a Spanish-French-Estonian group that discovered the traces of the cosmic sound, huge galaxy shells in the observed galaxy maps (Arnalte-Mur et al. 2011). As these are real-space features, they may give us extra information about the cosmic sound.

Supernovae and dark energy

Supernovae are death explosions of massive stars. As stars of the same mass should be similar everywhere and at all times (their evolution and death is governed by thermonuclear reactions), supernovae are thought to provide a standard candle, objects of known luminosity. A supernova is frequently more luminous than its home galaxy of hundreds of billions of stars, and can be observed from enormous distances. The only problem is to find a distant supernova – the explosions are rare and unpredictable. However, two special supernova observation projects were successfully carried out in mid-nineties, and used to measure the rate of expansion of the Universe. For most astronomers and physicists, the result was unexpected – in recent past, the Universe accelerated its expansion (Riess et al. 1998). If the expansion was due to the initial Big Bang, it should only be decelerating due to the gravity of the whole Universe; acceleration meant that some other mysterious force was at work. This force was called 'dark energy' and its nature is unclear yet, although more than ten years have passed since the discovery.

Together with dark matter, dark energy dominates the total mass-energy content of the Universe, and determines its evolution. I illustrate it in another iconic figure – the cosmic pie diagram (Fig. 4).

'Dark energy' might be a manifestation of new physics, and a multitude of hypotheses have been proposed to explain it. At the moment, none of them is completely logical and well connected with the present physics. But as dark energy accounts for 72% of the total energy content of the Universe, understanding its nature is of utmost importance.

Astroparticle physics

A new popular research direction in astronomy is astroparticle physics. It is the tightest connection between physics and astronomy these days, and it exploits the fact that the Universe and its constituents constitute the best physics laboratory possible. On the Earth, it is impossible (or almost impossible) to reach the temperatures, pressures, magnetic fields, matter densities, particle energies etc. that are common somewhere in the Universe. The

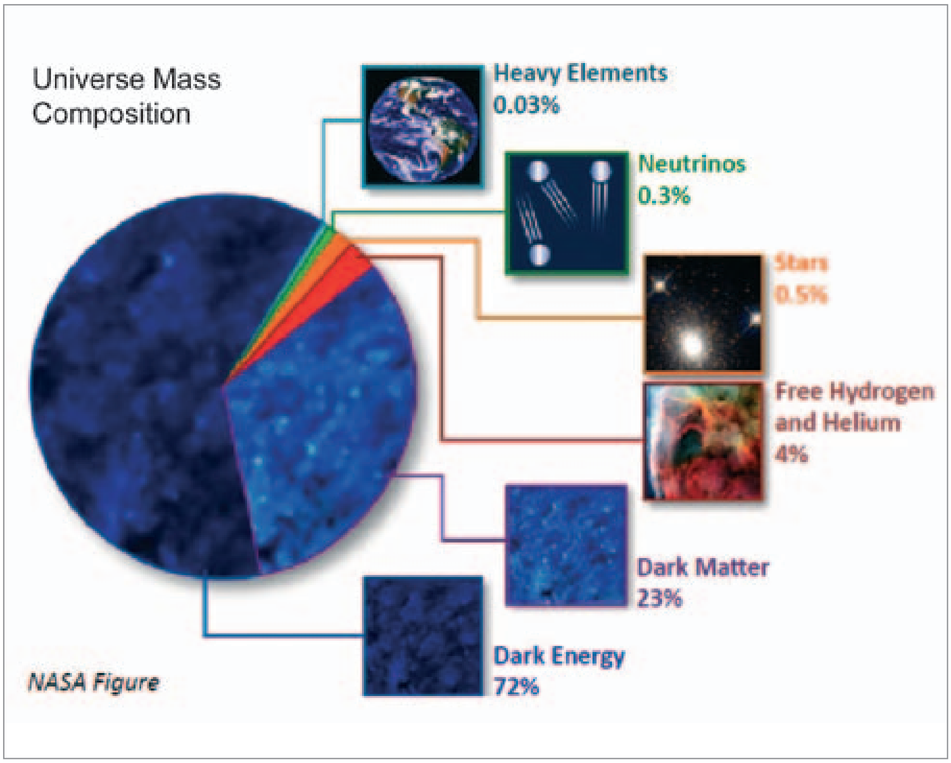


Figure 4.
The mass (energy) balance of the Universe. The main component is dark energy, and together with dark matter, these unknown components form 95% of the energy-content of the universe and determine its evolution. Credits: NASA.

downside to that is that experiments cannot be controlled, and all deductions have to be made on the basis of frequently indirect observations.

Such studies need tight collaboration between particle physicists, who know what to look for, and astronomers, who know where to look, and what the processes are that may wreck the experiment (observation). The most popular subjects of study in recent years have been the search for weakly interacting dark matter (e.g., its annihilation radiation), and the study of extremely high-energy particles that cannot be produced in present-day accelerators. In Estonia, we observe a start of interesting collaboration between particle physicists who participate in the CERN LHC dark matter experiment, and young cosmologists from Tartu Observatory, searching for traces of dark matter in recent satellite data (Hütsi et al. 2009). This collaboration will certainly widen, and will open interesting opportunities for both groups.

In the summary

How can we characterise the last 50 years of development of cosmology?

As described above, cosmology has matured, changing from theoretical speculations to a data-rich science. We have collected experience in conducting complex experiments and observational programs. 50 years ago we could not even imagine costly space missions that are especially designed to answer cosmological problems that take about 15 years to prepare and 5 years more to process the data; such missions are commonplace nowadays.

We have great galaxy catalogues, huge amounts of data on galaxies, their locations and properties. Our present catalogues of galaxy distances contain about a million of galaxies; compare that with the catalogues of a couple of hundreds of galaxies we had in 1960's.

We have powerful computers that cannot be compared with those we had 50 years ago. Computers and numerical modelling have become indispensable tools in all areas of science and in cosmology also. We need them to process the large amounts of data we have, to carry out cosmological simulations and to understand them. There is a wide field of work for that, as our computers will never be as powerful as the Universe is.

Our community has grown. There were fewer than a hundred cosmologists worldwide 50 years ago, and the papers were few. Nowadays cosmology is probably the main branch of astronomy, with scores of new papers appearing every day. We have enthusiastic graduate students and young PhD's who will carry on cosmological research and define its status in the near future.

FUTURE TRENDS IN COSMOLOGY

Predicting future trends in science is a non-scientific task. Real discoveries will always happen serendipitously, without any planning. The present project-based management of science tends to discard interesting ideas and focuses on projects where the result will certainly be positive. This leads inevitably to studies where the result is known, in principle, or can be estimated already now, and science will give only next decimal points. A good example of that are the latest (2010) US Academy of Sciences white papers on astronomy (*White Papers*), where leading experts in the world formulated the most important directions in astronomy. These papers include 117 cosmological projects and 103 projects on galaxy evolution, some of them overlapping. Most of them are continuation of present studies, with emphasis on better instruments, more data, smaller sample errors, etc. This is the only way to go in our project-based world. Unfortunately, this approach cannot promise real scientific breakthroughs. These will happen when there are crisis moments, when our data defy explanation by known physics; fortunately, there are a couple of such problems in cosmology, too.

There are, of course, research directions where long-term planning is necessary. These include space missions, and construction of large facilities (accelerators in particle physics, telescopes and their arrays in astronomy). New instruments carry always the possibility of completely new knowledge, and are thus instrumental for the progress of science. This is especially true in astronomy, where our eyes (telescopes) are always limited by the dimness of distant objects, by the instrumental resolution, by the wavelength interval, by the time needed to collect the signal. Any new instrument that improves on these limits will certainly see new objects and processes in the vast Universe, and will lead to serendipitous discoveries.

Another case of necessary planning is education. It will take five-six years for students to become scientists (post-docs, who, as the lore tells, do most of the work). But, as is the case with instruments, a talented and well-educated student can always become a next Nobel winner or a project leader.

Of course, science can be done in the present way, too, and as it is financed by the public, the public has to know how we plan to spend their money. The only question is whether it is the most cost-effective way. Supporting scientists who spend years pondering a single problem does not fit into this scheme; maybe private institutions can do that? A good example is the Perimeter Institute for Theoretical Physics in Canada, supported jointly by private companies and the government. Another example is the line of Kavli Institutes, financed by the Kavli Foundation. So, considering the whole planet, we can hope for both better data (by the project-based planning) and for unexpected ideas.

Having formulated my complaints, I have to start predicting the future of cosmology. I suspect that the present volume is meant to predict the development of science in Estonia. But science is truly international nowadays, and Estonian cosmology is Estonian participation in the world cosmology. As Estonia is small and its resources are limited, our participation is mainly personal collaboration with other groups and personal participation in international research projects. The future of cosmology in Estonia is the same future as that of the whole cosmology done on this planet (and on other planets, too, in our Universe).

So, which are the most promising trends in the present-day cosmology? I will divide these by the type.

Important problems

The most important problem is driven by the present crisis in cosmology. In fact, we have not one, but two separate crises, one of the nature of dark matter and another of dark energy (see the pie diagram above). Both of them together form about 95% of the total energy content of the Universe, and determine its history and evolution.

Although we have not seen dark matter particles yet, we know pretty much about it. We know where it is located and we know that it is either weakly interacting or not interacting at all. There are many hypotheses about its nature, many of them following from the present picture of elementary particles and fields, but many of them completely *ad hoc*. It seems that astronomers cannot do much more to clarify the picture; the ball is in the hands of particle physicists. They are taking the problem very seriously – several sophisticated high-volume detectors have been built or are close to completion, and a special experiment is running in CERN, on the LHC (Large Hadron Collider). There is strong Estonian participation in this experiment, a group from the National Institute of Chemical Physics and Biophysics (NICPB), Tallinn, led by Martti Raidal; he has certainly explained the problem in his chapter in the present book. The cosmology group from Tartu Observatory will offer their help, using astronomical observations (frequently by satellites) to check and prove or disprove phenomenological hypotheses of different manifestations of dark matter. This collaboration is barely a couple of years old, but it works already, and we plan to keep close contacts in the future.

Dark energy will certainly be the most important (although maybe not the best financed) research problem in cosmology in coming years. The project-based approach is to try to study in detail the history of the expansion of the Universe, all the deceleration-acceleration play. For observational cosmologists, this will give the possibility to check the multitude of theoretical predictions. The clearest traces of dark energy are seen in the varying speed of expansion of the Universe.

Expansion can be measured in many ways, one of them being the speed of growth of the structure. Another popular method is to use baryonic oscillations that give us a well-known physical scale, a standard ruler. There are many projects for that, several of them Earth-based, using large new telescopes, and new space missions have been proposed.

Different theories of dark energy lead also to different Universes, with a different large-scale cosmic web. These theories can be tested by comparing their Universes (computer-simulated) to the observed Universe. As the cosmic web is a complex phenomenon, it is not easy to reproduce it in full detail by our computers, and the possibilities to check and discard theories are many. Computer simulations have become an extremely useful tool lately, mostly thanks to the fast development of computers, their speed, usable memories, parallel computations, etc. Rapid advancement in many-dimensional visualisation also helps.

Among the first questions such simulations could solve is the dark energy equation-of-state (the relative role of dark energy in the evolution of the Universe at different epochs). Many dark energy hypotheses can be tested by comparing topological properties of the cosmic web in models and observations.

But the dark energy problem could also provide us an opportunity of unexpected, principal, theoretical development. The natural way to go is to get help from the string theory (theories), and/or to develop quantum gravity. If dark energy needs string theories and brane universes to explain it, new gravity theories might appear. Explaining dark energy might raise questions about the nature of gravity that nobody was willing to test before, and which we have seen and read only in science fiction books – antigravity, screening of gravity, etc. The problem of the very early evolution of the Universe (its birth) belongs also here. And, maybe, we can get a solid theoretical handle on multiverses? Such theoretical development might be the most interesting direction in the near-future cosmology.

An interesting direction of studies is checking for the present most popular structure formation assumption, the Gaussianity of the initial perturbations, the seeds of structure. It is the simplest physical assumption, and seems to hold in general. But it is an assumption, the minimum-knowledge starting point, and discovering departures from Gaussianity will force us to see much more interesting universes. The search for non-Gaussianity has begun, but it is a difficult task; there is only one Gaussianity hypothesis, but an infinite number of non-Gaussian possibilities. Presently, these are described by simple deviations; the right way would be to formulate a physical mechanism first and to explore the resulting distribution after that.

In order to check our theories, we compare observations with numerical simulations and theory. We observe galaxies that are complex aggregates of matter, with properties defined by the initial conditions in the early Universe and by the complex history of their formation. So, we need to model well both the evolution of dark energy and dark matter, and the formation and evolution of galaxies. Both research directions show considerable progress, but their problems are different. The physics of our dark-matter Universes is relatively simple, but it is extremely difficult to model large volumes of space, populated by tiny dark matter particles. So, we always end up with either a non-sufficient mass resolution (the masses of particles in our models are too large), or with a too small volume (much smaller than the observed part of the Universe). Progress in computer hardware and software will help, but demands hard work.

Modelling galaxy formation is completely different. Here the main problem is the complex input physics, the sheer number of different physical processes forming a galaxy, determining its overall structure, the distribution of gas, stars and stellar clusters, star formation and evolution. The present progress suggests that galaxy evolution theory will in time get similar to the present well-developed theory of stellar evolution, but it will certainly take a long time.

New instruments

The most important drivers for cosmology are observations, and new knowledge comes from new telescopes. Most of important cosmological information has so far been obtained by using dedicated instruments that work only on one program (e.g., galaxy redshift catalogues, deep catalogues of far-away galaxies, special programs dedicated for observing the traces of baryonic oscillations). A good point of telescopes is that all the raw data is preserved, and can be used later. This opens way for unexpected discoveries, as, e.g., the recent discovery of two extra bright distant supernovae by the Pan-STARRS telescope that works for Earth watch, discovering Earth-approaching objects and following their paths in space.

Surveys use usually medium-sized telescopes. Large telescopes are used for crucial observations of very weak or very distant objects, or of very fast processes, and have very good angular resolution. There are several projects to build large telescopes – GMT (Giant Magellan Telescope), consisting of seven 8.4m segments, at Las Campanas, Chile, TMT (Thirty Meter Telescope) at Mauna Kea, Hawaii, and E_ELT (European Extremely Large Telescope) of European Southern Observatory, with the diameter of 42 m, at Cerro Armazones, Chile (*ESO E_ELT*). These telescopes will certainly bring real discoveries.

In addition to optical telescopes, important information comes from telescopes working in other wavelengths. The largest of these are radio telescopes, and the most promising among these is ALMA (the Atacama Large Millimeter/submillimeter Array) on the Atacama high desert plateau in Chile (*ESO ALMA*). It is an international effort, and the final configuration of the telescope will be 66 antennas, most of them with a diameter of 12 meters. It will be perfect for observing cool gas in nearby galaxies and far-away young galaxies in the Universe.

Important data is obtained by space missions – telescopes in space. These are not as large as the telescopes on the Earth, but can observe the Universe shining in the radiation that is absorbed by our atmosphere. Good examples are X-ray and gamma-ray telescopes, and special radio telescopes meant for registering very weak signals (the CMB).

As space missions take about 15 years to plan and build, there are always many projects in progress, proposed and accepted, or waiting for approval. A very near future (a couple of years) will bring us final results of the Planck mission that will give a precise picture of the CMB. Many astronomers are waiting for the launch of GAIA (*ESA GAIA*), a space mission that will map a billion of stars (one per cent of all the stars in the Galaxy), and that is hoped to greatly enhance our knowledge about the Galaxy.

A very important future space telescope will be JWST (the James Webb Space Telescope), the NASA successor to the Hubble telescope, a 6.5 diameter mirror infrared space observatory. As the Universe expands, the light from distant galaxies shifts from the optical wavelengths to the infrared, and far-away galaxies are best seen in these wavelengths. This telescope is in construction, the expected cost is four times of that planned originally, and there are fears that the project might be terminated. This would be a great waste of work, ideas, and hope.

The first space mission to measure the cosmic web is EUCLID. It has a small 1.2 meter telescope, but will able to observe galaxies out to redshift of 2, much deeper than in Earth-based surveys; it will also map dark matter by observing gravitational lensing of far-away galaxies. The decision about its fate will come soon.

New computational facilities

Modelling the Universe is not an easy task. The main difficulties are the huge volumes that have to be modelled (the largest structures in the Universe, superclusters of galaxies and huge voids, are very large, more than 100 Megaparsecs), and the large variation in scale of different objects (even giant galaxies are of the size of 20 kiloparsecs, 5000 times smaller than the largest structures). If we want (and we have to) include star formation, we have to consider the sizes of star-forming clouds, yet thousands of times smaller than galaxies. So the dynamical range in mass (cube of the size range) can reach about 20 decades, demanding millions of petabytes of memory and corresponding speeds that clearly surpass all our present and near-future computing possibilities.

The way we have chosen now is to cheat the nature, by designing clever algorithms and ignoring possible resolution errors. We will continue to do so, but dangers lie this way. For example, our algorithms have become so clever and so complex that only two-three program packages for modelling the Universe are in use now, and people building different universes are only fiddling with specific initial conditions. If there is a bug hidden in one of these complex codes, our model universes will be bugged, too.

As the contemporary supercomputers are expensive to build and to use, international collaborations have appeared to construct the best possible numerical Universes. The results are made freely accessible, and many PhD students observe numerical universes only (the best example are the three 'Millennium' simulations). Such numerical collaborations will certainly provide the best future model universes. As leading-edge computing is getting more complex, people with special skills can be more easily hired by collaborations.

Estonian perspective

As I wrote already, Estonian cosmology is a (pretty small) part of the world cosmology, and our hopes and ways to achieve these are the same as everywhere in the world. As our number is small, we have to rely more on collaboration than other groups elsewhere, but this is only good. The possibilities of collaboration and participation in international projects are getting better – as Estonia is half-way on joining ESA, we can now officially participate in ESA missions (the astrophysics group of Tartu Observatory is already participating in the GAIA mission). The Estonian roadmap for the development of science includes joining ESO (European Southern Observatory, the equivalent of the European Union for astronomy); if this will happen, we will get much easier access to large telescopes.

At the domestic front, we have to keep our cosmology group large enough to be effective. Most cosmological projects demand participation of several people with different skills; we have these at present, but have to ensure that for the future, too. A successful group must also have something that makes it different from other cosmological groups. Our trademark has been so far good knowledge of observational data together with good theoretical background. Knowing observational data nowadays is only the question of the amount of time dedicated to learning and work; we are planning to extend that to propose more programs on larger telescopes. Improving the theoretical background includes tighter collaboration with physicists and mathematicians, and keeping up simulation skills. The latter demands access to supercomputers, and testing of ideas on smaller but yet powerful computers. At the Estonian level, access to computers is good, and we have used supercomputers in collaboration with colleagues in Finland, Germany, and Spain.

The short-time perspectives for cosmology in Estonia are good. Together with the particle physicists in the NICPB we have just won financing for an Estonian centre of excellence 'Dark matter in (astro)particle physics and cosmology'; this will ensure our close cooperation in this new research direction.

Ensuring the influx of good students is also important; participating in teaching and advertising how fascinating cosmology is will, hopefully, ensure that.

REFERENCES

Past

Arnalte-Mur, P., et al. 2011. Wavelet detection of Baryon Acoustic Structures in the galaxy distribution. ArXiv e-prints 1101.1911.

Boggess, N. W., et al. 1992. The COBE mission – its design and performance two years after launch. The Astrophysical Journal, 397, 420-429.

Colless, M. M., et al. 2001. The 2dF Galaxy Redshift Survey: spectra and redshifts. Monthly Notices of the Royal Astronomical Society, 328, 1039-1063.

Einasto, J., Jõeveer, M., Saar, E. 1980. Superclusters and galaxy formation. Nature, 283, 47-48.

Einasto, J., Kaasik, A., Saar, E. 1974. Dynamic evidence on massive coronas of galaxies. Nature, 250, 309-310.

Einasto, M., et al. 1997. The supercluster-void network I. The supercluster catalogue and large-scale distribution. Astronomy and Astrophysics, Supplement Series, 123, 119-133.

Eisenstein, D. J., et al. 2005. Detection of the baryon acoustic peak in the large-scale correlation function of SDSS Luminous Red Galaxies. The Astrophysical Journal, 633, 560-574.

Gramann, M. 1988. Structure and formation of superclusters. VIII. Evolution of structure in a model with cold dark matter and cosmological constant. Monthly Notices of the Royal Astronomical Society, 234, 569-582.

Hütsi, G. 2006. Acoustic oscillations in the SDSS DR4 luminous red galaxy sample power spectrum. Astronomy and Astrophysics, 449, 891-902.

Hütsi, G., Hektor, A., Raidal, M. 2009. Constraints on leptonically annihilating dark matter from reionization and extragalactic gamma background. Astronomy and Astrophysics, 505, 999-1005.

Jarosik, N., et al. 2011. Seven-year Wilkinson Microwave Anisortopy Probe (WMAP) observations: Sky maps, systematic errors, and basic results. The Astrophysical Journal, Supplement Series, 192, 14.

Ostriker, J. P., Peebles, P. J. E., Yahil, A. 1974. The size and mass of galaxies, and the mass of the universe. The Astrophysical Journal Letters, 193, L1-L4.

Penzias, A. A., Wilson R. W. 1965. A measurement of excess antenna temperature at 4080 Mc/s. The Astrophysical Journal, 142, 419-421.

Riess, A. G., et al. 1998. Observational evidence from supernovae for an accelerating universe and a cosmological constant. Astronomical Journal, 116, 1009-1038.

Rubin, V. C., Ford, W. K. Jr., Thonnard, N. 1980. Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 / R = 4 kpc/ to UGC 2885 /R = 122 kpc/. The Astrophysical Journal, 238, 471-487.

Tago, E., et al. 2010. Groups of galaxies in the SDSS Data Release 7. Flux- and volume-limited samples. Astronomy and Astrophysics, 514, A102.

Future

ESA GAIA: GAIA Mission.
http://sci.esa.int/science-e/www/area/index.cfm?fareaid=26

ESO ALMA: Atacama Large Millimeter/submillimeter Array.
http://www.eso.org/public/teles-instr/alma.html.

ESO E_ELT: The European Extremely Large Telescope.
http://www.eso.org/public/teles-instr/e-elt.html.

White Papers: US National Academy of Sciences. Astro 2010: The Astronomy and Astrophysics Decadal Survey. Science White Papers.
http://sites.nationalacademies.org/BPA/BPA_050603.

THE QUESTION OF MASS – ORIGIN, SEARCHES AND PROPERTIES

M. Raidal, E. Heinsalu, A. Hektor, K. Kannike, M. Müntel

INTRODUCTION

Precise determination of the energy budget of the Universe is one of the greatest scientific achievements of the last decade. According to WMAP only 4% of the mass in the Universe is in the known form of baryonic matter. What are the Dark Matter (DM) and Dark Energy (DE) that make up 20% and 76% of the Universe, respectively, remains a complete mystery. The evidence for their existence is purely gravitational. The question of the origin of mass is the most important fundamental scientific question to be answered in this decade.

In the Standard Model all known particles except neutrinos acquire masses via spontaneous electroweak symmetry breaking due to the Higgs mechanism. This mechanism is essentially relativistic Lorentz invariant non-Abelian realisation of the theory of superconductivity. It predicts the existence of new scalar spin-0 degree of freedom – the Higgs boson. However, no fundamental scalars have been discovered so far. Finding the Higgs boson and confirming the theoretical prediction of origin of mass in the standard model is one of the main purposes of the Large Hadron Collider (LHC) experiments at CERN. In addition, neutrinos are known to be massive. Because neutrino masses are tiny, it is believed that neutrino mass mechanism must be different from the Higgs mechanism. One possibility is that another scalar multiplet, triplet under the Standard Model gauge group, induces masses to neutrinos. This proposal must be tested at the LHC.

While the Higgs boson(s) can be found (or excluded) by the LHC experiments alone, discovering DM, studying its origin, interactions and distribution in the Universe requires a global multidisciplinary approach. At the time of writing this text scientific community is waiting for a new result from the DM direct detection experiment XENON100 that will check the present positive hints by DAMA, CoGeNT and CDMS II with unprecedented sensitivity. LHC experiments are just recording the very first collisions after the winter shut-down with the aim of direct discovery of DM in the form of missing energy. PAMELA satellite is just about to present their data on positron and electron fluxes in cosmic rays at unprecedented energies up to 300~GeV that could signal the annihilations or decays of DM particles in the halo of our Galaxy. Fermi satellite is continuously reporting on the measurements of gamma ray spectra that could originate from DM annihilations/decays in and outside our Galaxy. Atmospheric Cherenkov telescopes like HESS measure the high-energy cosmic ray spectra at high energies. New satellite based experiments like AMS II will further look for antimatter in cosmic rays. Planck mission has just published their first data and will soon report on their measurement of the Cosmic Microwave Background (CMB) that will test the standard cosmology, including the inflation paradigm and reheating of the Universe, with enormous precision. This is just an incomplete list of ongoing activities. Combining all those experimental efforts in cosmology and particle physics with computer modelling and with theoretical interpretation of the results will make the discovery of DM possible.

In this review we outline the research activities carried out in Estonia that attempt to answer those questions. We start by describing Estonian participation at CERN experiments and outline our major research projects at the LHC. After that we describe our studies of DM properties emphasising the global DM searches both on Earth and on satellite based experiments. We finish with describing some ideas of matter theory that connect fundamental physics of matter with the known concepts from material science.

CERN CMS EXPERIMENT AND COMPUTING CENTRE FOR NEW PHYSICS

Large Hadron Collider at CERN

The LHC at the European Organization of Nuclear Research (CERN) can accelerate protons to hitherto inaccessible energies. The protons, smashed together, produce spectacular bursts of particles from which researchers hope to find signals of new physics.

The main goal for the project is to try to discover the scalar Higgs boson predicted by the mechanism of electroweak symmetry breaking in the Standard Model of particle physics. However, everyone in the physical community hopes that LHC will find some unexpected physics beyond Standard Model.

The circumference of the collider is 27 km. The LHC is installed in the tunnel of the previous accelerator LEP (Large Electron Positron collider) and it began operation in October 2008 after more than a decade of construction. Its beam energy is designed to reach 7+7 TeV and luminosity up to $L = 10^{34}$ cm^{-2} s^{-1}, that is seven-fold increase in energy and a hundred-fold increase in integrated luminosity over the current hadron collider called Tevatron. There are four experiments located in the LHC tunnel: CMS (Compact Muon Solenoid), ATLAS (A Toroidal LHC ApparatuS), LHCb (Large Hadron Collider beauty) and ALICE (A Large Ion Collider Experiment). CMS and ATLAS are general multipurpose detectors for investigating electroweak symmetry breaking through Higgs boson, looking for phenomena beyond the Standard Model (e.g. supersymmetry), and to study the high-Q2 region in more detail. ALICE in designed to investigate heavy ion physics (quark-gluon plasma) and LHCb is dedicated to *b*-physics and *CP*-violation studies.

Compact Muon Solenoid

The CMS is a general-purpose particle detector with an emphasis on excellent detection of muons, the heavy cousins of electrons. Its total length is approximately 21 m, the diameter is 15 m and it weighs about 12 500 tons. The detector has almost full coverage of solid angle because of its cylindrical design and planar endcaps. The CMS detector consists of a number of different subdetectors, which are designed to identify different physical particles and particle bunches called jets. CMS detector has a superconducting solenoid providing a 4 T magnetic field parallel to the beam direction. The field bends the trajectories of charged particles in the transverse plane of the detector, enabling to measure particle momentum and charge. Inside the magnetic field are the inner tracking system and the calorimeter. The muon detectors are situated outside the coil, so the muons are exposed to a lower magnetic field. The magnetic flux of the solenoid is returned by a set of iron yokes. Between the yokes are the muon chambers. The inner tracking system has two different detectors: the silicon pixel detector and the silicon strip detector. These are used to identify bottom quarks by so called *b*-tagging and to reconstruct tracks and momenta. Next layers are calorimeters: the electromagnetic (ECAL) and hadronic (HCAL) calorimeters that measure the energies and positions of photons, electrons and hadrons respectively.

CMS Tier 2 Computing Centre at NICPB

Carrying out the research program of our group requires large computing power both for particle physics and for cosmological simulations. Our group is in possession of the largest computing centre in Estonia. The computing power available to particle physics and cosmology research consists of a 1400 core computing cluster and 750 TB of storage capacity. For the simulation of Dark Matter halo profiles and substructures we have in addition a shared memory system with 256GB of RAM and 32 compute cores as well as 16TB of near-storage. The computing centre is connected to GEANT network via 10 Gb/s optical link. The computing centre is part of Estonian scientific computing infrastructure that is one of the objects in Estonian Science Roadmap. The centre is compatible with WLCG and EGI Grid infrastructures. We are so called Tier 2 level computing centre at the CERN CMS experiment satisfying a set of certain stability and availability requirements. Our research group is pioneering a lot of Cloud computing concepts to further enhance the flexibility as well as resource utilisation efficiency. Therefore factorisable physics jobs can be run on additional tens of thousands of CPUs in the Grid depending on the problem at hand.

Access to a large scale computing facility allows for precision predictions for the future as more fine grain simulations can be carried out that previously were unattainable. In addition to this, the sheer scale of resources allows also the researchers to remain competitive in a front-line field, where paradigmatic shifts can occur in case a new experimental result signals of some significant new phenomena. As an example the computing resources available allowed the group members to recompute all Dark Matter decay channels in a large parameter space within days after the PAMELA positron anomaly was first reported giving the very first and most accurate fit in the context of the new results that has laid the groundwork for indirect detection fits in the past year.

Beyond Standard Model Physics at CMS: Doubly Charged Higgs Boson

The existence of non-zero neutrino masses is a firmly established signal of particle physics beyond the Standard Model. An explanation of the tiny neutrino masses is so called seesaw mechanism. The minimal see-saw model of type II is realised with one triplet scalar field with the $SU(2) \times U(1)$ that contains component dubbed the doubly charged Higgs boson. This particle carries double electric charge and decays to the same charged lepton pairs and therefore favours the seesaw mechanism of type II as the most promising neutrino mass mechanism that can directly be tested at the LHC experiments. The Yukawa coupling matrix of doubly charged Higgs is proportional to the light neutrino mass matrix, thus allowing testing of the neutrino mass mechanism by measuring the decay branching fractions of doubly charged Higgs. As a result the LHC experiments are able to reconstruct unknown neutrino parameters such as the absolute neutrino mass scale, the mass hierarchy and CP-violating phases that are not testable in the experiments of neutrino oscillation.

We have done first inclusive search for doubly charged Higgs in the CMS experiment based on the first available data from the experiment. We have studied both, the pair production process as well as the single production. So far, we can say we have excluded doubly charged Higgs below mass 150 GeV under the given assumptions. We will continue analysing all new data produced by the CMS detector. The exclusion mass will increase significantly as the collusion energy and rate of LHC are going to thrive during some next year.

PARTICLE DARK MATTER AND COSMOLOGY

More Dark Matter than the Ordinary Matter

When astronomers charted the speed of stars in galaxies *vs*. distance from galactic centre, they were surprised: the speed did not fall off, as predicted from the gravitational pull of visible matter – mainly stars – but stayed at a constant level. This implies that engulfing each galaxy there is a halo of dark (non-luminous) matter whose gravity keeps the stars speed constant. Later it became clear that Dark Matter is also involved in creation of large scale structures of the Universe: because it is not charged, it could form denser regions at an early time when ordinary matter was a hot plasma whose density fluctuations were blown off by radiation.

Combination of modelling the large scale structure and observations of the cosmic microwave background show that 72% of all matter is Dark Matter. However, the particle nature and exact origin of Dark Matter is not yet known.

Dark Matter is usually thought to be a thermal relic: in the early Universe, it was at thermal equilibrium. As the Universe expanded and temperature began to drop under the mass of Dark Matter particle, particle collisions were not energetic enough to produce new Dark Matter, but Dark Matter began to annihilate into lighter particles. It would have annihilated wholly but for further expansion of the Universe that diluted it. To be at thermal equilibrium with the rest of the matter, it must have some interaction besides gravitation. The most natural assumption is that it has weak interactions or can interact with the rest of the world via the Higgs particle. This puts the mass of Dark Matter roughly in the range from a few GeV to a TeV.

Detecting Dark Matter

Via weak scale interactions, Dark Matter could be directly detected in collisions with atomic nuclei. There are several detectors like CDMS, XENON or Edelweiss, placed under ground to minimise background from cosmic rays. The best sensitivity to this date was published recently with the results of XENON100. One of our aims is to study the implications of direct detection experiments for various models of Dark Matter, including the widely studied supersymmetric models like the Minimal Supersymmetric Standard Model, and several phenomenological models that could tie the properties of Dark Matter to other physical measurements like the anomalous magnetic moment of the muon or neutrino masses.

Why Is Dark Matter Stable?

To prevent Dark Matter from decaying, it must be stable. The simplest way to stabilise Dark Matter is to introduce mirror symmetry Z_2 with Dark Matter odd under it. The symmetry ensures that in each interaction, an even number of Dark Matter particles is destroyed or created. (Each interaction must be the same as its 'mirror image'. It takes an even number of odd particles to create a symmetric, even interaction.) This forbids Dark Matter decay – an interaction in which 'one' Dark Matter particle turns into Standard Model particles.

It is usual to add Z_2 symmetry by hand. Such global discrete symmetries, however, are broken by quantum gravity. The breaking is tiny, but large enough to cause Dark Matter to decay in contradiction with observations. A better way is to start with a local symmetry group like $U(1)_X$ – the symmetry of a circle on the plane – that defines a force similar to electromagnetism. If broken by a scalar whose X-charge is N, the Z_N subgroup of $U(1)$ is left intact. (The mirror symmetry Z_2 is the simplest such group; Z_3 is the rotation symmetry of equilateral triangle, Z_4 the rotation symmetry of square, etc.)

Dark Matter from Grand Unified Theory

The $U(1)$ group whose breaking gives the Z_2 parity can be embedded in a Grand Unified Theory (GUT) group along with the Standard Model gauge symmetry. One of the most plausible candidates for GUT is $SO(10)$, rotation group in ten dimensions. $SO(10)$ is especially plausible because one 16 multiplet of it contains just the Standard Model fermions plus right handed heavy neutrinos. The latter are needed to explain – via the seesaw mechanism – why the light neutrinos are indeed so light. Amazingly, the charges of these particles under the $U(1)_X$ are just right to make stable Dark Matter consisting of scalar particles that are odd under the new parity. In a series of papers, we studied dark sector that comprises a complex singlet and an 'inert' weak doublet belongs to a scalar 16 representation of $SO(10)$. These particles are the scalar analogues of the light left handed neutrino and the heavy right handed neutrino! There is a relatively strong interaction with the Higgs boson that would help to see Dark Matter through a 'Higgs portal' in detectors or make it at the LHC collider. There are also scalar analogues of quarks that could form so-called *R*-hadrons that give a strong signal in the LHC.

The extra scalar doublet, with the help of the 'Higgs portal', can break the electroweak symmetry and give all the particles their mass. This would help to explain why the Standard model Higgs mass and vacuum density are so low compared to the huge GUT energy scale.

Due to a residual symmetry, the mass of the next-to-lightest dark sector particle would be close to Dark Matter mass. This makes the NL particle more stable: if produced in the LHC, it can get away and give a distinct signal when it decays a few millimetres or centimetres away from the beam.

Non-Abelian Dark Matter

A simple mirror symmetry is certainly not the only possibility. Dark matter may be stabilised in more complex ways, by the rotation symmetry of an equilateral triangle or square, that is, Z_3 or Z_4. All Z_N symmetries have one common property: they are Abelian, meaning that if we consider two rotations on the plane, it does not matter in which order the rotations are made.

There are more complicated symmetries. Doing a half-turn with cube about the x-axis and then about the y-axis ends the cube in a different position than first doing a rotation about the y-axis and then about the x-axis. Such symmetries are called non-Abelian, and include the symmetries of the perfect solids (the tetrahedron, cube, octahedron, icosahedron and dodecahedron) and many more, in fact most of the possible symmetry groups.

Non-Abelian symmetries can be used to stabilise Dark Matter just like the Abelian ones, but there are only a few papers that use them. It is a brand new field waiting for the plough.

In fact, particle physicists are already using non-Abelian symmetries to a different purpose – in trying to explain the masses and mixing angles of different generations or flavours of leptons and quarks. An interesting possibility is to use one non-Abelian symmetry for both things.

It has been shown that the flavour symmetry must necessarily be broken. In this case Dark Matter decays – it is possible, granted that its half-life is about 10 billion years. But if one wants to explain the excess of cosmic ray positrons seen by PAMELA with Dark Matter, its lifetime must in fact be much longer, in the order of 10^{18} years.

Though we do not know the new forces which mediate such Dark Matter decay, we can parameterise our ignorance in terms of effective operators: we can write down all interactions that involve a Dark Matter particle and two, three, … various Standard Model particles. The more particles there are in the final state, the more suppressed the operator and the longer the lifetime of Dark Matter.

To explain the PAMELA result, it is enough to forbid decays up to effective operators of mass dimension six that can be done with a discrete group large enough. PAMELA requires Dark Matter to decay into leptons, not quarks. The discrete symmetry group – if, at the same time it is the flavour group – can indeed forbid decays into quarks and allow decays into leptons. (It has been shown that this requires use of a non-Abelian group.)

In our future research, we plan to study flavour groups, and their high energy origins, together with what the need to stabilise Dark Matter implies for flavour physics.

Discrete symmetry groups can arise from breaking continuous symmetry groups. The rotation symmetries of the tetrahedron or cube are obviously among the rotation group of the sphere $SO(3)$. Thus if there is a new force at high energies that turns particles of one generation into particles of another generation, its low energy remnant could be the large mixing angles in the neutrino mass matrix. Another possible continuous group in which they could be embedded is the special unitary group $SU(3)$, a kind of rotation group for complex valued vectors.

Non-Abelian Symmetries from Extra Dimensions

In fact there is another way, besides breaking gauge groups – orbifolds. String theory predicts that in addition to the three large dimensions that we see around us, space has several more dimensions that are rolled up and too small to see at low energies. But extra dimensions in fact do not presume string theory. In any case, their symmetry can give rise to many of the same discrete groups as breaking continuous groups. I this case, particles of different generations are literally situated in different places in the extra dimensions and changes between generations happen with rotation or reflection in the rolled-up dimensions.

While discrete symmetries can explain neutrino mixing very well, it is hard to put quarks in the same framework. The largest angle in the quark mixing matrix, the Cabibbo angle, is just over 13 degrees. An intriguing observation that could throw light on quark mixing is so-called quark-lepton complementarity: the sum of the Cabibbo angle and the corresponding neutrino mixing angle is very close to half the right angle. The same applies to the other angles. It may be just a numerical coincidence, but on the other hand, it may mean that the small Cabibbo angle is in fact the difference of two large angles.

LOOKING FOR DARK MATTER IN COMBINED EXPERIMENTAL DATA

Despite all the progress, to this date clear evidences for Dark Matter are based only on its gravitational interactions. Those evidences do not help to determine the composition of Dark Matter. There are three techniques to identify Dark Matter: (i) direct detection, (ii) indirect detection, and (iii) to produce them at particle colliders. Direct detection method is based on the idea that Dark Matter can rarely scatter off atomic nuclei of normal matter.

Indirect Detection of Dark Matter

Indirect detection is motivated by idea that if Dark Matter is produced by thermal freeze-out in the early Universe, it should annihilate to Standard Model particles in the present Universe. The annihilation products may be detected, but it is clear that the Dark Matter annihilation is very suppressed due to very weak coupling between the Dark Matter and Standard Model particles. If Dark Matter is composed of WIMPs it offers exciting possibilities for indirect detection. WIMPs are gathered gravitationally into the centres of galactic clusters, galaxies, the Sun, and Earth. In those regions with higher density of Dark Matter we can more easily detect annihilation signal and hopefully the signal can be distinguished from ordinary astrophysical processes.

The best indirect signatures of the Dark Matter annihilation are high-energy neutrinos, positrons and antiprotons because there are no other known intensive astrophysical sources of them. Gamma-ray signal is also excellent source of information as gamma-rays propagate directly and with low absorption in cosmic environment. Gamma-ray and charged cosmic-ray telescopes can easily pick up the unusual radiation of positrons, antiprotons and gamma-rays. The charged particles and gamma-rays from Dark Matter annihilation are partially absorbed in the environment of the Universe, which causes the reionisation and heating of the hydrogen and helium the environment is mainly comprised of. The reionisation and heating affects the Cosmic Microwave Background (CMB) and it can be measured with high accuracy.

If the mass of Dark Matter particle is below to the collision energy of a collider and it has weak interaction with Standard Model particles, it may be produced at particle colliders. Although Dark Matter particle is undetectable for a particle detector at collider the production of Dark Matter particle is typically accompanied by related production mechanisms. Thus the production events of Dark Matter particle can be observed by measuring the missing transverse momentum of collisions.

Satellite Experiments

Currently there are many experiments providing us with the probable indirect signatures of Dark Matter annihilation or decay in our Galaxy or in the whole Universe. We will list here some most famous ones: PAMELA (Payload for Antimatter Matter Exploration and Light-nuclei Astrophysics), Fermi LAT (Large Area Telescope), HESS (High Energy Stereoscopic System) and MAGIC. PAMELA is a satellite experiment to measure energy and types of high energy cosmic charged particles in the energy region from 1 to 200 GeV. Femi LAT is a satellite based gamma-ray telescope designed to measure gamma-rays from 100 MeV to 100 GeV. It can also measure electrons and positrons with similar energy. HESS and MAGIC are imaging atmospheric Cherenkov telescopes measuring the atmospheric Cherenkov traces of cosmic high energy particles. They can cover energy region from hundreds of GeV to tens of TeV. There are many smaller balloon-based experiments to measure charged cosmic rays in the upper atmosphere, e.g. ATIC, PPB-BETS. In recent years PAMELA and Fermi satellite measurements of the cosmic ray spectra have created a boom in indirect searches for Dark Matter annihilations/decays. The group members and their collaborators have played a major role in the analyses of this data (see the publications below). Our approach to Dark Matter annihilations/decays has been model independent studying all possible final states in the processes of Dark Matter annihilation/decay to Standard Model particles. This approach is very computing-power consuming and explicitly demonstrates the usefulness of applying HEP tools to solving astroparticle physics problems. Our major results in astroparticle physics have been combined; the tools for those analyses, simulated spectra and fits to the results are made available in the website http://www.marcocirelli.net/PPPC4dark matterID.html.

Combining Experimental Results

Computer simulations of the LSS formation, distribution and profiles of Dark Matter haloes are continuously getting more and more precise. This is just an incomplete list of ongoing activities. Combining all those experimental efforts in cosmology and particle physics with computer modelling and with theoretical interpretation of the results will make the discovery of Dark Matter possible.

The combined analyses of all the available data can answer the open questions: what is the nature of Dark Matter (scalar, Majorana or Dirac fermion?), what is its mass, what are the Dark Matter self-interactions and the interactions with matter, how is Dark Matter produced (thermal or non-thermal relic, symmetric or asymmetric?), how did Dark Matter influence the history of the Universe, how is Dark Matter distributed in the large and small scales of the Universe today etc.

Concrete details of our analyses as well as the expected results depend on the results of those experiments. If Dark Matter particles will be directly observed by XENON100, we derive the preferred Dark Matter mass and interaction strength with matter. From that information we get direct confirmation on the Dark Matter density and velocity in local scales that test the Dark Matter halo models of our Galaxy. We test the complementarity and consistency of this result with other experiments (DAMA, CoGeNT, CDMS) and predict Dark Matter observables at LHC. We are also able to predict cosmological fluxes of gamma rays, positrons and antiprotons depending on theoretical model and the production mechanism of Dark Matter. The measured cosmic ray spectra and the CMS measurements will allow us to test Dark Matter distribution in large cosmological scales. Of course, if Dark Matter will not directly discovered we can only study the complementarity of different observables and to derive new constraints on Dark Matter properties.

Theoretical Models of Dark Matter

The above described results and predictions depend strongly on the Dark Matter theory. Therefore we do not restrict ourselves to one particular scenario of Dark Matter but shall be as general as possible. We work on Dark Matter properties both model independently as well as study the most promising Dark Matter models. The former approach requires mastering HEP and cosmology tools and requires huge computing power. We do possess those skills and resources. In the latter case we study the most popular supersymmetric scenarios like CMSSM, NMSSM and also other popular possibilities like the scalar Dark Matter models, singlet fermion models etc. In our research we consider different Dark Matter halo profiles, distributions and substructures in order to test the cosmological observables. In particular the Dark Matter substructures may completely change the present interpretation of the Dark Matter indirect detection results. In this research computer simulations of the structure of our Galaxy will be of crucial importance. We consider different scenarios of Dark Matter generation (thermal and non-thermal relic abundance, Dark Matter due to asymmetry etc). The results of this research should single out the realistic Dark Matter scenarios and to rule out others.

Dark Matter in Structure Formation

An important aspect of our research is to search for the signals of Dark Matter in the CMS data. If Dark Matter is discovered in direct detection experiments, precise determination of its interactions, nature and mass can be done at LHC. This includes SUSY scenarios but also more exotic Dark Matter signatures like the displaced lepton vertex plus missing energy as predicted by the Grand Unified Theory of scalar Dark Matter. If, however, there will be no signals of Dark Matter from other experiments nor Dark Matter will be discovered by LHC, we obtain constraints on Dark Matter properties from our LHC searches.

We also focus on CMB tests of Dark Matter using the data by Planck mission including the polarisation measurements. This allows us to see the signals of Dark Matter annihilation in the distortion of CMB spectra. Using Planck data we plan to test the Dark Matter mass and annihilation cross section as well as the Dark Matter annihilation contribution to the reheating of the Universe. For the latter the polarisation measurements will be crucial. Dark matter determines both the formation and evolution of the LSS, of galaxies, their groups, clusters and superclusters. Different types of Dark Matter (e.g., cold, non-interacting particles, or weakly interacting Dark Matter particles) will lead to different cosmological structures, and so, studies of the properties of these structures can help us to determine the properties of Dark Matter. There are several known problems that may be solved by assuming special features of the Dark Matter, as the discrepancy between the observed and predicted abundance of satellites of galaxies, unusual galaxy formation times (early giant ellipticals), the possible non-Gaussianity of the LSS (observed early galaxy clusters and superclusters, and extremely rich present-time superclusters). To conclude, the LSS and galaxy studies are essential for determining the nature of Dark Matter (see Fig. bellow).

STOCHASTICS AS THE LINK BETWEEN SMALL AND LARGE

Diffusion of a particle in a magnetic field

Diffusion of charged particles in a magnetic field is of special relevance to many problems, ranging from technological applications to basic question of quantum mechanics and quantum Brownian motion. At a cosmological level, it is important to understand how the strong galactic magnetic fields affect the diffusion of elementary particles as well as that of macroscopic charged dust particles.

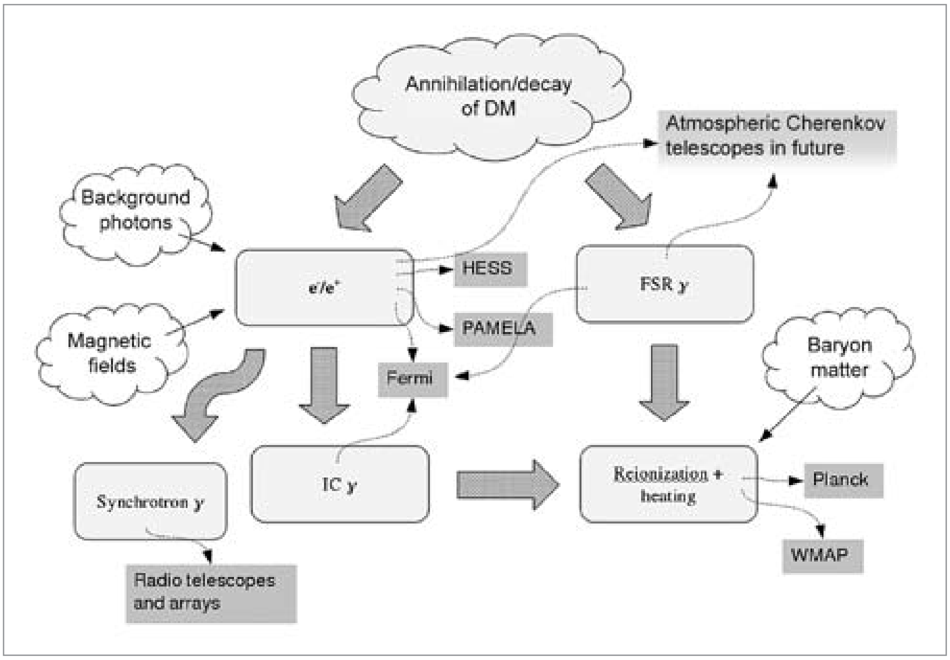


Figure 1.
The steps of the radiation transfer (bold arrows) from Dark Matter annihilation/decay (A/D) when only charged leptonic channels are allowed. In the first step, Dark Matter A/D produces charged leptons plus final state radiation photons. The heavier charged leptons decay quickly to electrons/positrons and neutrinos. In the next step, the electrons/positrons scatter from present photon background radiation (e.g. CMB) and/or present magnetic fields (e.g. the Galactic magnetic fields) producing inverse Compton photons and/or synchrotron radiation. Grey labels denote the experiments, which can measure the noticed radiation (dashed arrows).

We have studied classical as well quantum diffusion in the presence of a constant magnetic field. The effect of a magnetic field is multi-fold. Some of the results we found, in agreement with previous works, show that a magnetic field typically suppresses the diffusivity of particles. However, it also introduces unexpected correlations in phase space and can also partially inhibit decoherence process in a quantum system.

Quantum diffusion and stochastic mass

Quantum diffusion, quantum transport in stochastic potentials, describes a number of processes in Nature, from electron conductivity to light propagation in opaque environment. Interestingly, there are only few attempts to use the mathematical formalism of quantum diffusion developed in solid state physics for theoretical particle physics. For example, at low energy a diffusing massless particle acquires effective mass, 'stochastic effective mass' (SEM). In some extent, it can be considered as an alternative of the Higgs mechanism (aka mass from symmetry breaking). The LHC experiment at CERN is shedding light on the mass generating mechanism in some next years and if it is not the Higgs mechanism, one seriously needs an alternative. On the other hand, the SEM type mass mechanism can have wide set of applications in solid state physics. We are looking for an application of SEM in graphene. SEM allows to tune massless charge carriers in graphene to massive ones. Theoretically, one can drive conductivity or create suitable chemical potentials for charge carriers by tuning SEM of charge carriers in graphene.

Kramers' problem

The Kramer's problem of the escape of a particle over a potential barrier in the presence of thermal fluctuations is a basic problem of physics underlying the dynamics of different peculiar phenomena.

Examples of such phenomena are 'Stochastic Resonance', the unexpected appearance of a resonance as a function of noise level in the response of a nonlinear bistable system to an external periodic force, or the 'Ratchet Effect', i.e. the production of directed motion of a system along a periodic potential in a well-defined direction in the absence of any external bias.

The Kramer's problem, also due to its close analogy with the quantum-mechanical tunnel effect, is related to fundamental problems such as phase transitions of various types, ranging from nucleation phenomena in condensed matter to the mass generation through symmetry breaking in gauge theories.

We study various topics related to Kramer's problem, i.e. some general aspects of Brownian dynamics in the presence of a potential well, and some more application-oriented problems. In particular we have shown that in order for stochastic resonance to take place, it is necessary that the particle is bounded by a hard potential, i.e. a potential growing at large distances faster than a quadratic potential. In fact, stochastic resonance has been always studied using a prototypical quartic potential; the results obtained provide an immediate explanation of different known findings, in which e.g. linearly growing potentials or potentials which do not grow with distance (such as periodic potential) do not seem to produce stochastic resonance.

In addition, we have studied how stochastic resonance can appear even in situations where the form of the external potential does not allow it. This is the case for instance of non-elementary systems, e.g. systems with internal degrees of freedom. In these cases stochastic resonance can take place and be clearly observed in the internal degrees of freedom, as long as they are suitably coupled to the external force.

Thermal diffusion in a tilted periodic potential

The dynamics of a Brownian particle diffusing in a periodic potential in the presence of an external force (tilted periodic potential) represents a universal model which can describe a surprisingly long list of different physical systems, ranging from Josephson junctions, to polymer translocation as well as mechanisms leading to the matter-antimatter asymmetry of the Universe.

We have studied diffusion in a tilted periodic potential in different regimes: namely in the regime of normal diffusion – e.g. random walks with Gaussian random jumps and waiting times; and in the anomalous sub-diffusive regime, in which particles are observed to diffuse slower than in normal diffusion and are characterised by a mean square displacement growing slower than linearly in time. We have investigated the general dependence as well as the appearance of various resonant behaviours – e.g. optimal transport (resonant mobility) or diffusion (resonant diffusivity) – as a function of the main system parameters.

Super-diffusion and Levy flights

Besides sub-diffusion, the term 'anomalous diffusion' also serves to indicate 'super-diffusion', with its characteristic mean square displacement growing in time faster than the first power of time. As in the case of sub-diffusion, it used to be considered (and named accordingly) 'anomalous' but it is now known to characterise phenomena taking place at all scales, from microscopic scales associated to diffusion processes of charged particle through a plasma (one of the first known examples of super-diffusion) to biological and even ecological scales. In one of its first known forms, super-diffusion and reaction-diffusion processes in the presence of anomalous diffusion have received a renewed interest due to their connection with the dynamics of the plasma state of matter characterising early cosmological evolution.

We have studied the main differences between (reaction-)diffusion processes associated to the evolution of systems composed of units undergoing normal or super-diffusion (i.e. a random walk or Levy flights) in the presence of a finite-range interaction.

Power law distribution

The general power laws appearing in the laws for the mean square displacement of an anomalous diffusion process represent just one example of a more general feature of any systems characterised by the lack of a typical scale associated to the system dynamics. In fact, power laws also appear in distributions in place of the 'canonical' Gaussian or exponential tails through a wide range of systems and phenomena, from the macroscopic scale of social phenomena down to the microscopic scale of elementary particle collisions.

We have studied the formation of a power law tail under general conditions and have proposed a simple universal mechanism for their appearance, based on the heterogeneity of the interacting units. For instance, in the case of scattering phenomena, we have discussed how power-law tails naturally appear in the energy distribution of collision processes between molecules of a heterogeneous fluid. This is expected to hold whenever different scattering processes take place between particles which differ from each other, e.g. in the number of internal degrees of freedom. In fact, transverse momentum distributions in high energy hadron-hadron collisions as well as fragment size distributions in high energy nuclear collisions (as well as fragmentation process of macroscopic drops of many substances) present some universal power law tails.

SOLID STATE PHYSICS

V. Hizhnyakov, J. Kikas, A. Lushchik

A review on solid state research at the Institute of Physics of the University of Tartu is given. Noted in short is the investigation carried through in the initial period after foundation of the solid state research group. Recent results both in the solid state theory and in the experimental investigation have been presented in a greater detail. Future directions have briefly been mentioned.

INTRODUCTION

Solid state physicists in Estonia are predominantly concentrated into Institute of Physics, University of Tartu. The keywords of Solid State Physics in the Institute of Physics are dielectrics, also optical properties and processes in broad sense of optics as interaction of matter with electromagnetic field in its whole range of frequencies from radio waves to gamma radiation. In addition to dielectrics strongly correlated systems (including high-temperature superconductors, magnetically ordered solids etc.) too have been investigated in recent years. Activity in this area during the last 15 years has partially relatively continuously grown out of the earlier studies and the problems they raised, but of course also new themes have emerged, frequently in connection with expanding possibilities for international collaboration.

Creation of laboratory of luminescence (1951) in Tartu under the Institute of Physics and Astronomy of the Estonian Academy of Sciences, founded by the then rector of the University of Tartu, Feodor Klement can be considered the beginning of modern Solid State Physics in Estonia. Long-time head of the laboratory Cheslav Lushchik came from the University of St. Petersburg (then named Leningrad). Almost at the same time Karl Rebane, also graduate of Leningrad University, began assembling a group of solid state theory. Study of luminescence started in the early 50's soon developed into solid state physics in broader sense.

First objects of study were doped (i.e. containing impurities) halide crystals. Impurities in solids work here as microscopic probes, optical spectra of which provide information on the structure of the main substance and processes thereat in the impurity site. Solid state theory group developed the quantum mechanical theory of the impurity centres strongly interacting with vibrations of atoms. Original series of experiments performed in Tartu which investigated radiation spectra, kinetics and polarisation over a wide temperature range (from ultra-low up to high temperatures) and in very strong magnetic field confirmed the theory.

Together with solid state theoreticians from St. Petersburg, analogy of optical spectra of impurity centres /molecules with the Mössbauer effect was pointed out. There is a so-called zero-phonon line, which corresponds to a pure electronic transition. The line is highly sensitive to the interactions in solids allowing for a number of novel applications of solid state spectroscopy. Theory of zero-phonon line was developed and implemented in spectral hole burning, which is one of top results of Estonian physics (the phenomenon was discovered in 1974). Zero-phonon line gave rise to a new branch of spectroscopy – single molecule spectroscopy, as well as space-time holography – a promising method for information processing, which was realised at the Institute of Physics.

A remarkable achievement of Tartu theorists was the theory of secondary radiation, which unified scattering of light and luminescence. The theory was experimentally confirmed in the Institute of Physics and elsewhere. The theory indicated the possibility of hot luminescence as a component of resonant secondary radiation. Hot luminescence in solids was successfully experimentally observed in Tartu (officially recorded in 1981 as discovery by K.Rebane, P.Saari and V.Hizhnyakov). In the framework of the theory of secondary radiation the transform method of resonant Raman scattering was developed. The method is an extension for inelastic processes of the main foundation of optics, the optical theorem. This method found international application in Raman spectroscopy.

The theory of resonant secondary radiation was extended to resonant scattering of gamma radiation and neutrons in crystals. It predicted the slowdown of radiation energy propagation caused by the conversion of gamma-quanta into slow-moving nuclear polaritons. Those are quasiparticles consisting of the collective nuclear excitation (nuclear exciton) and a small amount of resonant gamma radiation adjoined to it. The concept of nuclear polaritons was developed in Tartu and it also found its experimental confirmation here.

These works were the first to demonstrate a million-fold slowdown of movement of electromagnetic waves in the matter.

Ultrashort laser pulses provided a unique opportunity to explore ultrafast electronic and vibrational processes directly in real time. Theory of time-dependent secondary radiation predicted a compensation effect in time-dependent spectra. This effect enables to achieve maximal possible spectral resolution, which is determined by the time period from excitation of the system until registration of the emitted photon. This result shows how the uncertainty principle for energy and time actually works, which is of general importance in physics: The compensation effect was confirmed experimentally in Tartu by measurements of time-dependent Mössbauer spectra.

Structural phase transitions which alter the symmetry of the crystal have always been an important research topic of the laboratory of solid state theory of the Institute. The substances under study were ferroelectrics – crystals, in which below a certain temperature (Curie temperature) a spontaneous electric moment appears. These materials have found wide application in technical devices (capacitors, modulators, optical storage devices, etc.). Institute of Physics has developed vibronic theory of structural phase transitions that successfully described properties of ferroelectric materials Structural phase transitions can be influenced by impurities first locally then globally. The theory of local phase transition near low-symmetry impurities was also developed in Tartu.

Recent successes of Tartu solid state theorists include the discovery of percolative phase separation in high-temperature superconductors (in collaboration with Nobel prize winner A. Müller) (Kremer et al. 1992); multiband theory of superconductivity (Kristoffel et al. 2003, 2008), study of superconducting fluctuations (Örd et al. 2009), theory of the magnetic incommensurability in cuprates (Sherman, Schreiber 2008), analytical theory of vibrational solitons in crystals (Hizhnyakov et al. 2006a), prediction of local modes associated with vibrational solitons (Hizhnyakov et al. 2006b); modelling of defect formation and increasing of radiation stability of crystals (Haas et al. 2010), nonperturbative theory of multiphonon relaxation of strong local vibrations (Hizhnyakov, Tehver 2006; Hizhnyakov 2010), theory of electronic transition with nonlinear vibronic coupling (Hizhnyakov 1996, 1999), theory of quantum diffusion of vacancies in solid He (Hizhnyakov, Benedek 2005); discovery of the effect of the zero-point motion of atoms in optical spectrum of superfluid droplet of ^{4}He (Hizhnyakov et al. 2009).

High-temperature superconductivity

Solid state theorists of Estonia joined the investigation of high-temperature superconductivity immediately after the discovery of the phenomenon. The laboratory of solid state theory was working on three models.

Collaboration with University of Stuttgart began in the end of 1980s (during the high-temperature superconductivity boom) and resulted in 'model of percolative phase separation' in high temperature superconductors. According to this model, when doping antiferromagnetically structured copper oxides – the initial material of high temperature superconductivity – charge-holes are generated in CuO_2 planes. Small spin-ordered hole-cluster (spin-polarons/ferrons) is generated around the holes. Increasing the concentration of holes leads to overlap of clusters, forming a percolation net. Below the critical temperature T_c superconductivity is possible within the percolation net. According to this mechanism, a microscopic strongly inhomogeneous electronic state is generated. This is a significant difference from standard superconductivity mechanism, where the electronic state is homogeneous throughout the crystal/metal volume. These ideas encouraged conducting of a series of experiments. Hole-clusters and the respective phase separation were observed by different methods: measurements of magnetic resistance and conductivity, electronic and nuclear magnetic resonance, neutron scattering etc. The results were published as joint articles, in which the discoverer of high temperature superconductivity K.A. Müller participated as a co-author. A new research direction was established and three international conferences under the common name "Phase separation in cuprate superconductors" (1992 and 1995 in Erice, Italy and 1993 in Cottbus, Germany) were devoted to it.

In the 'multiband scenario of superconductivity' a unified multi-gap ordering arises due to the electron-electron interaction between the actual electron bands. The different versions of interband mechanism were employed to explain the superconducting properties of cuprates, magnesium diboride, and also fullerene and graphite compounds.

The discovery of superconductivity in MgB_2 with unexpectedly high transition temperature of 39 K was reported in the beginning of the year 2001. In parallel with several other research groups, a relevant multi-channel model was developed in Tartu, taking into account intra- and interband interactions. The theory enabled to find superconductivity characteristics of MgB_2 in quantitative agreement with experiment and it explained the dependence of transition temperature on composition in the related compounds. A model with doping-created interband coupling, elaborated for cuprate superconductors, describes the behaviour of the properties of these systems in the phase diagrams.

In cooperation with universities of Chemnitz and Krasnoyarsk the strongly correlated systems were studied in the frames of 'tJ-model'. One actively discussed mechanism of superconductivity in formation of pairs of charge carriers (pairing) mediated by magnetic excitations. It was shown that in tJ-model, this mechanism does not lead to superconductivity. However, the interaction of holes and magnetic excitations causes a strong change in the spectrum of charge carriers, which significantly increases the transition temperature T_c when phonons are included into pairing. Another result of the interaction is a drastic change in the system at the 14–17% concentration of holes: effects of strong correlations disappear abruptly, and only phenomena characteristic to weak correlation remain. This transition has also been observed experimentally.

The effects of surface of strongly correlated systems were also studied. In particular, it was shown that the surface separates the crystal into two regions – near-boundary layers and the bulk – with widely different spin excitations. The existence of the depletion layer leads to considerable differences in spectral functions in the near-boundary region and in the bulk. This property of the crystals has promising perspectives for electronics.

A theory of the magnetic response in cuprate perovskites was developed, which takes into account all processes contributing to the susceptibility. The theory gives an explanation for the incommensurate response observed in p-type cuprates without additional assumptions of charge stripes or Fermi surface nesting. The obtained results allow one to connect the strengthening of the incommensurate elastic intensity by the applied magnetic field with the photoemission arcs in the pseudogap phase.

Nonlinear lattice dynamics

An important subject of research in modern physics is large-amplitude vibrations of atoms (ions) in defect-free crystals involving significant nonlinear effects – vibrational solitons which are also called breathers or intrinsic localised modes. The study of breathers began in collaboration with the institute of theoretical physics of Stuttgart University. Later the research was continued in collaboration with laboratory of atomic and solid state physics of Cornell University (in frames of "National Research Council Twinning Program"). Usually breathers are investigated by numerical computations. In Tartu a new method was developed that enables to study breathers analytically. The new phenomenon was predicted – coexistence of linear local modes with breathers. It was also shown that in metals the screening of atomic interaction by free electrons may enhance symmetric part of anharmonic forces. Therefore in metals may exist breathers with frequency above the phonon spectrum. Previously it was thought that the breather must have frequency smaller than the maximum frequency of phonons and therefore they cannot exist in solids, where there are no gaps in the phonon spectrum. It turned out that in metallic Ni and Nb where the gaps are absent, such breathers should occur.

Strong localised vibrations including breathers can relax (decay) due to quantum effects. Response for the decay interaction of the vibration with zero-point vibrations is often strong, so the perturbation theory used in quantum mechanics is not applicable. To describe the relaxation of local modes and breathers the nonperturbative quantum theory was developed. The theory predicted new phenomenon: an explosive growth of rate of multiphonon emission – a phonon burst – can occur at certain critical amplitude of the vibration. Existence of this phenomenon was experimentally confirmed by studies of hot luminescence of xenon crystal in Tartu.

A new numerical method was developed, which allows one to calculate motion of atoms and ions in solids taking simultaneously into account the short-range and long-range forces in the lattice. The method has been applied for modelling of creation of lattice defects by high-energy particles. The result was that long-range forces significantly facilitate formation of defects. This allows one to offer a new idea: turning off these forces in disperse substance makes possible to significantly increase the radiational stability of the material.

In 2003, together with Alex Sievers from Cornell University a school-conference on "Intrinsic Localised Modes and Discrete Breathers in Nonlinear Lattices" was organised in Erice Conference Centre in Sicily.

Vibronic transitions, quantum diffusion

One of important problems of solid state theory is multiphonon vibronic transitions in impurities. Characteristic of this problem is the participation of a very large number (of the order of the Avogadro number) of phonons in the transition. Change of the atomic bonds at the transition results in mixing of all phonons which strongly complicates the problem, but at the same time, leads to new important effects. Recently a new method was developed in Tartu which allows one to solve the problem. The important case of strong local softening of the phonon dynamics at the transition was also solved. The theory explains anomalous spectra of chlorin photoproducts in organic polymer matrices observed in Tartu.

A fundamentally important process in solids is diffusion of defects. According to classical physics, the diffusion is accelerating with rise of temperature. At low temperatures, properties of diffusion are radically changed: due to quantum laws diffusion accelerates with decreasing temperature, i.e. it behaves contrary to the classical law. Describing quantum diffusion, the change of phonon spectrum by the defect must be taken into account: the defect disturbs the phonon spectrum and this disturbance will be transmitted on with diffusion leaps. In framework of Cariplo Foundation project in cooperation with physicists from the University of Milano-Bicocca it was demonstrated that due to this disturbance quantum diffusion depends largely on the type of defects: quantum diffusion of vacancies is much slower than that of interstitials. Theory explains properties of quantum diffusion in helium crystals at temperature T ~ 1 K.

Quantum liquids ^{3}He and ^{4}He

At very low temperatures ~ 0.1 K ^{3}He and ^{4}He are respectively Fermi and Bose type quantum liquids with unique properties, being therefore the object of interest in modern physics. It has been demonstrated that these systems can be investigated by spectroscopic methods, especially via zero-phonon line, doping droplets of the quantum liquids with small molecules. In Tartu University in collaboration with University of Milano-Bicocca a theory of optical spectra of these systems has been proposed. It was found that optical spectrum of Na_2 complex on the surface of ^{4}He droplets contains direct information about the zero-point energy of neighbouring He atoms. That plays a decisive role in determining the unique properties of those quantum objects. Until that, the zero-point energy of atoms had been observed only indirectly. This work was highlighted in the journal Europhysics News.

The main distinguishing feature of ^{3}He quantum liquid is the existence of Fermi sea – the macroscopic subsystem of Fermi excitations. It has been shown that the particle-hole excitations of the Fermi sea lead to disappearance of the zero-phonon line in the spectrum already at zero temperature. Instead a strongly asymmetric narrow band appears in the spectrum, describing the electronic transitions assisted by the particle-hole excitations of the Fermi sea. The discovered phenomenon has some analogy with the well-known Kondo effect in metals. The proposed theory explains experimentally observed optical spectra of ^{3}He droplets doped by glyoxal molecules.

New 'phases'

One feature of the development of solid state physics during recent decades has been the discovery and studies of many new phases of condensed matter. Classical 'physics of crystals' has become a discipline in which the objects are glasses, incommensurate systems, quasicrystals, quantum liquids and several 'non-traditional' phases of solids.

Besides powerful bulk methods used in the investigation of solids – X-ray, neutron and Raman-scattering, magnetic resonance techniques – novel techniques such as single-molecule imaging (Plakhotnik, Palm 2001; Pärs et al. 2008), confocal Raman- and luminescence microscopy, photon correlation spectroscopy have been implemented (Basov et al. 2009). In recent years these methods have been complemented by atomic force microscopy and intensive numerical simulations (Brik et al. 2010; Romanov, Kolesnikova 2009). Pulsed laser deposition, sol-gel technology and focused ion beam lithography enable the fabrication of a large variety of meso- and nanosystems (Saal et al. 2011).

There is an important method of spectral hole burning developed in the Institute of Physics (the phenomenon was discovered in 1974). A rare combination of opportunities in global level – ultra-high spectral resolution + low temperatures (a few degrees above absolute zero) and high pressures (up to 10 000 atmosphere) – has enabled Institute of Physics to obtain unique information about changes of properties of substances under such conditions. High resolution laser spectroscopy studies of incommensurate molecular solid biphenyl initiated in cooperation with Technical University of Munich have shown that in such matter of both ordering typical to crystals and structural instability characteristic to glasses exist together .

Interesting results have been obtained in micro-physics of pressure-sensitive crystals and glasses. Sensitive spectroscopic techniques, such as the spectral hole burning, have enabled to show that already pressures of a few thousand atmospheres affect molecular movement in glasses (e.g. in polymer glass polystryrene) significantly, suppressing effects associated with anomalous (compared to crystals) degrees of freedom. In cooperation with solid state theoreticians of Institute of Physics a various observed phenomena were successfully explained from a common point of view complementing previously known model of microdynamics of glasses (by abandoning one assumption previously held as 'paradigmatic').

Radiation phenomena in wide gap functional materials

Inorganic wide-gap materials (WGM) are dielectrics with large width of the forbidden energy gap, E_g = 4–15 eV. Many of WGM are used nowadays for applications as phosphors in efficient light sources and displays, laser hosts, materials for optical components, safety inspection systems and nuclear energetics, highly sensitive fast detectors, dosimeters for medical and industrial purposes, etc. Under operating conditions WGM suffer the influence of different types of radiation: vacuum-ultraviolet (VUV) light, x- and γ-rays, electrons, protons, fast neutrons, swift ions. Interaction of WGM with radiation results in luminescence, creation of radiation defects and emission of electrons. Just last three phenomena were a major focus of interest for the Laboratory of Physics of Ionic Crystals for many years. The specific features of crystal lattice, behaviour of intrinsic electronic excitations – conduction electrons, valence holes and excitons – play a crucial role in the functioning of WGM. The growth methods of many alkali halide and some metal oxide (MgO, CaO, SrO) single crystals, both highly pure or doped with s^2-ions (Ga^+, Ge^+, In^+, Sn^{2+}, Tl^+, Pb^{2+}, Bi^{3+}) and other impurities, have been elaborated. The introduced impurity ions served as efficient luminescence centres or highly sensitive probes for the investigation of migration, self-trapping, multiplication of intrinsic electronic excitations as well as of the creation of radiation Frenkel defects (vacancies and interstitials). A complex of optical, EPR and thermoactivation spectroscopy methods allowing covering important action region on single crystals and thin films of 50 eV to 300 keV electrons and up to 22-eV photons (and further use of synchrotrons in Moscow, Lund and Hamburg in VUV-XUV spectral regions) has been elaborated as well.

The phenomenon of photon multiplication, when the excitation of a crystal with one VUV photon causes the luminescence with a quantum yield QY >1 was revealed in 1964 and stimulated later world-around the search for phosphors with the multiplication threshold of 7.2-8.4 eV, which can be used in environmental-benign (mercury-free) xenon-discharge luminescent tubes and display panels. Since 1964, the prospects of photon multiplication (photon cutting) have been analysed in Tartu for some hundreds of WGM (in 1997–2001 in cooperation with OSRAM Sylvania): binary metal oxides, complex tungstates, silicates, phosphates, sulphates aluminates, etc. Some application tasks were solved using Y_2O_3, YVO_4 and $BaMgAl_{10}O_{17}$ phosphors.

VUV radiation initially causes the formation of electronic excitations in WGM. Following fast relaxation, these excitations undergo both radiative and non-radiative decay (various types of luminescence) and nonradiative decay with heat release (phonons) or the creation of nanosize defects of a crystal lattice. In metals defects are predominantly formed by the creation of Frenkel defects at collisions of incident particles with the atom nuclei of a crystal. In WGM, besides this universal for solids impact (knock-out) mechanism, defects are created due to the decay of excitons or recombination between electrons and holes, formed during irradiation (non-impact mechanisms). Based on the EPR discovery of H interstitials (a halogen interstitial atom in a form of a X_2^- dihalide molecule situated at one anion site) by W. Känzig and the 'creation spectra' of F centres (Farbzentrum, an electron in the field of an anion vacancy) and H_A centres (an H centre nearby an impurity cation) measured by Estonian and Latvian physicists, it became clear in the middle of the 1960-ies that pairs of Frenkel defects in alkali halides are formed at the decay of self-trapped excitons or the recombination of electrons with self-trapped holes. The creation efficiency of Frenkel defects via these nonimpact mechanisms is hundreds times as high as at the irradiation by fast neutrons. In contrast to alkali halides with low formation energy of a Frenkel pair ($E_{FD} < E_g$), an inequality $E_{FD} > E_g$ is valid for a majority of metal oxides, where the contribution of the exciton decay or cold (after vibronic relaxation) electron-hole recombination to radiation damage is negligible.

Some recent important results obtained in close cooperation with many other research centres can be divided into several groups as follows (see selected papers Lushchik A. et al. 1996, 2006, 2007, 2008, 2011; Feldbach et al. 1997; Kirm et al. 1999, 2007, 2008, 2009; Lushchik Ch. et al. 2001; Dolgov et al. 2003; Nakonechnyi et al. 2006; Krasnikov et al. 2007; Babin et al. 2011).

Radiation effects in binary and complex metal oxides

(Dolgov et al. 2003; Lushchik et al. 2006, 2008, 2011; Babin et al. 2011)

According to our experimental recent data, the nonimpact creation mechanisms should be taken into account even in binary and complex metal oxides with $E_{FD} > E_g$ (MgO, Al_2O_3, SiO_2, $YAlO_3$, $Lu_3Al_5O_{12}$, etc.) under their irradiation with ~GeV swift heavy ions providing extremely high density of electronic excitations in ion tracks. It is experimentally proved that the efficiency of this nonimpact mechanism can be controlled, to some extent, either by a removal of some "unwanted" defects or by the introduction of certain luminescent ions that partly intercept the energy of hot electrons thus decreasing the number of their recombinations with creation of defects.

In MgO crystals (E_g = 7.85 eV) neither excitons nor electrons and holes undergo self-trapping in regular lattice regions, all these electronic excitations possess high mobility even at helium temperatures. It is shown that some impurity ions (Be^{2+}, Ca^{2+}, Al^{3+}-v_c-Al^{3+}) and as-grown structural defects (divacancies) serve as hole traps and have a large cross-section for the recombination with hot or cold conduction electrons. However, the energy released at such recombination in the bulk does not exceed E_{FD} under conditions of low-dense irradiation. The creation of F and F^+ centres (two or one electron in the field of an oxygen vacancy) and oxygen interstitials in the bulk has been detected only at the irradiation of MgO with swift heavy ions, i.e. under condition of extremely high density of electronic excitations. In order to change dielectric and electron emission characteristics, thin MgO films for plasma display panels were prepared varying concentration of certain impurity ions and using different methods. The project was performed in cooperation with SAMSUNG.

A complex study of electronic excitations has been performed in wide gap binary and complex metal oxides with varying number of ions per unit cell (from two in MgO to 160 in $LuAl_3O_{12}$, LuAG). In LuAG, oxygen hole polarons and excitons, the effective radius of which is smaller than the size of a unit cell, participate in the absorbed energy transfer to luminescence impurity centres via a hopping diffusion with an enormously low frequency factor as compared to that for free holes and excitons in high-symmetry MgO. At the irradiation of complex polycation oxides, a slow hopping diffusion of 'heavy' electronic excitations impedes their movement to the surface and subsequent nonradiative decay into structural defects. A similar situation is realised in highly radiation-resistant $MgAl_2O_4$ and especially $Y_3Al_5O_{12}$ with complex unit cells. In YAG, a slow hopping diffusion of small-radius polarons was verified by measuring the spectra of two-photon absorption.

Electronic excitations and Frenkel defects in metal fluorides

(Lushchik et al. 2006, 2007, 2008; Nakonechnyi et al. 2006)

Binary and more complex metal fluorides (LiF, CaF_2, BaF_2, LuF_3, GdF_3, YF_3, $LiLuF_4$, etc.) are widely used for many technical applications. In recent years, some optical materials containing rare earth ions (lanthanides) gained a special interest, as new VUV emitters based on d→f transitions in rare earth ions for the usage in ionisation chambers for high-energy particles. Luminescence in deep VUV region (~10 eV), which is due to interconfigurational $4f^{n-1}5d \rightarrow 4f^n$ transitions in Gd^{3+} and Lu^{3+} ions, has been detected, identified and studied in detail with high spectral and time resolution for a number of metal fluoride crystals. It was found that Gd^{3+} ions emit at $T < 100$ K nanosecond luminescence due to spin-allowed transitions, while Lu^{3+} ions emit at low temperature microsecond luminescence connected with spin-forbidden $5d$-$4f$ transitions. However, with increasing temperature, an admixture of spin-allowed luminescence also shows up.

LiF single crystals, highly pure or doped with luminescent impurities, are widely used as isotropic optical windows and personal dosimeters of γ-rays. It is shown that the formation energy of nonrelaxed one-halide excitons is very high (~13.6 eV), while two-halide self-trapped excitons possess significantly lower energy (emission at ~3.4 eV). So, the relaxation and decay of anion excitons in LiF tentatively occurs not only with the creation of both, anion and cation. The decay of a cation exciton (~62 eV) at 6 K causes the creation of a group (triplet) of spatially correlated defects. At the irradiation of LiF with 10-MeV heavy ions, the absorbed energy is approximately equally spent on the impact displacement of ions into interstitial sites due to their elastic collisions with high-energy particles and on the formation of electronic excitations (i.e. ionisation losses). Extremely high density of electronic excitations takes place in the tracks of 2-GeV gold ions, more than 99% energy of which goes to ionisation losses. The nonimpact mechanisms (e.g., hot electron-hole recombination) of defect creation are the dominant ones under such irradiation conditions.

Multiplication of electronic excitations in WGM
(Lushchik A. et al. 1996, 2011; Feldbach et al. 1997; Kirm et al. 1999; Lushchik Ch. et al. 2001)

The processes of multiplication of electronic excitations, when one exciting photon of 10-40 eV forms simultaneously up to 3-4 spatially close excitations have been thoroughly investigated in many wide gap single crystals. An enhanced local density of electronic excitations formed via the multiplication process causes the changes of luminescence quantum efficiency as well as facilitates the subsequent creation of groups of spatially correlated defects. The measurements of the excitation spectra for various intrinsic or impurity emissions in WGM additionally exposed to uniaxial stress or an electric field allowed us to separate clearly three main mechanisms of multiplication in WGM with relatively narrow valence bands. A hot conduction electron, formed at the absorption of an exciting photon, is able to create a secondary e-h pair, a secondary anion exciton or to provide a direct excitation of an impurity ion. The latter process is considered as the solid-state analogue of the Franck-Hertz effect in gases. The three mechanisms of electronic excitations multiplication differ also in the kinetics of fast and recombination emissions, appearance of tunnel recombination etc. In WGM with wide valence bands ($E_v > E_g$) Auger processes with the participation of hot electrons and hot valence hole are even more complicated. In metal oxides (e.g., Al_2O_3), the exciting photon of 30–33 eV that causes the ionisation of the $2s^2$ shell of an oxygen ions and the formation of a $3p$ conduction e is able to form even three electron-hole pairs.

It is worth noting that one more possibility to decrease the threshold energy for photon multiplication is connected with the use of rare-earth trivalent terbium ions. At relatively high concentration, the existence of two- or even three-terbium containing complex centres in $CaSO_4$ facilitates the resonant energy transfer between Tb^{3+} with identical energy structure resulting in quantum yield of impurity luminescence above unity at the matrix excitation by 6-8 eV photons.

Ultrafast phenomena and fast scintillators

(Krasnikov et al. 2007; Kirm et al. 2008, 2009; Babin et al. 2011)

Creation, relaxation dynamics and mutual interaction of excitons were studied in $CdWO_4$ and $CaWO_4$ scintillators under excitation by sub-picosecond laser pulses. Using the method of interferometry with 100-fs time resolution a very slow exciton relaxation in the range of tens picoseconds was revealed. Strong dependence of the relaxation time on temperature and excitation density indicates to substantial difference in effective masses of electrons and holes. Excitation by powerful VUV (free electron laser, high-order harmonics of a Ti-sapphire laser) or UV (OPO Topaz) pulses with duration 15-100 fs results in creation of small-radius Frenkel excitons at densities 10^{18}-10^{21} cm^{-3}, correspondingly, in recombination or resonant process. At such densities, which are comparable to the density of electronic excitations created in tracks of ionising radiation, the exciton emission undergoes partial quenching and its decay kinetics becomes non-exponential due to the dipole-dipole interaction of excitons. A theoretical model of the underlying processes was developed. The exciton-exciton interaction was shown to play a crucial role in non-proportional response of scintillators whose operation is based on intrinsic luminescence. The experiments were performed at the Laser Centres in Saclay, France and Vilnius University, Lithuania and FLASH facility at DESY, Hamburg, Germany.

The function mechanisms of scintillators for various applications (high-energy physics, medical imaging, safety inspection, etc.) based on fast intrinsic and impurity emissions has been investigated in Tartu over many years. The prospects for use of two kinds of fast intrinsic emissions – crossluminescence (~ ns) and intraband luminescence (~100 ps) – in fast scintillation detectors have been analysed. These temperature-stable emissions have been investigated in detail under the excitation of many WGM by single nanosecond 300-keV electron pulses of Koval'chuyk-Mesyats type generator. A complex study has been concentrated on the increasing of scintillation efficiency, emission decay time shortening and the enhancement of radiation resistance of the materials. In cooperation with Czech, Italian, Russian, Ukraine, Hungary and German scientists the elaboration of novel fast scintillators and the improvement of the existing ones have been performed. The function mechanisms of new-generation scintillators based on Ce- and Pr-doped Y-Al, Lu-Al perovskites and garnets were clarified. The scintillation characteristics of these single crystals and crystalline films as well as of a number of tungstates (PWO is used in Large Hardon Collider in CERN) were studied and optimised. In collaboration with PHILIPS (EC project "STRING"), the systematic investigation on the dynamics and spectroscopy of radiative *d–f* electronic transitions was conducted as a function of excitation energy, temperature and dopant concentration for wide variety of host lattices for a new generation of PET scanners.

Nanomaterials

One new and successfully developing direction in physics of materials is the development of nanomaterials and methods for their characterisation. Scanning probe microscopy allows not only investigation of such materials, but it can also be a technological tool for creation and modification of nanostructures. On the other hand, development of the method itself requires solving of a number of problems in materials science. Notable results in this direction are materials (produced by sol-gel method) that enable combination of functionalities of tunneling microscope and near-field optical microscopy.

Scientists of the Institute of Physics are involved in nanomaterials research for the European Science Foundation "Nanotribology" program and the COST P13 program "MOLSIMU". Numerical simulation of atomic level processes is an important tool for investigating nanomaterials. In this area the activity of Estonian physicists has been significantly associated with international cooperation, which has provided access to necessary computing resources. Specific example is cooperation with Uppsala University in molecular-dynamical ion mobility simulation in polyethylene oxide doped with nanoparticles – a promising material for new generation of electrical accumulators.

FUTURE DIRECTIONS

Solid state physics and materials science will focus on developing new functional (nano)materials and methods for their formation and testing. A special attention will be paid to mesoscopic systems and phenomena which lie in between the macro world and atomic dimensions in order to 'bridge' the gap between the atomic and macro dimensions. Such spatial scale is a challenge for the current theories. On the other hand, the meso-range of dimensions and processes hides potential for numerous innovative approaches for new materials, sensorics and informatics.

In the theory of condensed matter, research will be continued in areas of nonlinear dynamics of macroscopic, nano- and mesosystems, strongly correlated solids with electronic and structural ordering, high-temperature superconductivity, nonlinear and singular optics, nonlinear plasmon-light interaction in metal-dielectric interface (plasmons are collective oscillations of electron density). Theoretical studies of critical phenomena and stochastic processes are related to these areas of study. Complex studies of meso- and nanosystems are planned for years 2011–2015 in framework of Centre of Excellence "Mesosystems: theory and applications" (project funded by the European Union through the European Regional Development Fund).

Expected results: development of methods of simulations of ultrafast nonlinear dynamics of mesosystems, development of new models of mesosystems, development of theory of nonlinear quantum optics and nonlinear plasmon-photonics on discrete photon level, prediction and study of new superconducting and magnetic mesosystems, and new materials stable against radiational damage. Further research together with the other groups includes the study of superconducting, magnetic and electronic mesoscale orderings; stochastic phenomena in multi-gap superconductors (Josephson junctions, vortex transport); phase transitions and nonlinear dynamics in spatially restricted systems; the development of new models of mesosystems. Properties of materials in extreme conditions (cryogenic and high temperatures, strong electric and magnetic fields, high pressures, very high excitation densities) are studied. Experimental studies are carried out in close collaboration with fundamental theoretical research in solid state theory.

In methods of synthesis, the focus is on atomic layer deposition, laser ablation and sol-gel technologies. New opportunities for nanostructuration of materials will open up with introduction of focused ion beam technique. As methods of analysis, modern methods of X-ray diffraction, scattering and reflection, electronography, optical (including laser-) microspectroscopy (including single molecule spectroscopy), also different methods of X-ray-, radio-, electron- and dielectric spectroscopy will be used and developed. Modern microscopy techniques and equipment for characterisation of materials are developed. Potential results of these studies are new materials and methods for sensorics, information and communications technology and other areas of applications.

REFERENCES

Babin, V., et al. 2011. Time-resolved spectroscopy of exciton-related states in single crystals and single crystalline films of $Lu_3Al_5O_{12}$ and $Lu_3Al_5O_{12}$:Ce. Physica Status Solidi B, 248, 6, 1505-1512.

Basov, A. et al. 2009. Spatial localization of Si-vacancy photoluminescent centres in a thin CVD nanodiamond film. Physica Status Solidi A: Applications and Materials Science, 206, 9, 2009-2011.

Brik, M. G., Sildos, I., Kiisk, V. 2010. First-principles calculations of optical and electronic properties of pure and Sm3+-doped TiO2. Physica B: Condensed Matter, 405, 10, 2450-2456.

Dolgov, S. A. et al. 2003. Electron paramagnetic resonance of the $[Be]^+$ centre in MgO:Be. Journal of Physics: Condensed Matter, 15, 10, 6871-6878.

Feldbach, E. et al. 1997. Direct excitation of impurity ions by hot photoelectrons in wide-gap crystals. Physical Review B, 56, 21, 13908-13915.

Haas, M. et al. 2010. Effects of long-range forces in nonlinear dynamics of crystals: creation of defects and self-localized vibrations. IOP Conf. Series: Materials Science and Engineering, 15, 012045.

Hizhnyakov, V. 1996. Relaxation jumps of strong vibration. Physical Review B, 53, 21, 13981-13984.

Hizhnyakov, V. 1999. Multiphonon anharmonic decay of a quantum mode. Europhysics Letters, 45, 4, 508-513.

Hizhnyakov, V. 2010. Zero-phonon line: effect of quadratic electron-phonon coupling. Chemical Physics Letters, 493, 191-194.

Hizhnyakov, V., Benedek, G. 2005. Quantum diffusion: effect of defect-localized phonon dynamics. European Physical Journal B, 43, 4, 431-438.

Hizhnyakov, V., Shelkan, A., Klopov, M. 2006a. Self-consistent theory of intrinsic localized modes: Application to monatomic chain . Physics Letters A, 357, 4-5, 393-396.

Hizhnyakov, V., Tehver, I. 2006. Optical transition in an impurity centre of a crystal: mixing of phonons. Chemical Physics Letters, 422, 4-6, 299-302.

Hizhnyakov, V., Tehver, I., Benedek, G. 2009. Theory of the optical spectrum of Na_2 on ^4He droplets: effect of the zero-point energy of the nearest atoms. European Physical Journal B, 70, 507-512.

Hizhnyakov, V. et al. 2006b. Linear local modes induced by intrinsic localized modes in a monoatomic chain. Physical Review B, 73, 224302-1-224302-5.

Kirm, M. et al. 1999. Self-trapping and multiplication of electronic excitations in Al_2O_3 and Al_2O_3:Sc crystals. Physical Review B, 60, 1, 502-510.

Kirm, M., et al. 2007. Vacuum ultraviolet 5d-4f luminescence of Gd^{3+} and Lu^{3+} ions in fluoride matrices. Physical Review B, 75, 7, 075111.

Kirm, M., et al. 2008. Behaviour of scintillators under XUV free electron laser radiation. Journal of Luminescence, 128, 5/6, 732-734.

Kirm, M., et al. 2009. Exciton-exciton interactions in $CdWO_4$ irradiated by intense femtosecond vacuum ultraviolet pulses. Physical Review B, 79, 23, 233103.

Krasnikov, A. et al. 2007. Localized excitons and their decay into electron and hole centres in the $PbWO_4$ single crystals grown by the Bridgman method. Journal of Physics: Condensed Matter, 19, 30, 306202.

Kremer, R. K., et al. 1992. Percolative phase-separation in La_2 CuO_4+ _d and La_2-x Sr_x CuO_4. Zeitschrift für Physik B: Condensed Matter, 86, 3, 319-324.

Kristoffel, N., Örd, T., Rägo, K. 2003. MgB2 two-gap superconductivity with intra- and interband couplings. Europhysics Letters, 61, 109-115.

Kristoffel, N., Rubin, P., Örd, T. 2008. Multiband model of cuprate superconductivity. International Journal of Modern Physics B, 22, 5299-5327.

Lushchik, A. et al. 1996. Secondary excitons in alkali halide crystals. Physical Review B, 53, 9, 5379-5387.

Lushchik, A. et al. 2006. Defect creation caused by the decay of cation excitons and hot electron-hole recombination in wide-gap dielectrics. Nuclear Instruments and Methods in Physics Research B, 250, 1-2, 330-336.

Lushchik, A., et al. 2007. Creation of nanosize defects in LiF crystals under 5 -and 10-MeV Au ion irradiation at room temperature. Physical Review B, 76, 5, 054114.

Lushchik, A. et al. 2008. Stabilization and annealing of interstitials formed by radiation in binary metal oxides and fluorides. Nuclear Instruments and Methods in Physics Research B, 266, 12-13, 2868-2871.

Lushchik, A., et al. 2011. Low-temperature study of electronic excitations in wide-gap materials using luminescent RE^{3+} and Cr^{3+} impurity ions. Optics and Spectroscopy, 111, 3, 434-440.

Lushchik, Ch. et al. 2001. Creation of F centres and multiplication of electronic excitations in $Na_6Al_6Si_6O_{24}(NaBr)_{2x}$ optical ceramics under VUV irradiation. Journal of Physics: Condensed Matter, 13, 27, 6133-6149.

Nakonechnyi, S. et al. 2006. Low-temperature excitonic, electron-hole and interstitial-vacancy processes in LiF single crystals. Journal of Physics: Condensed Matter, 18, 2, 379-394.

Örd, T., Rägo, K., Vargunin, A. 2009. Relaxation of superconducting fluctuations in a two-component scenario with intra-and interband pairings. Journal of Superconductivity and Novel Magnetism, 22, 85-88.

Pärs, M. et al. 2008. Room-temperature single-molecule fluorescence imaging for terrylene in biphenyl single crystals. Journal of Luminescence, 128, 5-6, 838-841.

Plakhotnik, T., Palm, V. 2001. Interferometric signatures of single molecules. Physical Review Letters, 87, 18, 183602-1/4.

Romanov, A. E., Kolesnikova, A. L. 2009. Application of disclination concept to solid structures. Progress in Materials Science, 54, 6, 740-769.

Saal, K. et al. 2011. Micro- and nanoscale structures by sol-gel processing. International Journal of Materials and Product Technology, 40, 2-14.

Sherman, A., Schreiber, M. 2008. Fluctuating charge density waves in the Hubbard model. Physical Review B, 77, 155117 .

ULTRATHIN FILMS FOR APPLICATIONS FROM NANOELECTRONICS TO CORROSION PROTECTION

J. Aarik, H. Alles, K. Kukli, V. Sammelselg

INTRODUCTION

Rapid development of microelectronic and information technology during recent decade has created a strong need for thin solid layers (thin films) with thicknesses down to a few nanometres. Field effect transistors (FETs), random access and flash memories, and devices with spin-sensitive tunnel junctions are just a few applications where ultrathin films with well-controlled properties are needed. In addition, using thin-film surface coatings one can increase the sensitivity of chemical sensors, modify the reflectivity and transmittance of optical elements, increase the conversion efficiency and durability of solar cells, prepare surfaces with self-cleaning performance, markedly increase the corrosion resistance of materials, etc.

Aggressive development of integrated circuits that has continued more than 40 years has resulted in several problems limiting future progress in this field. Complications faced are primarily due to the increase in the areal density and, correspondingly, the decrease in the lateral dimensions of the components. For instance, prototypes of silicon-based FETs with 6-8 nm channel length have been prepared (Wong, Iwai 2006). However, it is expected that the exponential progress of silicon-based semiconductor industry (well described by Moore's Law) is most likely going to end in this decade due to fundamental limitations set by quantum mechanics. Moreover, even without these limitations, reproducible production of the smallest prototype transistors in high volumes may turn out to be too complex. Along with shrinking dimensions of the transistors, their constituent materials have already been replaced by some alternatives. Polysilicon-based gates have been changed to metals, traditional gate dielectrics (SiONx) with high-permittivity (high-*k*) metal oxides, e.g. Hf-based mixed oxides, and the research is being conducted worldwide to find reliable substitutes for the silicon-based channels of FETs.

The most popular concepts of transistor architecture anticipate exploiting strained silicon-on-insulator, silicon-germanium, and also 3D channels built up of silicon blocks (so-called FinFET). The latest, 22 nm generation of processors is already based on 3D transistors[1] and the development is sup-

[1] http://newsroom.intel.com/community/intel_newsroom/blog/2011/05/04/intel-reinvents-transistors-using-new-3-d-structure

posed to continue. This would allow more efficient use of the channel area and faster switching. Another way to increase the channel mobility is the introduction of III-V materials such as GaN. Before the validation as production-ready processes, all these concepts require, however, solving of many problems connected to the growth of uniform films on 3D substrates, elimination of interfacial regions of poorly controlled properties between different material layers, finding inexpensive and reliable methods for robust and safe deposition of multilayer stacks in a common temperature window, etc.

Besides silicon, silicon-germanium alloys, III-V compounds and other related materials, carbon-based nanostructures are considered as a very prospective research field for materials science. A lot of studies have been devoted to the investigations on transistors, switches, memories and sensors based on carbon nanotubes (CNT). Carbon nanotube is a 3D structure consisting of rolled-up and closed sheet of carbon atoms. Due to its extremely high electronic conductivity, CNT might offer a prospective for creation of high-speed transistors and switches, when serving as a transistor channel instead of silicon. CNT is not the only choice among the 3D nanostructures, as e.g. nanorods of ZnO and other materials have also been studied as alternatives. This kind of nanostructures can be grown by chemical vapour deposition (CVD) methods. It is imperative, however, that before the actual exploiting the nanorods, nanowires, and nanotubes as base structures for electronic devices, one has first to learn to deposit them in an absolutely controlled way, i.e. exactly in the desired size and direction, just because these structures are hardly modifiable by post-deposition/lithographical means.

Actually the most promising material that could replace silicon in future nanoelectronic devices and possibly in several other applications is 'graphene'. Graphene, the first 2D crystal ever discovered (Novoselov et al. 2004), is a two-dimensional sheet of carbon atoms arranged in a hexagonal lattice. Graphene can also be imagined as a carbon nanotube cut and rolled-out on a planar surface. The studies of graphene belong currently to the hottest topics in physics due to unique electronic properties of this material.

The charge carriers, electrons and holes, have zero effective mass in graphene at low energies and hence, their behaviour can be described by the Dirac relativistic equation instead of usual Schrödinger equation. Graphene as a conductor has comparable or even better electronic mobility than carbon nanotubes. Furthermore, graphene possesses high current-carrying capabilities, transparency, thermal conductivity and mechanical stability. During recent years serious efforts have been made in order to learn to prepare high-quality graphene reliably and reproducibly on large areas. The most promising methods in that respect are the epitaxial growth of graphene on

SiC (de Heer et al. 2007) and CVD of graphene on transition metals like Ni (Kim et al. 2009) or Cu (Li et al. 2009). In a longer perspective, graphene-based applications can be anticipated in different fields from nanoelectronics to supercapacitors (Zhang et al. 2010), chemical sensors (Schedin et al. 2007) and may be even in anticorrosion coatings (Chen et al. 2011).

Development of the capacitive components of computer memories, such as dynamic random access memory (DRAM) cells, proceeds in a somewhat analogous way with that of FETs. Electrodes based on (poly)silicon have already been replaced by conducting nitrides (TiN) or are to be exchanged either by metals (e.g. Ru) or conductive metal oxides (e.g. RuO_2). The dielectric layers based on silicon oxynitride have been replaced by high-*k* metal oxides (HfO_2, ZrO_2, ZrO_2-Al_2O_3-ZrO_2). 3D cells of DRAMs are by now constructed on either trenched silicon wafers or columnar structures built up above the wafer level in order to increase the electrode area, capacitance of the cell and/or lateral density of the cells. Processes and novel materials enabling the deposition of more uniform metal layers as well as dielectrics with increased dielectric constants are actively searched for.

For preparation of thin and ultrathin layers of semiconductors, molecular beam epitaxy and different modifications of CVD methods are widely used. As mentioned already, the CVD methods can be applied also for preparations of nanotubes, nanorods and nanowires as well as graphene. In order to improve material quality deposited by CVD methods, plasma enhanced processes are frequently exploited. Plasma-enhanced CVD has been applied for deposition of dielectric films too. In addition, various types of ion sputtering and vacuum evaporation methods (including, e.g. laser ablation techniques) are used in electronic industry. However, all these methods have faced problems when coatings of uniform thickness should be deposited on 3D-processed surfaces.

Atomic layer deposition (ALD) is a method that allows deposition of ultrathin films with well-controlled thickness on highly profiled 3D solid surfaces as well. The ALD technique is based on saturative surface reactions of alternatively supplied precursors and was invented and initially developed already in 1960s and 1970s independently in Russia and Finland (see e.g. Puurunen 2005). The interest in ALD was limited till the end of the last century mainly due to relatively low deposition rate (typically 10-300 nm/h). The slowness of this method has gradually lost significance because of fast downscaling the component sizes in integrated circuits. After the focused attention of big electronic companies (Sharp, Philips, Samsung, IBM, Motorola, Freescale, Intel and others) to ALD in the end of 1990s the interest in this method increased significantly worldwide. Currently the ALD method is used e.g. in production of read-heads of magnetic discs (Kautzky et al. 2008), for deposition of gate oxide (HfO_2) for 45 nm (year 2007) and 32 nm (year 2009) CMOS devices, and for deposition of dielectric layers of DRAM cells (Cho et al. 2007).

An interesting trend observed in connection with thin-film technologies is alternative application of methods initially developed mainly for electronic industry. For instance, it has been demonstrated recently (Lee et al. 2009) that ultrathin coatings that have been used as dielectrics in electronic devices can markedly increase strength of fibres of biological origin. Even more important is exploiting this kind of coatings for improving performance of packaging materials (Hirvikorpi et al. 2010). In this application, coatings with a thickness of only 25 nm deposited by ALD have dramatically reduced the diffusion of gases and moisture through the materials that could be used, e.g. in food industry.

DEVELOPMENT OF THIN-FILM DEPOSITION METHODS AT THE UNIVERSITY OF TARTU

Methods for real-time characterisation of thin-film growth

Over about 25 years, a considerable experience in thin-film technology has been developed at the Institute of Physics of the University of Tartu (IPUT). Many works of our materials scientists are cited e.g. in a quite recent review on ALD (Puurunen 2005). This research has been concentrated on dielectrics such as Al_2O_3 (Aarik et al. 1990), Ta_2O_5 (Aarik et al. 1994), TiO_2 (Rosental et al. 1997), HfO_2 (Aarik et al. 1999), and ZrO_2 (Kukli et al. 2001), as well as sensor materials e.g. SnO_2, (Rosental et al. 2003) and Cr_2O_3 (Tarre et al. 2008). The studies contributed to better understanding of ALD processes and development of new deposition routes based on novel precursor combinations investigated by using different types of ALD reactors (Fig. 1) designed and constructed at the University of Tartu. A specific feature of the research performed was extensive use and advancement of the quartz crystal microbalance (QCM) method, which allowed real-time characterisation of the thin-film growth with a sensitivity well below the atomic monolayer level in a wide range of substrate temperatures (Aarik et al. 1990, 1994, 1999). In addition to development of novel ALD routes, determination of adsorption probabilities of precursors is an example for application of this method (Aarik, Siimon 1994).

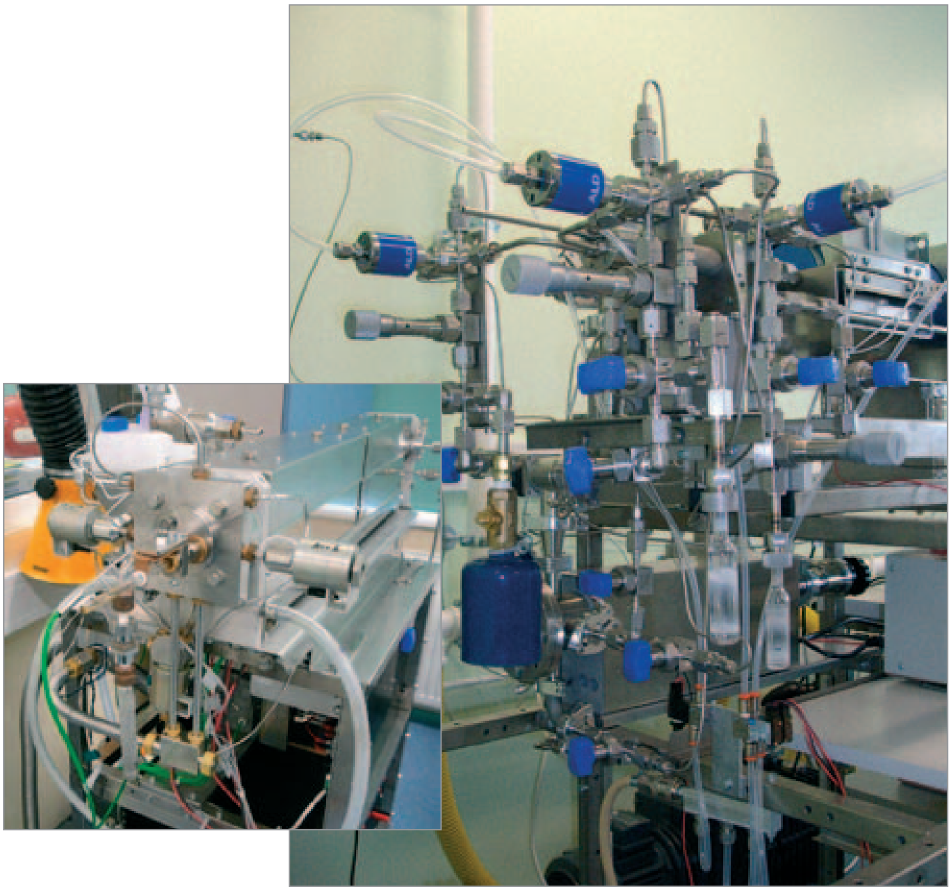


Figure 1.
ALD reactors designed at the University of Tartu for development and characterisation on thin-film deposition processes.

Another method that has successfully been used at IPUT is the incremental dielectric reflection technique based on the reflection of the polarised light from the surface of the growing film (Rosental et al. 1997). This method also allows real-time monitoring of adsorption with a resolution of less than one atomic layer during ALD. Although no mass changes can be detected, recording of the changes in reflectance and optical thickness of a film enables one to select the optimum growth regimes. Differently from QCM that can be used for investigation of deposition processes only on the mass sensor of QCM, the incremental dielectric reflection characterises adsorption processes on the surface of a real substrate and, thus, yields information about nucleation processes on different substrates (Tarre et al. 2008). A similar approach has been applied at IPUT for characterisation of pulsed laser deposition

(PLD) processes as well (Kodu et al. 2011). Due to the access to high-sensitive real-time characterisation methods, development of new thin-film processing routes based on ALD and PLD techniques will probably be an important research area also in the future. Successful co-operation with producers of precursor chemicals forms a good basis for this research direction.

Metals and metal oxides for micro- and nanoelectronic applications

The studies aimed at micro- and nanoelectronic applications have been performed at IPUT in extensive international co-operation. Electrical characterisation of films deposited at IPUT, comprising the capacitance and leakage current evaluation, clarification of dominant conduction mechanisms, estimation of the density of bulk and interface electronic trap densities via deep level transient spectroscopy and related methods, together with high-resolution transmission microscopy, have been carried out in collaboration with partners from the University of Valladolid, Uppsala University and University of Helsinki (Kukli et al. 2005b; Dueñas et al. 2006). During these international and interdisciplinary studies, several aspects of the growth and physical quality of thin films have been explored. Possibilities to control the phase composition and electrical properties, e.g. leakage current and electronic defects at interfaces, have been revealed. It has been demonstrated that the application of double-temperature growth regime, i.e. deposition of a seed layer at low temperature (150–300 °C) followed by the deposition of the rest of the film at higher temperatures (400–600 °C) can lead to solid layers with improved structural homogeneity (Aarik et al. 2004) and lower interface trap density (Dueñas et al. 2006). The double-temperature growth recipe developed at IPUT later turned out to be very useful in the studies devoted to the growth of high-k dielectrics on graphene (Alles et al. 2011) but further investigation in this direction might lead to many other applications as well.

During recent years, the studies of dielectrics potentially relevant to electronics have been gradually shifted from the thin films deposited on silicon to those grown on alternative substrates, such as ruthenium, platinum and iridium serving as bottom electrodes of capacitor structures. It was found in our experiments performed in co-operation with partners from the University of Helsinki that all these metals ensure higher effective permittivity (and capacitance) of the HfO_2-based capacitor stacks compared to the layers grown on silicon (Kukli et al. 2005a). Ruthenium is, however, the most promising material for bottom electrodes because the ruthenium precursors would probably be the cheapest choice among the noble metal sources.

Since the ruthenium oxide (RuO_2) is almost as good conductor as metallic Ru and the smoothness of RuO_2 may well exceed that of Ru, RuO_2 electrodes have also become of great interest. The RuO_2 layers can be deposited using either ALD or metal-organic chemical vapour deposition (MOCVD). Crystalline RuO_2 has the rutile structure. Thus, the dielectrics to be contacted to RuO_2 electrodes in capacitor structures of memory cells should be selected among the high permittivity materials crystallising also in the rutile phase because interfaces of the best quality are formed between the material layers with matched crystal lattices. To avoid detrimental effect of interface layers and find ways to increase the capacity of DRAMs, possibilities for preparation of epitaxial RuO_2-TiO_2-RuO_2 structures were studied in our co-operation with the Institute of Electrical Engineering of Slovak Academy of Sciences (IEESAS). The RuO_2 films were deposited by CVD at IEESAS and the TiO_2 films were grown by ALD at IPUT. In these structures, we obtained the highest dielectric constant values (up to 150) ever reported for ultrathin TiO_2 films (Fröhlich et al. 2009). Very recent results obtained show that a marked progress can also be expected in reduction of the leakage current densities and the continuation of the research in this direction may lead to even more interesting results.

Because of the encouraging results obtained, growth of dielectrics (TiO_2, Al_2O_3, HfO_2 etc.) on metals (Ru, Ir etc.) and oxides of metallic conductivity (e.g. RuO_2) as well as the growth of metals on dielectrics are currently being investigated in order to find reliable methods for processing epitaxial metal-dielectric structures with layer thicknesses of 1-20 nm on planar and nano-profiled substrates. Particular attention is paid to structures grown on nanopatterned templates and in nanosize trenches. In this connection, local epitaxy of dielectrics on polycrystalline electrodes would be of significant importance to realise advantages of epitaxial structures on nanoprofiled substrates that have prospective applications in nanoelectronic devices (e.g. in DRAMs) of the next generation. In these studies, further co-operation of IPUT with IEESAS in preparation and characterisation of epitaxial structures on RuO_2 electrodes, Linköping University in transmission electron microscopy studies, Praxair Inc. (USA) and Air Liquide (France) in investigation of new precursors, Fraunhofer Center Nanoelectronic Technologies in Dresden (Germany) in evaluation of high-permittivity oxides and metal electrodes, University of Helsinki and University of Valladolid in post-growth material studies and with National Center for Scientific Research Demokritos (Athens, Greece) in application of nanodots in flash memories is anticipated.

A new type of memory devices that has been extensively studied during last few years is based on switching of resistivity in metal-dielectric-metal structures. These structures are very similar to those used in capacitor cells of DRAMs. Hence, basing on knowledge obtained in the latter devices, research in the field of resistive switching might also be an interesting and promising field of development at IPUT. The first results obtained in co-operation of researches of IEESAS are encouraging (Hudec et al. 2011) suggesting continuation of this research direction as well.

As mentioned already, Ru and RuO_2 are very attractive electrode materials for micro- and nanoelectronic applications. This has stimulated investigation of corresponding deposition processes, particularly those yielding films with uniform thickness also on nanoprofiled substrates needed for next-generation electronic devices. Therefore ALD of ruthenium from novel precursors provided by Praxair, Ltd. was studied thoroughly at IPUT in co-operation with other partners (Kukli et al. 2011). Deposition of uniform Ru thin films has been realised on a variety of substrates, including Si and different metal oxides. The precursors allowed one to deposit metallic Ru already at as low temperatures as 250–275 °C but the nucleation and growth rate turned out to be very sensitive to the composition of the substrate material. Although the conformal deposition of a metal on 3D substrates is generally quite a complicated task, the first trials carried out at IPUT (Kukli et al. 2011) were very promising in terms of the thickness uniformity achieved. If the electrode metal deposition process on 3D substrates could be united with well-working processes of high-permittivity metal oxide materials based e.g. on ZrO_2 (Tamm et al. 2010), the 3D memory cells based on Ru electrodes and high-*k* oxides could be realised. However, there are still serious obstacles to overcome before the problems related to successive deposition of metal and dielectric layers on 3D substrates will be solved. Hence, the corresponding research will certainly be continued worldwide and at IPUT too.

In these studies, improved control on the deposition of ultrathin metal films will be of particular significance. As already stated above, smooth and homogeneous electrode layers are required for capacitor structures of DRAMs. On the other hand, growth of well-separated metal nanoclusters is of great interest, as these nanoclusters can be applied in flash memories, efficient electron emitters, nano-optical devices, chemical sensors etc. The methods that will be advanced should provide great opportunities for growing smooth homogeneous films as well as nanoclusters of well-controlled sizes. In order to realise these challenges, a detailed study of nucleation on different surfaces is needed using different commercial and non-commercial precursors in ALD and variation of process parameters in physical deposition (e.g. electron beam evaporation, magnetron sputtering and PLD) experiments. Access to appropriate real-time and post-growth characterisation methods is of crucial importance in this research. Fortunately, a significant improvement in possibilities for characterisation of this kind of materials is expected at IPUT already in this year after installation of modern X-ray analysis equipment allowing, for instance, grazing incidence small-angle X-ray scattering (SAXS) studies.

Thin films for solar cells and sensors

Although most of the metal oxides mentioned above have been of major interest for micro- and nano-electronics, ultrathin films of these materials can also be used as optical and/or protective coatings, sensor materials, etc. For instance, TiO_2 chemical barrier layers with a thickness of a few nanometres have been used in solar cell structures with nanostructured ZnO bottom electrode prepared in co-operation of IPUT and Tallinn University of Technology (Oja-Acik et al. 2009). Surface passivation of high-efficient solar cells is also a prospective application field of very thin dielectric films. In international co-operation, corresponding deposition technologies are currently being developed also at IPUT.

It has been shown that the optimum thickness of gas-sensitive SnO_2 layers in chemical sensors might be as low as 10 nm (Rosental et al. 2003). Moreover, in order to increase the sensitivity and selectivity of the sensors, surface activation with catalyst layers with sub-nanometre thickness is of major interest. Besides SnO_2, chemically sensitive materials, such as Co_3O_4 (Kärkkänen et al. 2009), $CoWO_4$ (Kärkkänen et al. 2010), and Cr_2O_3-TiO_2 (Jantson et al. 2005) have been investigated at IPUT. An interesting effect observed in very thin epitaxial Cr_2O_3 films deposited on sapphire substrates is a change of crystallographic orientation at the film surface, when a critical thickness is reached (Tarre et al. 2008). This effect might lead to marked changes in the gas sensitivity of the material and after further studies could find attractive applications in sensor technology.

A material, which can potentially form a basis for chemical as well as magnetic sensors, is iron oxide. Since the transparent conductive phase of iron oxide – magnetite (Fe_3O_4) – is of particular interest as possible host or dopant for spintronic materials, including field sensors based on tunnel junctions, studies are being conducted in order to clarify the conditions favouring the formation of magnetite. It has been found in our experiments that the exploitation of low process temperatures and/or appropriate substrate materials such as SnO_2 promotes the stabilisation of magnetite without specific reduction procedures. In addition, promising experiments have been carried out in order to achieve conformal growth of Fe_2O_3/Fe_3O_4 on 3D substrates (Fig. 2).

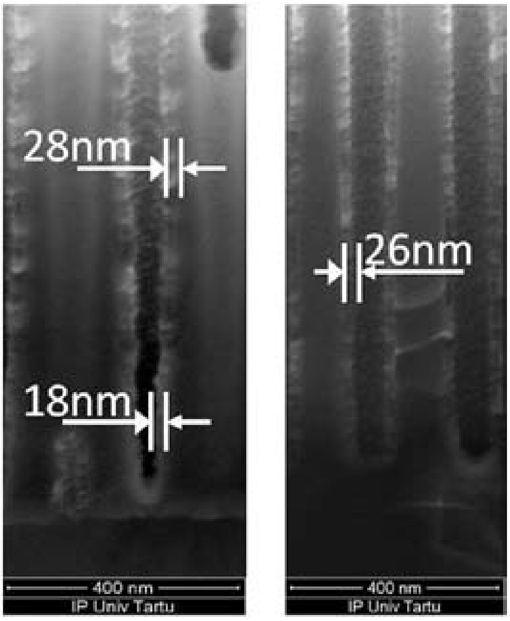


Figure 2.
Iron oxide thin films deposited by ALD on the walls of trenches with depth to diameter ratios of 20 (left panel) and 40 (right panel). Before deposition of iron oxide the surface was covered with SnO_2 nanospheres pre-pared by sol-gel technology.

Magnetometric measurements, performed in co-operation with National Institute of Chemical Physics and Biophysics, have shown that magnetite-rich films grown by ALD possess markedly high saturation magnetisation and well-defined coercivity (Tamm et al. 2011). ALD of magnetic materials as well as magnetometry on very thin solid films grown by this method is a relatively new direction worldwide, not only in Estonia, and will offer an appreciably wide basis for studies in future together with large selection of potentially interesting materials. Furthermore, films such as Ni-doped TiO_2 (Kisand et al. 2010) or Co-doped TiO_2 (Pärna et al. 2011) deposited at IPUT by the sol-gel method with the target composition promising chemical response to gases might, actually, also demonstrate ferri/ferromagnetic response to external fields. Magnetic studies on these films are yet to be carried out.

Graphene studies

Graphene studies were started at IPUT in collaboration with the Low Temperature Laboratory of Helsinki University of Technology (currently Aalto University) in 2008. First, possibilities for deposition of thin (10-30 nm thick) high-quality dielectric layers of high-k oxides (TiO_2 and HfO_2) on graphene prepared by micromechanical exfoliation technique were investigated. In order to obtain films with the highest dielectric properties, the ALD method was chosen for thin film deposition although it was known that deposition of uniform thin films on graphene by this method is not a simple task because on the defect-free graphene surface, there are no free chemical bonds, which are needed for ALD reactions. Nevertheless, using two-temperature growth processes described above, thin dielectric films were successfully deposited directly on graphene without surface functionalisation that frequently results in degradation of the electrical characteristics of graphene. In our experiments, good sticking of the monoclinic HfO_2 films to high-quality graphene was achieved as the Raman spectroscopy studies revealed biaxial strain but no generation of defects in graphene (Alles et al. 2011).

Currently the materials scientists of IPUT are participating in two international graphene projects. In the first one, "Entangled spin pairs in graphene", launched in the framework of the European Collaborative Research program EuroGRAPHENE, the materials scientists of IPUT work together with researchers from Aalto University, University of Würzburg, University of Geneva, University of Basel and University of Leiden. The goal of this project is the demonstration of entangled spin pairs in graphene, which would open up several interesting possibilities in the use of graphene in spintronics. The task of IPUT is to develop methods for preparation of tunnel barriers and ferromagnetic layers, which are necessary in order to guide the movement of charge carriers in graphene. ALD and electron beam evaporation are the main methods to be used. The expected results of the work are applicable also beyond graphene studies.

Another ongoing international graphene project is "Suspended graphene nanostructures" (RODIN, FP7) in which researchers of IPUT together with colleagues from Chalmers University of Technology, Aalto University, Technical University of Delft, University of Cambridge and Catalan Institute of Nanotechnology and industrial partners from Diarc-Technology Oy and Nokia Research Center investigate suspended single- and few-layer graphene nanostructures and annealed diamond-like carbon films. The main goal of the project is the fabrication and demonstration of a tuneable graphene resonator with electronic readout. Immediate impact area for devices using such resonators is the subarea of radiofrequency components used in front ends of radio communication devices as band-selection filters, switches and antennas, which are rapidly becoming a major obstacle in cost- and size-efficient production of mobile phones.

For efficient contribution to the graphene-related projects, large-area graphene samples are needed. Therefore, in order to fabricate such samples, a CVD method was very recently successfully taken into use in IPUT. In February 2011 the first high-quality CVD graphene sheets (Fig. 3) with an area of several square centimetres were produced in a home-built reactor on copper foils.

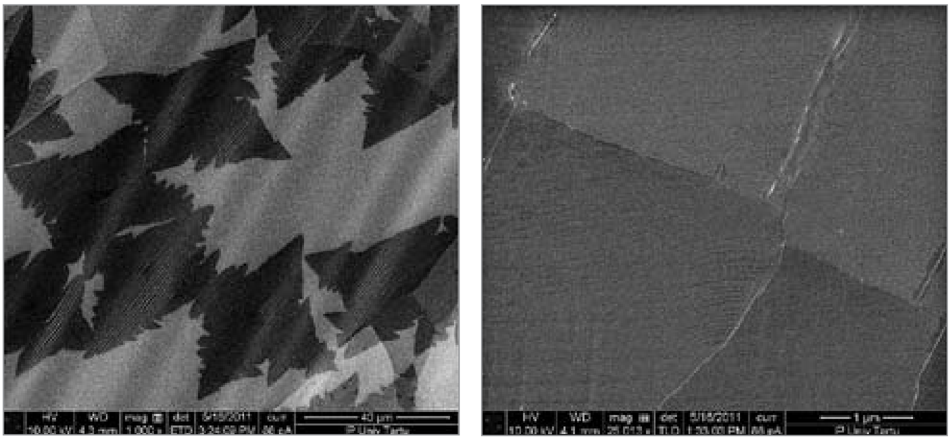


Figure 3.
High-resolution scanning electron microscopy (SEM) images of graphene prepared by CVD method on copper foil. According to Raman spectroscopy data, the sample contained significant amount of monolayer graphene. SEM images show, however, formation of multilayer graphene (darker areas on images) as well.

The technology for transferring of the graphene sheets from copper foils to other substrates (Si, SiO_2, Al_2O_3) for further studies has also been implemented at IPUT by now. One of the possible applications of the CVD graphene is related to development of chemical sensors in the framework of a national research project "Manufacturing, processing and characterisation of graphene-based nanostructures" which has been running in IPUT since 2009 thanks to the European Social Fund (Mobilitas-programme). It's worth mentioning that the graphene-based sensors might be very sensitive, having detection limits in the order of one part per billion (Schedin et al. 2007). Nevertheless, much work has to be done in order to produce graphene-based gas sensors suitable for routine environmental monitoring.

Anti-corrosion coatings

Corrosion leads to the destruction of materials and is usually associated with changes in the oxidation state of a metal due to electrochemical or chemical reactions in a reactive environment containing water and oxygen with or without salts or acid/alkali additions. There are several methods used for protecting metal surfaces against corrosion (Roberge 2000). Surface passivation (followed by painting or varnishing) and galvanic and sprayed coatings are some examples of those. Sometimes corroding material can be encapsulated into a non-corrosive jacket, e.g. ceramic tube, but there are severe limitations for using this method. An alternative solution, less studied for anti-corrosion purposes so far, is coating the material with a thin but very dense sealing film, e.g. by ALD (Díaz et al. 2011).

The materials scientists from IPUT are currently participating in the 7th FP project "Corrosion protection with perfect atomic layers" (CORRAL) that is focused on protection of some industrial metal alloys against corrosion. The project involves universities and research institutions from Germany, France, Netherlands, Finland, Switzerland, Hungary, United Kingdom and Estonia, and several medium and large European companies, like Hauzer Techno Coatings BV, Messier-Bugatti and Linde AG, and should result in development of methods for preparation of high-density, defect-free ultrathin sealing coatings with excellent barrier properties and improved corrosion resistance. The research is concentrated on novel nanostructured coatings, such as nanoscale multilayers (Fig. 4), mixed and composite coatings. These impermeable sealing layers should be able to block the ion exchange between the substrate material and an aggressive environment, thus offering an efficient long-term protection against corrosion. Four alternative vapour deposition techniques, including ALD, have been planned to be used to demonstrate the feasibility of corrosion protection in some targeted industrial sectors producing e.g. high precision bearings, aerospace components (break

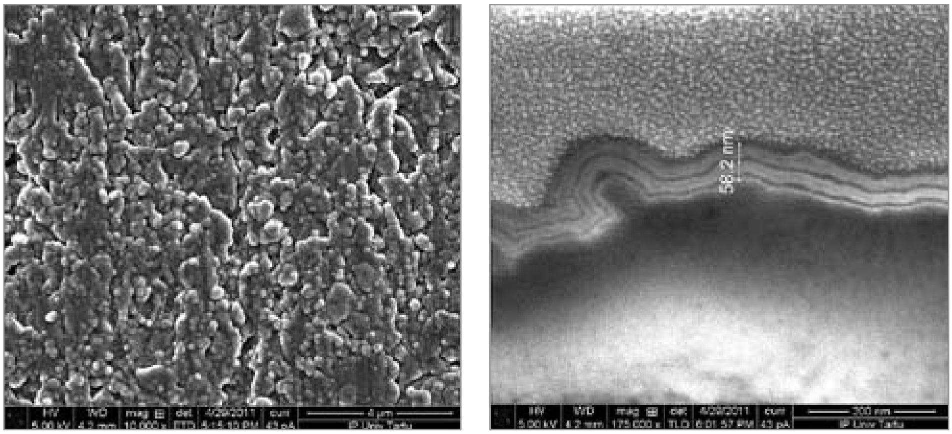


Figure 4.
High-resolution SEM images of anti-corrosion coating consisting of 5 Al_2O_3 and TiO_2 double layers with a total thickness of 58.2 nm on pre-treated surface of Al-alloy. Left panel shows topography of the surface. Cross-section (right panel) shows Al-alloy (lower part of the image) covered with conformal coating (in the middle of the image) deposited by ALD method and with an amorphous Pt-layer (upper part of the image) deposited on top of the coating to protect the structure during ion beam preparation of the cross section.

systems) and gas handling components. Researchers of IPUT have designed and constructed an ALD reactor (Fig. 1, right panel) for coating the gas handling components and have demonstrated the applicability of the reactor and the ALD method for this purpose. Currently industrial tests of the coatings are underway.

Coatings for preventing metal corrosion in reactive environments at elevated temperatures, e.g. in reactors or funnels, are being developed in cooperation of IPUT with Estonian Nanotechnology Competence Centre. For that purpose, a novel two-step ALD method has been introduced (Aarik et al. 2011). The further development in the area of corrosion protection will evidently be related to investigation of composite and hybrid anti-corrosion nanocoatings, in which carbon nanoparticles and/or 2D structures, polymer thin films as well as ceramic ultrathin films prepared by low-temperature ALD will be used.

CONCLUDING REMARKS AND FUTURE OUTLOOK

The studies of Estonian materials scientists have contributed to different applications of thin solid films ranging from functional components of electronic devices and sensors to protective coatings. Since the early 1990s, research teams of the University of Tartu have developed and advertised real-time characterisation of the thin-film deposition processes, and contributed to deeper understanding the growth mechanism of ultrathin inorganic layers. Parallel studies have been and still are devoted to the thorough structural and morphological characterisation of thin solid films deposited in the ranges of process parameters exceeding those commonly investigated or published by other groups. As a result, the research in this field of materials science, i.e. in thin solid films, has enabled creation of international collaboration with the materials scientists, chemists, physicists and engineers from many countries. Some well-working contacts to industrial partners have been established in the form of either bilateral collaboration or within the frameworks of projects funded by European Union.

Different thin-film deposition methods such as ALD, PLD, electron beam evaporation, magnetron sputtering and sol-gel technology, have been successfully implemented at IPUT and an additional synergy can be gained from combining of the methods in the future. For example, PLD belonging to physical vapour deposition methods allows growth of thin films much faster and with markedly higher chemical purity than ALD does. Sol-gel as a chemical solution-based method enables very fast growth of films and, differently from PLD, on large-area substrates as well. Although the films usually contain considerable amounts of impurities not belonging to the target composition, they have a number of interesting applications due to their relatively low density, considerable porosity and/or ability to fill (larger) pores in the substrate. On the other hand, ALD is, differently from PLD and sol-gel methods, based on surface-controlled adsorption of gaseous species and yields thin solid films with very high density and precise and uniform thicknesses even on 3D substrates. Thus, combination of these methods could allow covering very wide application range clearly exceeding that discussed in this short review. Particularly, more extensive application of the methods for modification and functionalisation of nanoporous materials in technology of fuel cells and supercapacitors can be anticipated.

A marked positive effect on research in the field of thin-film technology is expected due to improvement of the infrastructure in the last few years. Purchase of modern dual-beam equipment Helios Nanolab 600 (FEI) in 2009 and installation of spectroscopic ellipsometer GES-5E (Semilab Sopra) in 2011 are few examples of the infrastructure upgrade. This process will be further continued to follow the contemporary and future needs for precise and fast measurements and analysis. For instance, a new X-ray fluorescence spectrometer and modern X-ray reflection, X-ray diffraction and SAXS equipment will be purchased to characterise more adequately thin films with thicknesses down to few nanometers and nanoclusters. For more detailed electrical characterisation of potential components of nanoelectronic devices, IPUT will be equipped with a contemporary measurement station consisting of a new induction-capacitance-resistance meter and an analyser of current-voltage characteristics. In order to improve the conditions for material processing, thin-film deposition equipment has been and will be upgraded. A marked support for this kind of infrastructure development has been obtained from the European Regional Development Fund.

REFERENCES

Aarik, J., et al. 1990. Properties of amorphous Al_2O_3 films grown by ALE. Acta Polytechnica Scandinavica, Chemical Technology and Metallurgy Series, 195, 201-209.

Aarik, J., et al. 1994. Deposition and etching of tantalum oxide films in atomic layer epitaxy process. Journal of Crystal Growth, 144, 116-119.

Aarik, J., et al. 1999. Influence of substrate temperature on atomic layer deposition of HfO_2 thin films. Thin Solid Films, 340, 110-116.

Aarik, J., et al. 2004. Effects of precursors on nucleation in atomic layer deposition of HfO_2. Applied Surface Science, 230, 292-300.

Aarik, L., et al. 2011. High-temperature corrosion Resistive ALD coatings for nickel alloys. European Materials Research Society (E-MRS) Spring Meeting 2011, May 9-13, Nice, France.

Aarik, J., Siimon, H. 1994. Characterization of adsorption in flow type atomic layer epitaxy reactor. Applied Surface Science, 81, 281-287.

Alles, H., et al. 2011. Atomic layer deposition of HfO_2 on graphene from $HfCl_4$ and H_2O. Central European Journal of Physics, 9, 319-324.

Chen, S., et al. 2011. Oxidation resistance of graphene-coated Cu and Cu/Ni alloy. ACS Nano, 5, 1321-1327.

Cho, H. J., et al. 2007. New TIT capacitor with $ZrO_2/Al_2O_3/ZrO_2$ dielectrics for 60 nm and below DRAMs. Solid-State Electronics, 51, 1529-1533.

Díaz, B., et al. 2011. Low-temperature atomic layer deposition of Al_2O_3 thin coatings for corrosion protection of steel: Surface and electrochemical analysis. Corrosion Science, 53, 2168-2175.

Dueñas, S., et al. 2006. Influence of single and double deposition temperatures on the interface quality of atomic layer deposited Al_2O_3 dielectric thin films on silicon. Journal of Applied Physics, 99, 054902.

Fröhlich, K., et al. 2009. Epitaxial growth of high k TiO_2 rutile films on RuO_2 electrodes. Journal of Vacuum Science & Technology B, 27, 266-270.

de Heer, W.A., et al. 2007. Epitaxial graphene. Solid State Communications, 143, 92-100.

Hirvikorpi, T., et al. 2010. Effect of corona pretreatment on the performance of gas barrier layers applied by atomic layer deposition onto polymer-coated paperboard. Applied Surface Science, 257, 736-740.

Hudec, B., et al. 2011. Resistive switching in $RuO_2/TiO_2/RuO_2$ MIM structures for non-volatile memory application. ASDAM 2010, Conference Proceedings, (in press).

Jantson, T., et al. 2005. Nanocrystalline Cr_2O_3-TiO_2 thin films by pulsed laser deposition. Sensors and Actuators B, 109, 24-31.

Kautzky, M. C., et al. 2008. Atomic layer deposition of Al_2O_3 films for permanent magnet isolation in TMR read heads. IEEE Transactions of Magnetics, 44, 3576-3579.

Kim, K. S., et al. 2009. Large-scale pattern growth of graphene films for stretchable transparent electrodes. Nature, 457, 706-710.

Kisand, V., et al. 2010. Influence of the heating temperature on the properties of nickel doped TiO_2 films prepared by sol–gel method. Applied Surface Science, 256, 4538-4542.

Kodu, M., et al. 2011. Hydrogen doping of MgO thin films prepared by pulsed laser deposition. Applied Surface Science, 257, 5328-5331.

Kukli, K., et al. 2001. Atomic layer deposition of zirconium oxide from zirconium tetraiodide, water and hydrogen peroxide. Journal of Crystal Growth, 231, 262-272.

Kukli, K., et al. 2005a. Atomic layer deposition and characterization of HfO_2 films on noble metal film substrates. Journal of the Electrochemical Society, 152, F75-F82.

Kukli, K., et al. 2005b. Engineering structure and properties of hafnium oxide films by atomic layer deposition temperature. Thin Solid Films, 479, 1-11.

Kukli, K., et al. 2011. Atomic layer deposition of ruthenium films from (ethylcyclopentadienyl)(pyrrolyl)ruthenium and oxygen. Journal of the Electrochemical Society, 158, D158-D165.

Kärkkänen, I., et al. 2009. Modelling and characterisation of Co_3O_4 thin film gas sensors. Procedia Chemistry, 1, 654-657.

Kärkkänen, I., et al. 2010. Sensitivity of $CoWO_4$ thin films to CO. Procedia Engineering 5, 160-163.

Lee, S.-M., et al. 2009. Greatly increased toughness of infiltrated spider silk. Science, 324, 488-492.

Li, X., et al. 2009. Large-area synthesis of high-quality and uniform graphene films on copper foils. Science, 324, 1312-1314.

Novoselov, K. S., et al. 2004. Electric field effect in atomically thin carbon films. Science, 306, 666-669.

Oja-Acik, I., et al. 2009. Nanostructured solar cell by spray pyrolysis: Effect of titania barrier layer on the cell performance. Thin Solid Films, 517, 2443-2447.

Puurunen, R. 2005. Surface chemistry of atomic layer deposition: A case study for the trimethylaluminum/water process. Journal of Applied Physics, 97, 12301.

Pärna, R., et al. 2011. Effect of cobalt doping and annealing on properties of titania thin films prepared by sol-gel process. Applied Surface Science, 257, 6897-6907.

Roberge, P. R. 2000. Handbook of Corrosion Engineering. McGraw-Hill Publ. Co.

Rosental, A., et al. 1997. Atomic layer deposition in travelling-wave reactor: In situ diagnostics by optical reflection. Applied Surface Science, 112, 82-86.

Rosental, A., et al. 2003. Gas sensing properties of epitaxial SnO_2 thin films prepared by atomic layer deposition. Sensors and Actuators B, 93, 552-555.

Schedin, F., et al. 2007. Detection of individual gas molecules adsorbed on graphene. Nature Materials, 6, 652-655.

Tamm, A., et al. 2010. Atomic layer deposition and characterization of erbium oxide-doped zirconium oxide thin films. Journal of The Electrochemical Society, 157, G193-G201.

Tamm, A., et al. 2011. Atomic layer deposition of magnetic iron oxide films on planar and 3D substrates with SnO_2 nanoparticles. European Materials Research Society (E-MRS) Spring Meeting 2011, May 9-13, Nice, France.

Tarre, A., et al. 2008. Atomic layer deposition of Cr_2O_3 thin films: Effect of crystallization on growth and properties. Applied Surface Science, 254, 5149-5156.

Wong, H., Iwai, H. 2006. On the scaling issues and high-k replacement of ultrathin gate dielectrics for nanoscale MOS transistors. Microelectronic Engineering, 83, 1867-1904.

Zhang, L., Zhou, R., Zhao, X. S. 2010. Graphene-based materials as supercapacitor electrodes. Journal of Materials Chemistry, 20, 5983-5992.

PHYSICAL BIOLOGY: *EN ROUTE* TO QUANTUM BIOLOGY?

A. Freiberg

Complex relationships between the quantum realms of molecular processes and the classical world of living organisms are outlined. Recent validation of the survival of photosynthetic excitons at functional temperatures is yet another example bridging the gap between quantum physics and biology.

INTRODUCTION

Not long ago, a physicist and a layman alike would think, "Physical biology, what nonsense!" Indeed, from the middle and high school experience, nothing seems to be farther away from each other than these two disciplines. Physics deals with inanimate matter, while biology, with the living world. Biology is wet, warm, and messy, functioning at ambient temperatures; physics is solid, cold, and neat, rather studied at cryogenic temperatures. Last, but not least, to understand physics one needs math, while biology can be figured out 'naturally'. This confusion is an unfortunate result of the extreme specialisation of modern science. Once called 'natural philosophy', the scientific study of all of nature, the science began to divide into separate disciplines with increasing pace since the Renaissance. Great advantages of the narrow specialism of various scientific disciplines are obvious. As a result of that, hidden secrets of nature at the scales of both the very large and the very small have been revealed with unprecedented details. At the same time, pointless barriers have been raised which largely prevent scientists from different fields from acknowledging common grounds of their subject matters and from recognising broader perspectives. Yet science develops cyclically and this state of affairs is bound to change. Surfacing of the various mixed two-word subjects such as the one used in the title of this treatise precisely reflects this trend. The reductionist approach has been enormously successful and literally dominated research in biology for over the last 50 years. However, complexity of cellular processes reinforces an opposite, integrated method that has not long ago materialised in a discipline called systems biology. The physical biology, known also as biological physics or biophysics, combines essential knowledge from physics and biology of the living state of matter. One of the lessons taught by the physical biology is that on the very basic level the biology is as quantitative a science as the physics is. The emerging fields of physical and systems biology are clearly among the most exciting frontiers of the 21st century science (Zewail 2008).

Physics has fundamentally a quantum origin. Photosynthesis, on the other hand, is a basic biological process, a major source of life on Earth. However, since photosynthesis begins with absorption of solar quanta, it can be considered a subject at the borderline of classical and quantum realms. The physical concept of 'exciton', a collective electronic excitation of condensed matter introduced by J. Frenkel almost exactly eighty years ago, is by now well established in photosynthesis (Van Amerongen et al. 2000). It is believed that absorption of a solar photon by chlorophyll and other chromophores in light-harvesting (LH) antenna complexes creates a coherent exciton that transfers its energy very efficiently to the photochemical reaction centres (RC), where it is subsequently transformed via vectorial electron transfer steps into potential chemical energy. Yet this insight has primarily arisen from interpretations of the low-temperature spectroscopic measurements. Experimental evidence for the excitons at physiological temperatures is scarce at best. Thorough spectroscopic studies of *in vitro* and *in vivo* (bacterio) chlorophyll molecules performed in our laboratory over a broad temperature range from liquid helium to ambient temperatures provided first clear proof for the presence of excitons, hence of long-lasting (stationary) quantum coherences, in photosynthetic systems at functional temperatures.

RECENT ADVANCES IN UNDERSTANDING QUANTUM EXCITATIONS OF *IN VITRO* (BACTERIO) CHLOROPHYLLS

Chlorophylls are ubiquitous in plants and algae, similar to bacteriochlorophylls in bacterial photosynthetic systems (Blankenship 2002). The useful functioning of the (bacterio) chlorophyll molecules is based on their unique quantum-optical and redox properties. Being the lowest-energy electronic transitions in the visible range, the Q_y singlet electronic transitions are instrumental in all photophysical and -chemical processes these chromophores are involved in. Their properties in dependence on solvent polarity, axial coordination of the central magnesium atom, pigment aggregation, temperature, external pressure and electric field, etc. have experimentally been studied using almost all conceivable spectroscopic techniques. Despite this intensive scrutiny, optical spectroscopy of the (bacterio) chlorophyll molecules is still not wholly understood, neither experimentally nor theoretically.

The couplings between a molecular electronic transition and intramolecular nuclear vibrations determine the vibronic structure of optical spectra of individual chromophores. The chromophore-host bath interactions present in the solvent phase give rise to the electron-phonon structure of the spectra. Phonons are collective vibrational modes (standing waves) of the surrounding matrix in which all the matrix particles simultaneously participate. The electron-phonon and intramolecular vibronic interactions together shape the so-called homogenous spectra of impurity molecules. Avarmaa and co-workers (Avarmaa, Rebane 1985) were the first to obtain high-resolution spectra related to the Q_y electronic state of the impurity chlorophyll *a* (Chl *a*) and bacteriochlorophyll *a* (Bchl *a*) molecules. Yet these low-temperature results had rather qualitative than quantitative value, since experimental limitations prohibited determination of the coupling strengths for both the lattice phonons and individual intramolecular vibrations. Zazubovich et al. (2001) achieved this task for the excited electronic state using hole burning in a triethylamine glass. Comparable results for the ground electronic state have been published by us only recently (Rätsep et al. 2009a, 2011).

Fig. 1 demonstrates the homogeneous spectral profiles revealing detailed phonon and vibronic structures of the fluorescence spectra of Chl *a* and Bchl *a* molecules in low-temperature glass matrices. The high-resolution spectra, which comprise a narrow zero-phonon line (ZPL) at the spectral origin, a phonon sideband (PSB) attached to the ZPL from the low-energy side, and up to 50 vibronic replicas distributed over a wide frequency range between ~80 and 1700 cm^{-1} are obtained using the novel difference fluorescence-line narrowing method (ΔFLN). The coupling between the Q_y electronic transition and the intramolecular vibrational degrees of freedom is weak; therefore, the replicas represent single-quantum band profiles for the local vibrational modes. ΔFLN as developed in our laboratory (Rätsep, Freiberg 2003) is a highly selective technique, which combines the strengths of both hole burning and traditional fluorescence-line narrowing methods.

Whether or not asymmetry exists between the Q_y absorption and emission spectra of the Chl *a* and Bchl *a* molecules was another unresolved question. Experimental data concerning this problem were not only scarce but also controversial. Mirror symmetry between the conjugate absorption and emission spectra is theoretically expected within the model framework of the crude adiabatic, harmonic, linear electron-phonon and vibronic coupling, and Condon approximations (also known as the basic model (Rebane 1970), together with an assumption of a fast excited state relaxation (the Kasha-Vavilov rule). While quantitative quantum chemical modelling of the absorption and emission spectra should reveal the nature of the electronic transitions and the origin of any spectral asymmetry, this is a challenging task for the chlorophyll type molecules owing to their rather large size and extended conjugation.

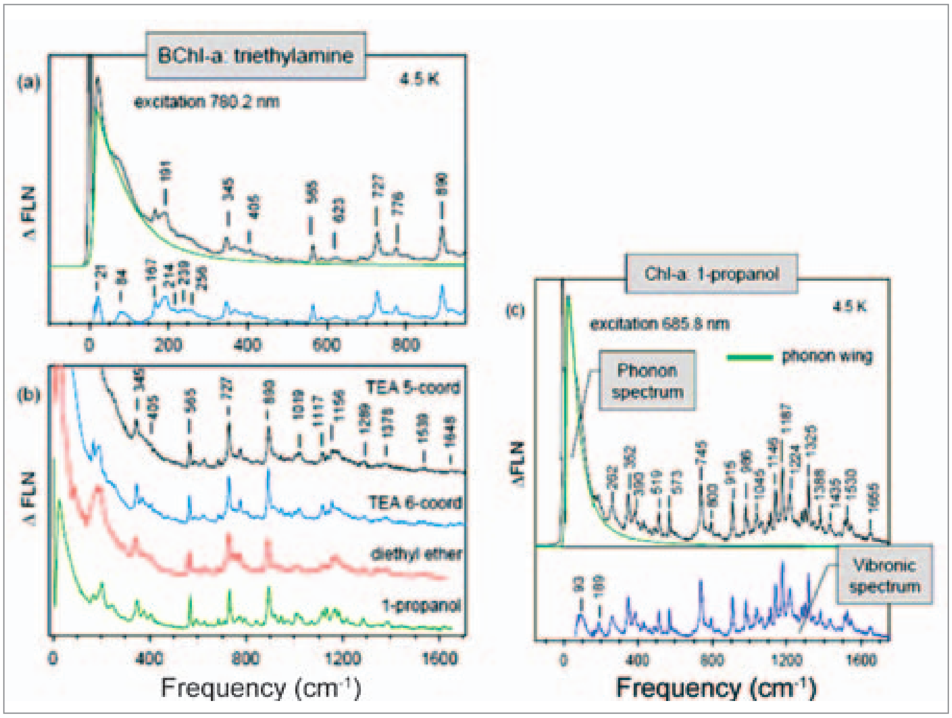


Figure 1.
ΔFLN spectra of Buhl *a* in triethylamine (a) and Chl *a* in 1-propanol (c) at 4.5 K. The narrow ZPL at spectral origins (780.2 nm for Buhl *a* and 685.8 nm for Chl *a*) is cut off at ~5% intensity level to amplify the sideband structure. The numbers indicate the main vibronic frequencies in wave numbers. The modelled phonon sidebands are shown with green curves. The lower blue curves are the difference between the black and green curves offset for clarity. (b) Vibronic region of the ΔFLN spectra of BChl *a* in various five fold- and six fold-coordinated matrices at 4.5 K.

We have experimentally studied the deviations from mirror symmetry in optical absorption and fluorescence emission spectra of Chl *a* and Bchl *a* molecules using different solvents to induce either penta- or hexa-coordination of the central Mg atom (Rätsep et al. 2009a, 2011). Various semi-empirical, density-functional, and *ab initio* methods to find a reliable scheme for the *a priori* prediction of absorption and emission spectra of the (bacterio) chlorophylls have also been investigated. Based on the data obtained at different temperatures, one can safely conclude that mirror symmetry does not hold at any temperature between 4.5 K and ambient temperature neither in case of the Chl *a* nor Bchl *a* molecules. In the Bchl *a* emission, for example, the low-resolution vibronic sideband structure is centred on modes at ~900 cm^{-1}, while in absorption this band moves out to 1100–1200 cm^{-1}. In environments with six fold-coordinated Mg, the emission sideband is considerably stronger, resulting in the corresponding reorganisation energy λ_E slightly larger than the reorganisation energy λ_A in absorption. In five fold coordinated environments, however, λ_A is 50% larger than λ_E. Most computational methods predict the spectral asymmetry to be very much larger than that actually observed; only the long-range corrected density functional CAM-B3LYP reproduced detailed experimental results for Bchl *a* and predicted asymmetry of the correct magnitude. These simulations also revealed that the spectral asymmetry arises from subtle consequences of the Duschinsky rotation on the high-frequency modes, allowing these vibrations to maintain their form on excitation and yet have very different normal-mode displacements through correctly phased minor contributions coming from a large number of low-frequency vibrational modes. In the case of the Chl *a* molecule, breakdown of the Condon approximation in inducing asymmetry appears to be significant. These works have thus established the first complete description of the Q_y absorption and fluorescence spectra of the (bacterio) chlorophyll molecules.

Inhomogeneous broadening of the optical spectra of impurity molecules in solids is a well-known phenomenon. The static spatial fluctuations of the environment can be formally looked at as a certain kind of a random noise that modulates the transition energies, i.e., the spectral positions of ZPL. Possible modifications of the homogenous shapes of the spectra, which are related to electron-phonon and/or vibronic couplings, are ignored in this approach. This turns out to be oversimplification. Fig. 2 demonstrates a significant increase of the linear electron-phonon coupling strength as a function of the excitation wavelength through the inhomogeneously broadened origin bands in the low-temperature absorption spectra of the glassy samples of Chl *a* and Bchl *a* (Rätsep et al. 2009b; Renge et al. 2011).

The linear electron-phonon coupling strength is associated with the displacement of the equilibrium positions of the nuclei upon a photo-excitation of the chromophore. Information about this parameter can be obtained from the so-called Huang-Rhys factor, *S*, which may be experimentally evaluated as: $\exp(-S) = I_{ZPL} / (I_{ZPL} + I_{PSB})$. Here, I_{ZPL} and I_{PSB} refer to the integrated intensities of the ZPL and the PSB, respectively. Physically, the Huang-Rhys factor is a measure of the average number of phonons that accompany a particular

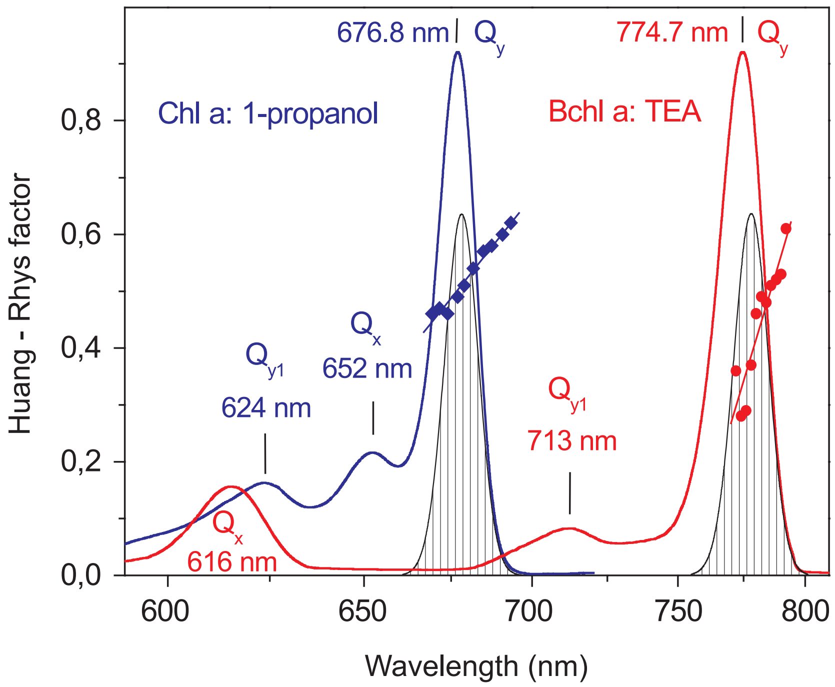


Figure 2.
Wavelength dependence of the electron–phonon coupling strength *S* for the Chl *a*-doped 1-propanol (blue rombs) and for the Bchl *a*-doped triethylamine (red rings) solutions at 4.5 K. The lines connecting the scattered data points present linear fits of the data. Plotted on the background with continuous coloured curves are the corresponding absorption spectra. The areas with drop lines denote inhomogeneous distribution functions of the ZPL related to the Q_y electronic transitions. A reciprocal (linear in energy) presentation of the wavelength scale is used. Q_y and Q_x refer to the two lowest-energy singlet electronic transitions; Q_{y1} indicates vibronic sidebands related to Q_y transitions.

electronic transition. The observed quasi-linear increase of the Huang-Rhys factor with the excitation wavelength suggests a strong correlation between the electron-phonon coupling strength and the solvent shift. Such correlation has been rationalised within a two-particle Lennard-Jones model of intermolecular interactions (Renge et al. 2011). This model also allows changes of the quadratic couplings, which show up in the widths of the ZPL. The latter prediction still waits to be proven experimentally.

QUANTUM COHERENCE SETS THE STAGE FOR EFFICIENT SOLAR ENERGY HARNESSING BY PHOTOSYNTHETIC ANTENNA NANOSTRUCTURES

Nanoscale systems, including molecular aggregates of biological origin, provide a transition between single molecules and bulk materials. A natural question then arises, "Are their inner workings also intermediate to those of the individual molecules and bulk systems?" This fundamental query, although addressed in the past (Kasha 1963), attracted relatively little attention until the last decade. Since then, there has been an explosive wake of interest in small quantum systems due to the advent of refined experimental procedures with which to fabricate and study the molecular nanostructures of great potential technical importance. Notable examples include plastic electronics and photonics, organic solar cells, fluorescent (bio) markers, and artificial photosynthetic complexes. A well-founded general prospect is that in nanoscale the solid-state materials acquire molecular-like features, first of all discrete spectroscopic transitions. The aggregate formation is also often accompanied by rather spectacular changes in optical spectra whose origin is coherent excitation extending over many molecular sites – an exciton.

Since the inner biological machinery uniquely controls the composition and size of the biological nano-aggregates, they are ideally suited for investigations of the basic spectral characteristics of the excitons in nanoscale systems. Great variety of such aggregates are available that include from a few to ~100,000 chromophores. Diversity of available genetic manipulation techniques only adds to the attractiveness of these samples. As an example, herein, we deal with a peripheral light-harvesting 2 (LH2) pigment-protein complex from purple photosynthetic bacteria. The high-resolution crystal structure of the LH2 complex from *Rhodopseudomonas (Rps.) acidophila* (strain 10050) is shown in Fig. 3 (McDermott et al. 1995). It reveals a cyclic array of nine transmembrane α, β – polypeptide pairs. Each polypeptide pair binds two Bchl *a* molecules at the outer membrane surface and one molecule on its intracytoplasmic side, forming two concentric molecular circles of nearly C_9-symmetry. The inner ring of 6-7 nm diameter consists of 18 strongly coupled Bchl *a* molecules, which feature intermolecular distances of less than 1 nm. The molecules in this ring give rise to the intense absorption band around 850-870 nm (B850 band) in the various species of purple bacteria. The other ring consists of 9 largely monomeric Bchl *a* molecules (separation > 2 nm) that absorb around 800 nm (B800 band).

While free in organic solvents, the Q_y transition of the Bchl *a* molecules is at ~775 nm (Fig. 2). The related absorptions of the B800 and, especially, the B850 chromophores appear thus strongly red-shifted with respect to the tran-

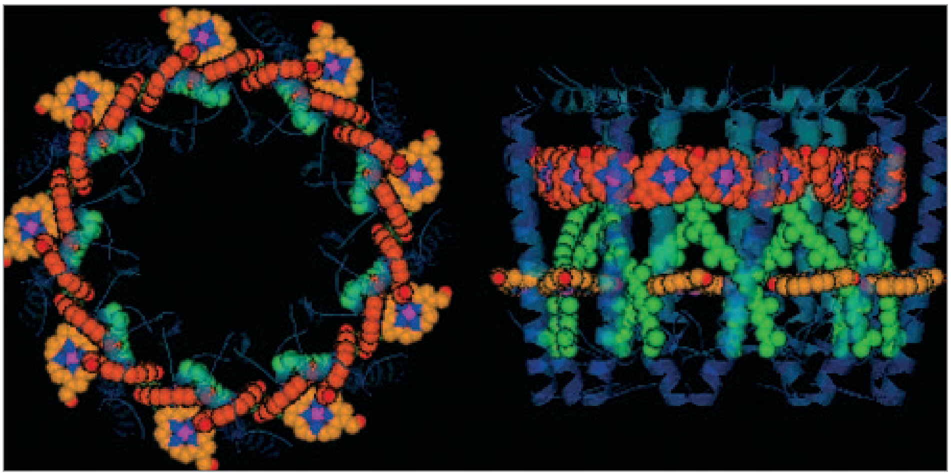


Figure 3.
Molecular structure of the LH2 light-harvesting complex from *Rps. acidophila* viewed top-down (left) and sidewise (right) with respect to the imaginable photosynthetic double-layer membrane. The red-coloured Bchl *a* molecules of the B850 arrangement in the top circle of the sidewise view are squeezed between the two protein walls formed by the membrane-spanning α-helix ribbons. The top-down view reveals dimerisation of the pigments in this ring. The Bchl *a* molecules of the B800 system located at opposite side of the membrane are shown in yellow. The green carotenoid molecules make close contact simultaneously with both the B850 and B800 chromophores. Here and on successive figures, only the porphyrin macrocycles of the Bchl *a* pigments are exposed for clarity. (Figure courtesy of R. Cogdell.)

sition of an individual molecule. This is because the spectral positions of the B850 and B800 bands are determined not only by the transition energies of the specific Bchl *a* sites but also by the dipolar sensitively depend on distances and mutual orientations of the molecules with respect to each other. The site energies of the chromophores can be fine-tuned by subtle interactions with the surrounding protein matrix. The main differences in the absorptions of the Bchl *a* molecules that are organised in the B800 ring and those assembled in the B850 ring stem from the strong (300-400 cm^{-1}, depending on the species (Freiberg, Trinkunas 2009; Pajusalu et al. 2011)) resonant coupling between the transition dipole moments of the closely packed B850 Bchl *a* molecules, leading to the formation of the exciton states. For the B800 molecules the inter-chromophore interactions are much weaker, in the order of 20-30 cm^{-1}, and can be neglected in most applications. In the following, we will predominantly talk about the B850 transitions where the quantum coherence effects cannot be overlooked.

In the simplest approximation, disregarding the symmetry-lowering disorders of both static (inhomogeneous) and dynamic (homogeneous) origin, the excited-state manifold for the C_9-symmetric assembly of the 18 Bchl *a* sites features two non-degenerate (denoted as $k = 0$ and $k = 9$) and eight pairwise degenerate ($k = \pm1$, $k = \pm2$, . . . , $k = \pm8$) exciton states (Fig. 4). Because of the dimeric sub-structure of the Bchl *a* molecular chain imposed by the surrounding protein scaffold, these exciton states are divided into two so-called Davydov sub-groups, 9 states in each sub-group (Davydov 1971). In solid-state theory, the splitting between the Davydov sub-bands is considered a measure of the exciton coupling strength. Due to the circular assembly of the sites only the exciton states $k = 0$, $k = \pm1$, $k = \pm8$, and $k = 9$ can be photo-excited. Since the transition-dipole moments of the individual B850 pigments are oriented mainly in the plane of the ring, just a small part of the total oscillator strength is associated with the non-degenerate exciton states $k = 0$ and $k = 9$. Given the head-to-tail arrangement of the transition dipole moments of the adjacent Bchl *a* molecules nearly all the oscillator strength is concentrated in the $k = \pm1$ exciton states, being reflected in a strong electronic absorption band at about 850 nm. The upper exciton components, $k = \pm8$, carry less than

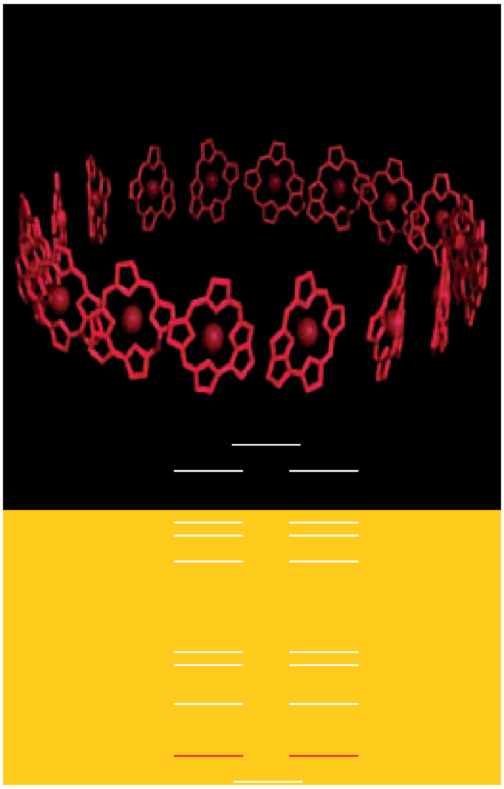


Figure 4.
Exciton level structure (white and red bars) for the 18 Bchl *a* pigments of the B850 arrange-ment shown on top. The two $k = \pm1$ exciton states, which possess nearly all the oscillator strength for the transitions to the ground state are highlighted with red. (Figure courtesy of D. Rutkauskas.)

3% of the total oscillator strength and give rise to very weak absorptions in the spectral range from 750 nm to 780 nm. Furthermore, in the systems with strong exciton coupling an additional contribution that determines the exact spectral position of the absorption band is given by the displacement energy (Davydov 1971). The latter energy accounts for the difference of the Coulomb interaction between neighbouring pigments in the electronically excited state and in the ground state. An *ab initio* estimation of the displacement energy for multi-chromophore systems is still a demanding theoretical task. From these considerations it follows that changes of the spectral position of the B800 band, in the first approximation, just reveal the interactions that involve site energies of the chromophores, whereas the origin of the B850 band spectral position change is more complex and reflects contributions from interactions that affect both the site energies and the coherent exciton interactions between the chromophores.

When suitably combined with the inevitable static energetic disorder, the coherent exciton theory briefly introduced above qualitatively explains all the major spectroscopic features of the LH2 light-harvesting complexes (Sundström et al. 1999; Van Amerongen et al. 2000; Cogdell et al. 2006). The only problem is that 'the numbers don't work'. Indeed, a detailed comparison of the absorption and emission spectra of ensembles of the LH2 complexes measured at low temperature with the predictions of the established disordered Frenkel exciton model reveals several serious discrepancies (Freiberg et al. 1999; Timpmann et al. 2001; Freiberg, Trinkunas 2009).

PHOTOSYNTHETIC EXCITON POLARONS IN SOFT PROTEIN ENVIRONMENT

Quantum coherence is all about phase relationships. The idealised exciton states in deformable lattices interact with the local vibrations of the molecular units themselves as well as with the collective phonons of the surroundings, leading (like in individual molecules albeit more complex ways) to dissipation of excitations as well as to vibronic and electron-phonon spectral structures, respectively. The same interactions, which increase with temperature, also govern decoherence and transport dynamics of excitons. To envision the latter effects, one could picture a moving exciton dragging behind a cloud of phonons, an exciton polaron. The thicker the cloud, the slower moves the entity exciton plus cloud. In the limiting scenario the exciton may stop moving altogether localising on a small region of the lattice. This phenomenon is called exciton self-trapping.

A strong ($S \geq 2$) exciton-phonon coupling in the fluorescence emission spectra of the B850 Bchl *a* aggregates has been discovered (Freiberg et al. 2003a, 2009; Timpmann et al. 2004b), several times greater than observed for monomeric chromophores in similar protein environment (Rätsep, Freiberg 2007). Visible consequences of such a strong interaction are broad multi-phonon sidebands, as demonstrated in Fig. 5. This is unexpected, since coupling for the exciton should be several times smaller than that for the respective localised excitation due to 'motional reduction' (Toyozawa 2003).

The Frenkel model as well as its various modifications such as the Redfield theory, which assume negligible exciton-lattice interactions cannot capture strong coupling effects. This is also the most obvious source of the above-noted inconsistencies between the Frenkel theory and experiment. A theoretical model involving dynamic exciton polaron formation in the spirit

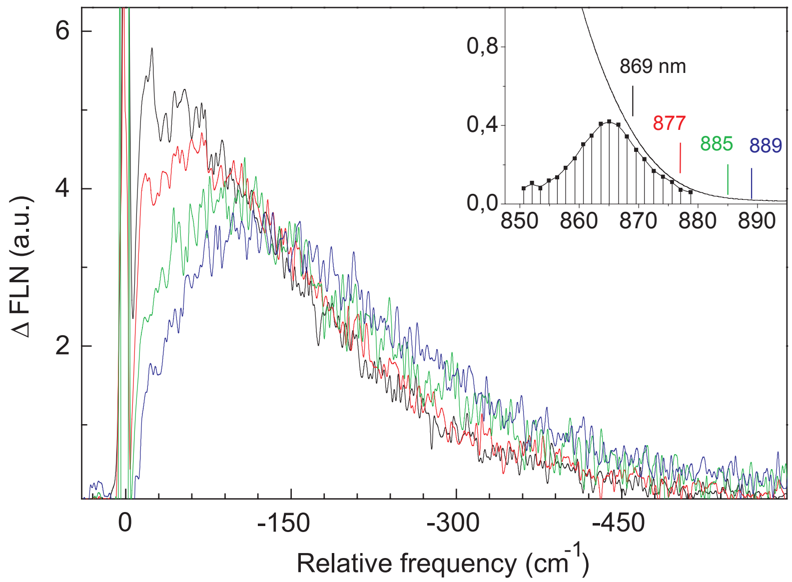


Figure 5.
ΔFLN spectra of isolated LH2 complexes from *Rhodobacter (Rb.) sphaeroides* at 4.5 K. For better analysis of the phonon sideband changes, the spectra measured at different excitation wavelengths (as indicated with the color-coded vertical lines in the inset) are shifted together at the ZPL origin and normalised by the sideband area. The narrow ZPL are cut off at a few percent intensity levels. The inset shows the inhomogeneous distribution function of the lowest-energy exciton states (drop lines) and the long-wavelength tale of the absorption spectrum (solid line).

of Holstein, Heeger and others has, therefore, been developed by us to explain the conflicting observations in case of the LH2 complexes (Freiberg et al. 2003b; Trinkunas, Freiberg 2006; Freiberg, Trinkunas 2009; Pajusalu et al. 2011). The fundamental nature of this model is that due to strong exciton-phonon interaction a structural reorganisation of the immediate environment is induced by the exciton, thereby lowering the exciton energy and causing its trapping. In one-dimensional electronic systems this kind of exciton self-trapping is taking place at any non-vanishing coupling (Sumi, Sumi 1994), being thus unavoidable in all real structures. The situation is rather different in higher dimensional organisations, where the free and self-trapped excitons are separated by an energetic barrier. Since the B850 ring can be considered with certain reservations a one-dimensional crystal made from the Bchl *a* molecules that strongly couple with their environment, exciton polarons rather than Frenkel excitons are generic photoexcitations of the LH2 antenna complex.

A configuration coordinate diagram (Fig. 6) illustrates this issue. Shown in the figure are the adiabatic potential energy surfaces that correspond to the ground state (GS, gray), the self-trapped exciton state (STE, green), and the three lowest ($k = 0, \pm 1$) light-absorbing Frenkel exciton states (Exciton, blue of different shades). The transitions between the ground and excited states are linearly coupled via two interaction coordinates. The horizontal Q_2 coordinate accounts for the diagonal (energetic) coupling, while the out-of-plain Q_1 coordinate, for the non-diagonal (structural) interaction. The straight and curly arrows represent the optical transitions and relaxation channels, respectively. The reorganisation energy related to the diagonal exciton-phonon bath couplings in the LH2 complexes from *Rps. acidophila* is 285 cm^{-1}. The same parameter for the off-diagonal couplings is 73 cm^{-1} (Pajusalu et al. 2011).

Since just the ground state and the Frenkel exciton states from one side and the self-trapped exciton state and the ground state from another side are directly coupled by photon transitions, it should be clear from Fig. 6 that the absorption and emission spectra in such system deliver different information about the excited states of the matter. While the absorption spectra characterise the exciton immediately after its creation, the emission spectra describe the exciton behaviour immediately before its annihilation (Toyozawa 2003). This situation is qualitatively different from that for localised impurities in solids.

Fig. 7 demonstrates major effects of the polaronic coupling on the fluorescence spectral response of the LH2 antenna complexes. Compared in this figure at two temperatures are the simulated fluorescence emission spectra of LH2 complexes using either a Frenkel exciton model or the exciton polaron

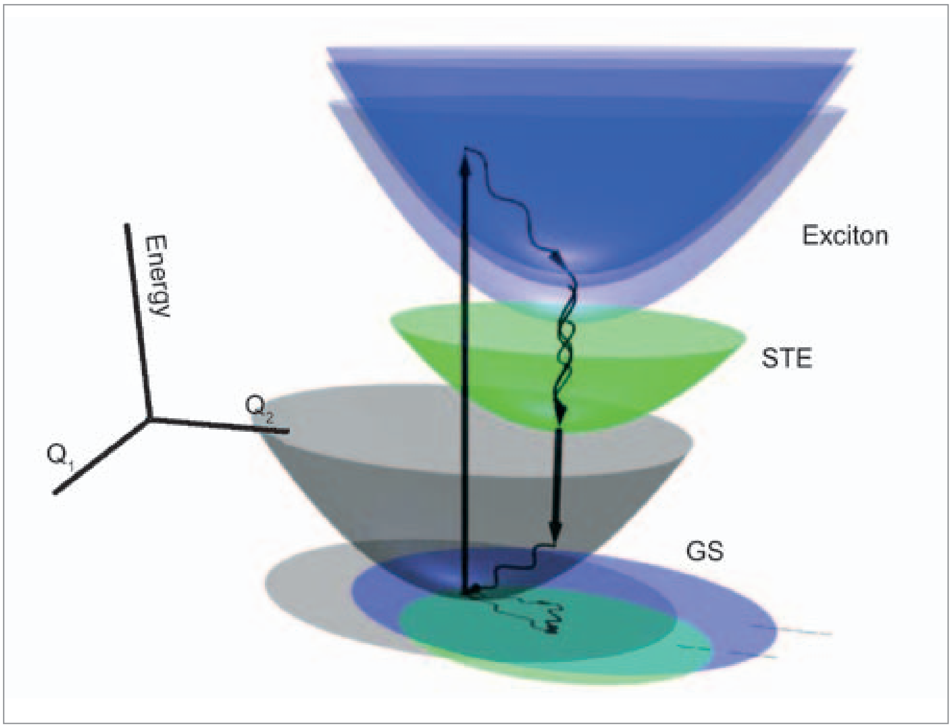


Figure 6.
The configuration coordinate model with two (off-diagonal Q_1 and diagonal Q_2) interaction coordinates for the excitons in the B850 aggregate of LH2. The coloured surfaces counted from the bottom of the figure towards the growing potential energy (not in scale) correspond to the ground state (GS, gray), self-trapped exciton (STE, green), and Frenkel exciton states $k = 0$ and $k = \pm 1$ (Exciton, blue of different shades). Upward and downward bold arrows indicate the photon absorption and fluorescence emission transitions, respectively. The curly arrows designate relaxation channels.

model. As one can see, the spectra according to the two models significantly deviate from each other at all temperatures both in terms of the fluorescence band position and its shape. The exciton polaron spectrum is generally broader as well as redder. These differences grow larger with temperature due to thermal occupation of the higher-energy states. Only the exciton polaron picture is in agreement with experiment (Pajusalu et al. 2011).

We have noted above (Fig. 2) that the Huang-Rhys factors (consequently, the homogeneous spectral shapes) that correspond to the impurity centres of Bchl *a* depend on excitation energy. As can be seen in Fig. 5, similar behaviour is characteristic for the LH2 antenna exciton polarons as

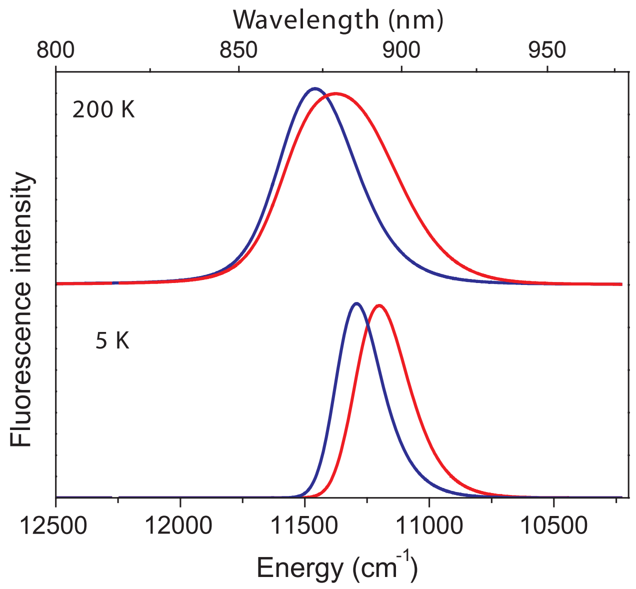


Figure 7.
Simulated fluorescence emission spectra for LH2 complexes from *Rps. acidophila* according to Frenkel exciton (blue line) and exciton polaron (red line) models. The spectra calculated at two indicated temperatures are normalised by peak intensity. (Figure adapted from Pajusalu, Rätsep et al. 2011.)

well (Timpmann et al. 2004b; Freiberg et al. 2009). With increasing excitation wavelength the peak of the phonon sideband in the emission spectrum shifts toward higher energies. The shift is accompanied with the sideband broadening and loss of its fine structure. All these changes point to progressively increasing exciton-phonon coupling toward lower-energy systems. Quantitative evaluation of the exciton-phonon coupling strength supports this conclusion; the Huang-Rhys factor grows from ~1.9 to ~3.0 between 869 and 889 nm. The excitations in the sub-group of the LH2 complexes that absorb high-energy light thus appear more exciton-like, while they are rather exciton polaron/self-trapped exciton-type when absorbing lower-energy light. In the recent single molecule experiments (paper under preparation) the fluorescence excitation and emission spectra from individual antenna complexes were measured. These experiments notably confirm the above insight.

PHOTOSYNTHETIC EXCITONS AT FUNCTIONAL TEMPERATURES

As already noted in Introduction, the exciton concept in photosynthesis heavily relies on spectroscopic measurements carried out at low temperatures. Except for circular dichroism data, which are sometimes demanding to interpret, there is not much experimental evidence for the relevance of excitons in functional photosynthesis that operates at ambient temperatures. Therefore, we shall next dwell on the following fundamental question, "Do the excitons in the photosynthetic light-harvesting complexes survive functional temperatures?" The essence of this problem is that temperature by virtue of inducing noise into the system tends to smear out the firm relationships between the phases of the wavefunctions that correspond to photoexcitations on individual molecular sites of the light-harvesting complex. Broadly, it might be expected that at $k_B T >> V$ (where $k_B T$ is the average thermal energy of environmental vibrations, and V is the nearest-neighbour exciton coupling energy) the phases of the site excitations may become randomised by thermal noise and the excitons cease to exist. More rigorously, however, V should be compared with the thermal energy due to the coupling to the environment (reorganisation energy λ) rather than with $k_B T$.

To check what really happens, one would like to investigate the bandwidth of the B850 excitons as a function of temperature over the whole range from liquid He to ambient temperatures. For that, it would be sufficient to measure energetic separation of the two light-absorbing exciton states ($k = \pm 1$ and $k = \pm 8$), which can be considered as markers of the exciton band edges. Yet the $k = \pm 8$ absorption is so weak that it is totally covered by the much stronger absorption tail of the B800 molecules. This obstruction can be overcome by exploiting the polarised fluorescence excitation spectroscopy (Timpmann et al. 2004a; Trinkunas, Freiberg 2006). Since the transition dipole moments within the pairs $k = \pm 1$ and $k = \pm 8$ are mutually orthogonal, there exists an energy for each pair of states where they are excited with equal probability. The fluorescence anisotropy will then feature a minimum at those energies, effectively designating spectral positions of the respective pairs of the exciton states (Pajusalu et al. 2011). The fluorescence anisotropy, r, is defined as: $r = (I_{vv} - I_{vh})/(I_{vv} + 2I_{vh})$, where I_{vv} and I_{vh} are the emission intensities that are polarised parallel (vv) and perpendicular (vh) with respect to the vertical orientation of the electric field vector of the linearly polarised excitation light.

The fluorescence anisotropy excitation spectra of the LH2 complexes from *Rps. acidophila* at the low (4.5 K) and high (263 K) temperature limits are shown in Fig. 8. At both temperatures the anisotropy is low at short wavelengths and rises across the B850 absorption band toward the theoretical limit of 0.4 at the long-wavelength boundary of the spectra. The most striking feature of the anisotropy curves, however, is the two distinct minima, one visible between 760 and 770 nm, another around 860 nm. The minima at high temperatures are similar albeit much shallower than at low temperatures. The dip wavelengths very well correlate with the expected positions of the $k = \pm 1$ and $k = \pm 8$ exciton states, thus marking the searched for edges of the B850 exciton state manifold. The energetic separation, ΔE, of the dips is 1447 cm^{-1} at 4.5 K and 1259 cm^{-1} at 263 K. Increasing temperature from deep cryogenic to physiological temperatures thus results in only ~ 13% narrowing of the exciton bandwidth. Providing definitive evidence for survival of excitons at

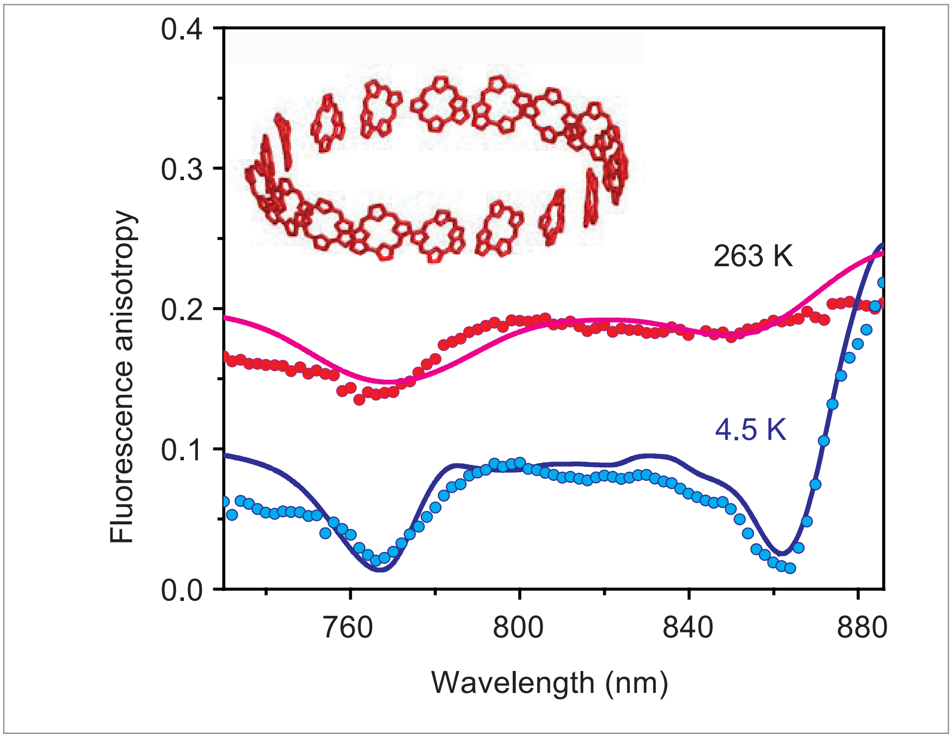


Figure 8.
Effect of temperature on the fluorescence anisotropy excitation spectra of the LH2 light-harvesting complexes from *Rps. acidophila*. Scattered dots represent experimental data at the two indicated temperatures, while their theoretical fits are drawn with continuous lines. The two curves are vertically shifted with respect to each other for better observation. The inset shows the B850 pigment ring supporting coherent excitons. (Figure adapted from Pajusalu, Rätsep et al. 2011.)

functional temperatures, the present data strongly suggest that the collective coherent electronic excitations may indeed play profound role in the functional photosynthetic light-harvesting process.

The single most obvious explanation for the observed robustness of the light-harvesting excitons against temperature is the strong intermolecular coupling in the B850 ring. A straightforward estimate of the nearest-neighbour coupling energy from the exciton bandwidth yields $V \approx \Delta E/4 = 362$ cm^{1}, while a more elaborate theory (Pajusalu et al. 2011) returns $V = 374$ cm^{-1}. These numbers are greater than $k_B T$ at ambient temperature amounting to ~205 cm^{-1}. However, the reorganisation energy, $\lambda = 358$ cm^{-1}, is bigger and practically equals V. Proposals have been made that the protein scaffold is someway special in protecting electronic state coherences. This is unlikely. A more sound idea came around only recently (Galve et al. 2010).

According to this study the common relationship between V and λ is only valid if the system is in thermal equilibrium with its environment. If not, the temperature no longer provides the relevant energy scale against which to compare quantum behaviour of the system. This role is taken over by an effective temperature, which can be much lower than the absolute temperature. Similar magnitude of electronic and exciton-lattice couplings imply that exciton processes in light-harvesting antennas might indeed take place in non-equilibrium conditions, at least partially explaining the above conundrum.

Two groups have recently reported recording the quantum coherence in photosynthetic systems at ambient temperature, one group dealing with the bacterial LH complexes similar to those discussed in this work (Panitchayangkoon et al. 2010), another with the cryptophyte algae (Collini et al. 2010). Using femtosecond pulse laser excitation electronic coherences that lasted a few hundred femtoseconds were demonstrated, proving an old theoretical prediction (Nedbal, Szöcs 1986). It is still not clear, however, whether these oscillations are long enough to be useful for photosynthetic energy transfer, which predominantly takes place in picosecond range. Our experiments detect exciton eigenstates, which by definition are stationary states. This is a definitive distinction. Since steady-state exciton states are present, there must be no doubt that quantum coherence is involved with functional photosynthesis.

NEW LIGHT ON LIGHT-HARVESTING: COHERENT STATES IN NATIVE PHOTOSYNTHETIC MEMBRANES

Photosynthetic chromatophore vesicles of nearly spherical shape of 50-60 nm diameter found in some purple bacteria (Fig. 9) constitute one of the simplest light-harvesting membrane systems in nature. The overall architecture of the chromatophore vesicles and the vesicle function remain poorly understood despite structural information being available on individual constituent proteins. The knowledge about relative spatial arrangement of the LH and RC complexes has historically mainly obtained by spectroscopic methods, more recently by atomic force microscopy, and notably in these days

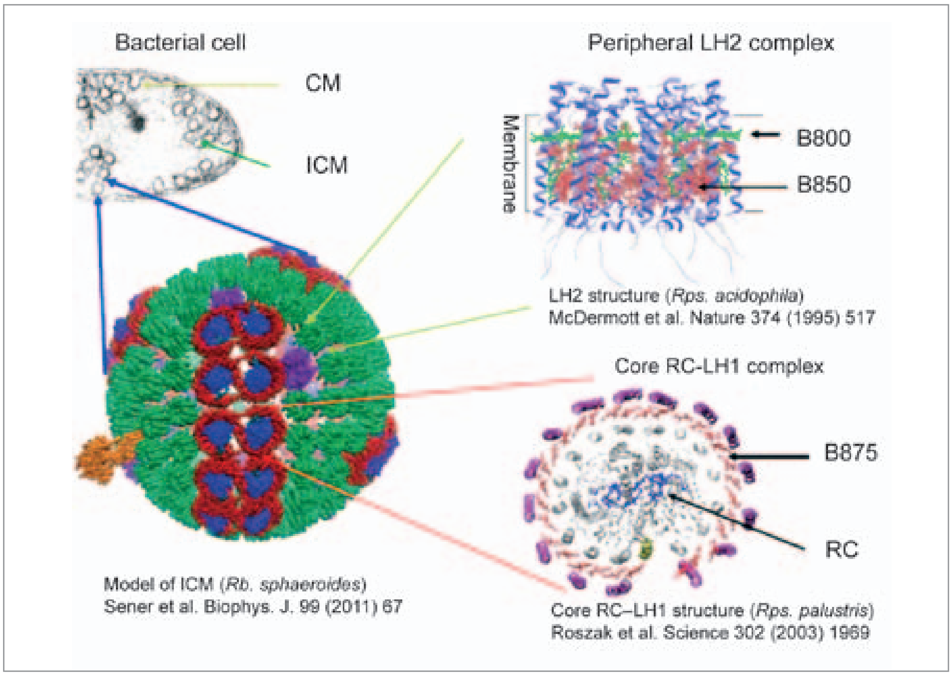


Figure 9.
In *Rb. sphaeroides*, the photosynthetic apparatus is organised into 50-60 nm diameter intracytoplasmic chromatophore vesicles (a spherical structure shown on the lower left corner), which are mainly populated by the peripheral LH2 light-harvesting complexes (drawn in green), along with a smaller number of core LH1 (red) and RC complexes (blue). The core antennas may appear in variable shapes, encircling either single (as in *Rps. palustris*) or double (wild type *Rb. sphaeroides*) RC. CM denotes the cytoplascmic membrane; ICM, the intracytoplasmic membrane; B875, the assembly of Bchl *a* molecules in LH1.

computationally. An all-atom computational structure model for an entire chromatophore vesicle of *Rb. sphaeroides* was lately presented in (Sener et al. 2010). The model improved upon earlier models by taking into account the stoichiometry of the core RC-LH1 and peripheral LH2 complexes in intact vesicles, as well as the curvature-inducing proper-ties of the dimeric core complex. The absorption spectrum of low-light-adapted vesicles was shown to correspond to a LH2 to RC ratio of 3 : 1. Confirmed also by massive computations of exciton properties was the hitherto assumed principle that the success of photosynthesis depends on ultrafast events, in which up to hundreds of membrane proteins are cooperating in a multi-step process of solar excitation energy funnelling to the RC (Freiberg 1995).

Quantum coherence requires great deal of order to manifest measurably. The biological membranes carrying photosynthesis machinery, on contrary, are highly disordered and dynamic structures. It wouldn't be too surprising if the coherent quantum effects apparent in individual light-harvesting complexes do not materialise in the native membranes. Therefore, we next assess, again using fluorescence anisotropy excitation spectroscopy, to what extent the coherent exciton concept holds in fully developed photosynthetic membranes.

Fig. 10 compares the anisotropy excitation spectrum of isolated LH2 complexes with this for the native membrane of *Rps. acidophila* complete with LH and RC complexes. In order to disentangle individual contributions to the anisotropy from the LH2 and LH1 complexes, the emission were detected either around 890-910 nm (the predominant emission range of LH2) or 920-950 nm (LH1). Thereby, clear signals corresponding to the exciton band edges of the LH2 and LH1 complexes have been observed.

Most remarkably, however, the exciton bandwidths of the LH2 complexes in membranes and in detergent-isolated complexes appear the same, meaning that excitons in these two rather different environments are basically identical. This is very good news for protein crystallographers, indicating that the crystals of detergent-isolated membrane proteins may indeed represent native membrane conformation and provide relevant structural information. Notable also is that the high-energy limits of the exciton state manifolds in the LH1 and LH2 antennas are tuned together, so that the broader LH1 exciton band (ΔE = 1951 cm^{-1}) totally engulfs the narrower LH2 band. Similar situation is observed in the membranes of *Rb. sphaeroides*. The perfect overlap of the high-energy exciton states of the peripheral and core antennas must be advantageous for efficient tunnelling of the excitation energy toward the RC. This is a wonderful example of optimisation of solar light harnessing by natural selection!

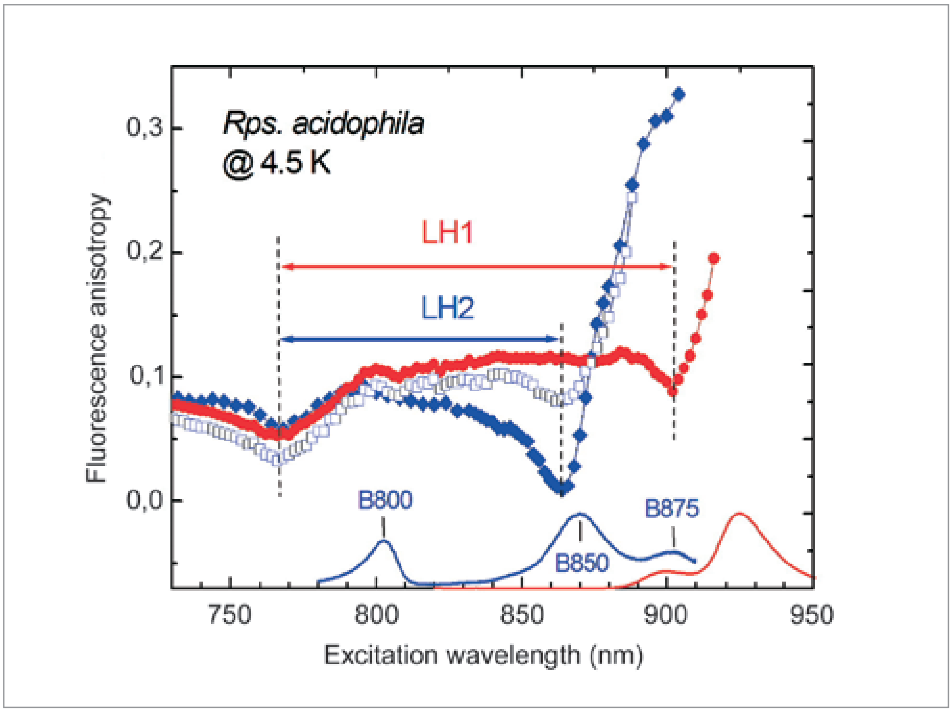


Figure 10.
Fluorescence anisotropy excitation spectra for native membranes (red dots and open blue squares) and isolated LH2 complexes (filled blue diamonds) from *Rps. acidophila* at 4.5 K. The reference fluorescence emission and fluorescence excitation spectra shown at the bottom of the figure are drawn with red and blue lines, correspondingly. The two peaks of the fluorescence spectrum at ~900 and ~925 nm relate, respectively, to the B850 and B875 bands in the excitation spectrum. The double arrows mark the bandwidths for the B850 and B875 exciton manifolds.

WHAT IS THE ANSWER?

The single essential quality that distinguishes quantum theory from classical mechanics is the phase of the wavefunction. The validation of stationary coherent photosynthetic excitations at ambient temperatures is yet another piece of evidence that helps to bridge the gap between quantum physics and biology. Other more or less established examples are vision, bioluminescence, tunnelling of electrons and protons in enzymatic reactions, magnetoreception of birds and animals, and weak van der Waals interactions between the molecules. The ubiquity of the quantum biological phenomena appears to suggest that biological systems might exploit the wave properties of matter for optimisation of their functionality. This hint should be taken seriously in expectation of a future society powered mostly by solar energy.

Do we indeed live in the quantum world? This question dates back to the influential work of Schrödinger, one of the founders of quantum mechanics, for more than half a century ago (Schrödinger 1944). Discussion goes on (Davydov 1982; Abbott et al. 2008) but there is little doubt that at the end this as well as the title question should be answered affirmatively. This is not to say that all problems have been solved. The major challenge still is to understand how our everyday classical world emerges from the counterintuitive realm of quantum mechanics (Ball 2008). A non-trivial revelation of the recent studies is that the quantum-classical transition is not so much a matter of size, but of time. The quantum laws manage equally well the phenomena at the scales of both the very small such as atoms and molecules and the very large such as black holes and, according to modern cosmology (Carroll 2011), even the whole universe.

Acknowledgements

There are many co-workers, collaborators, and simply friends in science around the world who directly (as can be seen from the reference list) or less so, just by casual conversations and interactions anywhere inspired this work. I am grateful to M.Rätsep and M.Pajusalu for critical reading the manuscript and preparing Fig. 1 and 6, respectively. Fig. 5 was designed by K.Timpmann. To Niina, for love and care.

REFERENCES

Abbott, D., et al. (eds.) 2008. Quantum Aspects of Life. Imperial College Press, London.

Avarmaa, R. A., Rebane, K. K. 1985. High-resolution optical spectra of chlorophyll molecules. Spectrochimica Acta A, 41, 1365-1380.

Ball, P. 2008. Quantum all the way. Nature, 453, 22-25.

Blankenship, R. E. 2002. Molecular Mechanisms of Photosynthesis. Blackwell Science, Oxford.

Carroll, S. 2011. From Eternity to Here. Oneworld Publications, Oxford.

Cogdell, R. J., et al. 2006. The architecture and function of the light-harvesting apparatus of purple bacteria: from single molecules to in vivo membranes. Quarterly Review of Biophysics, 39, 227-324.

Collini, E., et al. 2010. Coherently wired light-harvesting in photosynthetic marine algae at ambient temperature. Nature, 463, 644-648.

Davydov, A. S. 1971. Theory of Molecular Excitons. Plenum Press, New York.

Davydov, A. S. 1982. Biology and Quantum Mechanics. Pergamon, New York.

Freiberg, A. 1995. Coupling of antennas to reaction centres. Blankenship, R. E., Madigan, M. T., Bauer, C. E. (eds.). Anoxygenic Photosynthetic Bacteria. Kluwer Academic Publishers, Dordrecht, 2, 385-398.

Freiberg, A., et al. 1999. Disordered exciton analysis of linear and nonlinear absorption spectra of antenna bacteriochlorophyll aggregates: LH2-only mutant chromatophores of *Rhodobacter sphaeroides* at 8 K under spectrally selective excitation. Journal of Physical Chemistry B, 103, 45, 10032-10041.

Freiberg, A., et al. 2003a. Self-trapped excitons in circular bacteriochlorophyll antenna complexes. Journal of Luminescence, 102-103, 363-368.

Freiberg, A., et al. 2003b. Self-trapped excitons in LH2 antenna complexes between 5 K and ambient temperature. Journal of Physical Chemistry B, 107, 11510-11519.

Freiberg, A., et al. 2009. Excitonic polarons in quasi-one-dimensional LH1 and LH2 bacteriochlorophyll a antenna aggregates from photosynthetic bacteria: A wavelength-dependent selective spectroscopy study. Chemical Physics, 357, 102-112.

Freiberg, A., Trinkunas, G. 2009. Unraveling the hidden nature of antenna excitations. Laisk, A., Nedbal, L., Govindjee (eds.) Photosynthesis in Silico. Understanding Complexity From Molecules to Ecosystems. Springer, Heidelberg, 55-82.

Galve, F., et al. 2010. Bringing entanglement to the high temperature limit. Physical Review Letters, 105, 180501.

Kasha, M. 1963. Energy transfer mechanisms and the molecular exciton model for molecular aggregates. Radiation Research, 20, 55-71.

McDermott, G., et al. 1995. Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. Nature, 374, 6522, 517-521.

Nedbal, L., Szöcs, V. 1986. How long does excitonic motion in the photosynthetic unit remain coherent? Journal of Theoretical Biology, 120, 411-418.

Pajusalu, M., et al. 2011. Davydov splitting of excitons in cyclic bacteriochlorophyll a nanoaggregates of bacterial light-harvesting complexes between 4.5 and 263 K. European Journal of Chemical Physics and Physical Chemistry, 12, 634-644.

Panitchayangkoon, G., et al. 2010. Long-lived quantum coherence in photosynthetic complexes at physiological temperature. Proceedings of the National Academy of Sciences of the USA, 107, 12766-12770.

Rebane, K. K. 1970. Impurity Spectra of Solids. Plenum Press, New York.

Renge, I., et al. 2011. Intermolecular repulsive-dispersive potentials explain properties of impurity spectra in soft solids. Journal of Luminescence, 131, 262-265.

Rätsep, M., et al. 2009a. Mirror symmetry and vibrational structure in optical spectra of chlorophyll a. Journal of Chemical Physics, 130, 194501.

Rätsep, M., et al. 2009b. Wavelength-dependent electron–phonon coupling in impurity glasses. Chemical Physics Letters, 479, 140-143.

Rätsep, M., et al. 2011. Demonstration and interpretation of significant asymmetry in the low-resolution and high-resolution Q_y fluorescence and absorption spectra of bacteriochlorophyll a. Journal of Chemical Physics, 134, 024506.

Rätsep, M., Freiberg, A. 2003. Resonant emission from the B870 exciton state and electron-phonon coupling in the LH2 antenna chromoprotein. Chemical Physics Letters, 377, 371-376.

Rätsep, M., Freiberg, A. 2007. Electron-phonon and vibronic couplings in the FMO bacteriochlorophyll a antenna complex studied by difference fluorescence line narrowing. Journal of Luminescence, 127, 251-259.

Schrödinger, E. 1944. What is Life? Cambridge University Press, Cambridge.

Sener, M., et al. 2010. Photosynthetic vesicle architecture and constraints on efficient energy transfer. Biophysical Journal, 99, 67-75.

Sumi, H., Sumi, A. 1994. Dimensionality dependence in self-trapping of excitons. Journal of the Physical Society of Japan, 63, 637-657.

Sundström, V., et al. 1999. Photosynthetic light-harvesting: Reconciling dynamics and structure of purple bacterial LH2 reveals function of photosynthetic unit. The Journal of Physical Chemistry B, 103, 13, 2327-2346.

Timpmann, K., et al. 2001. Exciton self-trapping in one-dimensional photosynthetic antennas. The Journal of Physical Chemistry B, 105, 12223-12225.

Timpmann, K., et al. 2004a. Bandwidth of excitons in LH2 bacterial antenna chromoproteins. Chemical Physical Letters, 398, 384-388.

Timpmann, K., et al. 2004b. Emitting excitonic polaron states in core LH1 and peripheral LH2 bacterial light-harvesting complexes. The Journal of Physical Chemistry B, 108, 10581-10588.

Toyozawa, Y. 2003. Optical Processes in Solids. Cambridge University Press, Cambridge.

Trinkunas, G., Freiberg, A. 2006. A disordered polaron model for polarized fluorescence excitation spectra of LH1 and LH2 bacteriochlorophyll antenna aggregates. Journal of Luminescence, 119-120, 105-110.

Van Amerongen, H., et al. 2000. Photosynthetic Excitons. World Scientific, Singapore.

Zazubovich, V., et al. 2001. Bacteriochlorophyll a Frank-Condon factors for the S0 S1(Q_y) transition. The Journal of Physical Chemistry B, 105, 12410-12417.

Zewail, A. H. (ed.) 2008. Physical Biology. From Atoms to Medicine. Imperial College Press, Singapore.

ENERGY MATERIALS

E. Mellikov, A. Öpik, E. Lust, A. Jänes

Insecurity of power supplies and pollution of air environment in megacities and industrial regions are among the greatest problems for mankind at the moment and ways out of those problems need to be found within only a few decades. Estonia, being a member of the European Union, has and will have some problems connected with the fulfilling of the EU regulations, connected with the energetically ineffective shale-oil industry and exploitation of the thermo-mechanical cycle based technology to produce electricity in the future. Taking into account the very low upper heating value for oil-shale, there are no elegant solutions based on the old historical thermodynamically limited electricity production conception applied in AS Eesti Energia. In addition, the residual heat as co-product of electricity cannot be used effectively. There is no doubt that solar cells, fuel cells, Li-ion batteries (LIB), electrical double layer capacitors (EDLC) and hybrid supercapacitors (HSC) based on the designed nano/microstructural materials and combined with wind turbines for hybrid energy storage/generation systems will be the most perspective environmentally 'green' energy systems in the future. The fuel cells, especially solid oxide fuel cells (SOFC) are the energy conversion systems with 2…2.5 times higher efficiency compared with traditional internal combustion engines. In order to reduce the application of the fossil energy sources, a continuous growth of PV by at least 20% per year will be needed during the coming decades that will lead to the installation of at least ten TW_p before 2050 (Vision 2004).

In this paper, results of the research in modern energy concentration/generation/accumulation devices at Estonian Universities will be presented.

Development of new materials for solar cells

At present, the photovoltaic industry primarily uses polycrystalline or single-crystalline Si wafers, which have a wafer thickness in the range of 100-300 µm. The Si solar cell technology is very expensive not only in terms of money but also in terms of energy input (Luque, Hegedus 2006). The technical parameters of thin film solar cells are worse than those of crystalline solar cells, but thin film technologies are, in general, much cheaper than monocrystalline technologies. The most widely currently used thin film solar cell materials, CdTe and $CuInGaSe_2$, $CuInGaS_2$ (CIGS), contain resource-limited elements (Te and In) (Feltrin, Freindlich 2008). To realise advantages of thin film technologies in the future new SC materials that are abundant but much cheaper to produce than Si are needed. The most promising new materials are Cu_2ZnSnS_4, $Cu_2ZnSnSe_4$ and $Cu_2ZnSn(S_xSe_{1-x})_4$ (CZTS) with their abundant and nontoxic constituents (Ito, Nakazawa 1988). CZTS materials have high absorption coefficient and ability to undergo band gap engineering through alloy formation. Thin film technologies here used are similar to those used for CIGS. These are mainly vacuum methods based on the deposition of metallic (Cu, Zn, Sn) constituent films with different sequences with the following selenisation (Wibowo et al. 2007). The efficiency of developed CZTS SC-s is much lower than theoretically possible. However, Mitzi et al achieved a 9.6% efficiency confirmed by NREL (Mitzi et al. 2008). All this is well below the reported best efficiencies for CIGS SC-s (up to 20% in the lab but it is encouraging as a starting point).

Perspective alternative to prepare SC-s is the use of powder materials (Mellikov et al. 2007). Powder technologies are the cheapest technologies but powders have not found widespread use in PV yet due to lower parameters of SC-s on their base. It has been shown by us that isothermal recrystallisation of initial powders in molten fluxes appears to be a relatively simple, inexpensive method to produce CZTS powders with an improved crystal structure (Altosaar, Mellikov 2000). In monograin layer SC-s developed by us the powders replace monocrystalline wafers or thin films. The resulting separation of materials formation from the module fabrication is an additional very important advantage.

Aim of the research was the investigation of formation of absorber materials in different thermal and chemical processes and the development of low-cost materials and PV cells on their basis. To reach these objectives a multidisciplinary approach was used, where expertise from materials science, condensed matter physics, chemistry and nanotechnology are merged to provide skills in materials synthesis.

EXPERIMENTAL RESULTS

Method

The research in PV materials and SC-s bases on synthesis of CZTS either as monograins, or as thin films. A suitable choice of stoichiometry of CZTS, or the possibility to grow nanostructured CZTS films, provides us the flexibility to adapt the band gap of adsorber as well as the band positions relative to a contacting layer, which is ultimately designed for increase of PV device efficiency.

Monograin powder materials were synthesised from metal binaries (CuSe(S), ZnSe(S), SnSe(S) and elemental Se or S) in a molten flux using an isothermal recrystallisation process. Monograin layer (MGL) SC-s (graphite/CZTSSe/CdS/ZnO) production is described in detail in (Mellikov et al. 2008a). In use of thin films the precursor films with a different sequence of consistent binary chalcogenides (CuSe, ZnSe, SnSe) or metals (Cu, Zn, Sn) were deposited onto Mo covered glass and annealed with elemental Se vapour in isothermal sealed quartz ampoules (Volobujeva et al. 2009a). The CdSe nanostructured films were electrodeposited on indiumtinoxide (ITO) coated glass substrates from aqueous electrolytic bath.

The evolution of the crystal shape and morphology of materials were studied using a high-resolution scanning electron microscope (SEM) Zeiss ULTRA55. The chemical composition was determined using an energy dispersive X-ray analysis (EDX). Bulk structure and phase compositions were studied using X-ray diffraction (XRD) and Raman spectroscopy. For photoluminescence (PL) measurements (10 K), the 441 nm He-Cd laser line was used. For the temperature dependent (T = 10 – 300 K) quantum efficiency measurements SC-s were mounted in a closed-cycle He cryostat. The UV-Vis transmission and reflection were measured in the wavelength region from 250 to 1400 nm. The thickness and refractive indices were obtained from specular reflectance measurements.

Fundamental studies

Fundamental studies are concentrated on the opto-electronic properties of various chalcopyrite and kesterite compounds. Main effort has been put to PL studies of chalcopyrites such as $CuInSe_2$, $CuInS_2$, $CuGaSe_2$, $CuGaTe_2$, $CuInTe_2$, and kesterites $Cu_2ZnSnSe_4$ and Cu_2ZnSnS_4 (Altosaar et al. 2008). The PL spectra of all these materials had shapes from broad asymmetric bands, which are characteristic of heavily doped semiconductors. Recombination related to deep donor-acceptor pairs, Fano-type resonances, excitons bound to an isoelectronic defect have also been observed. Based on the temperature dependencies of the PL bands, a correction was made to the theoretical temperature dependence of the PL intensity (Krustok et al. 1997).

Additional information about the electronically active defects in the absorber was obtained from admittance spectroscopy. From capacitance measurements, information about the charge carrier concentration, mobility of the carriers, defect activation energies, bulk and interface defects etc. were obtained. The vibrational properties and the phase composition of the materials were studied by Raman spectroscopy. We determined the formation of the ordered vacancy compounds $CuGa_3Se_5$ and $CuGa_5Se_8$ in the Cu-Ga-Se system

(Grossberg et al. 2009). We were the first to publish Raman spectra of kesterites $Cu_2ZnSnSe_4$ and Cu_2ZnSnS_4 and their solid solutions $Cu_2ZnSn(Se_xS_{1-x})_4$ (Altosaar et al. 2008). Currently, a challenge for researches is the phase analysis of the kesterites. Unlike of chalcopyrites, it is rather difficult to grow a single phase kesterite materials due to very small chemical potential region for stable single phase material. We have proposed a novel method to study spatial potential fluctuations in compensated materials (Krustok et al. 2010). The method is based on the analysis of the temperature dependence of quantum efficiency curves in SC-s (Fig. 1).

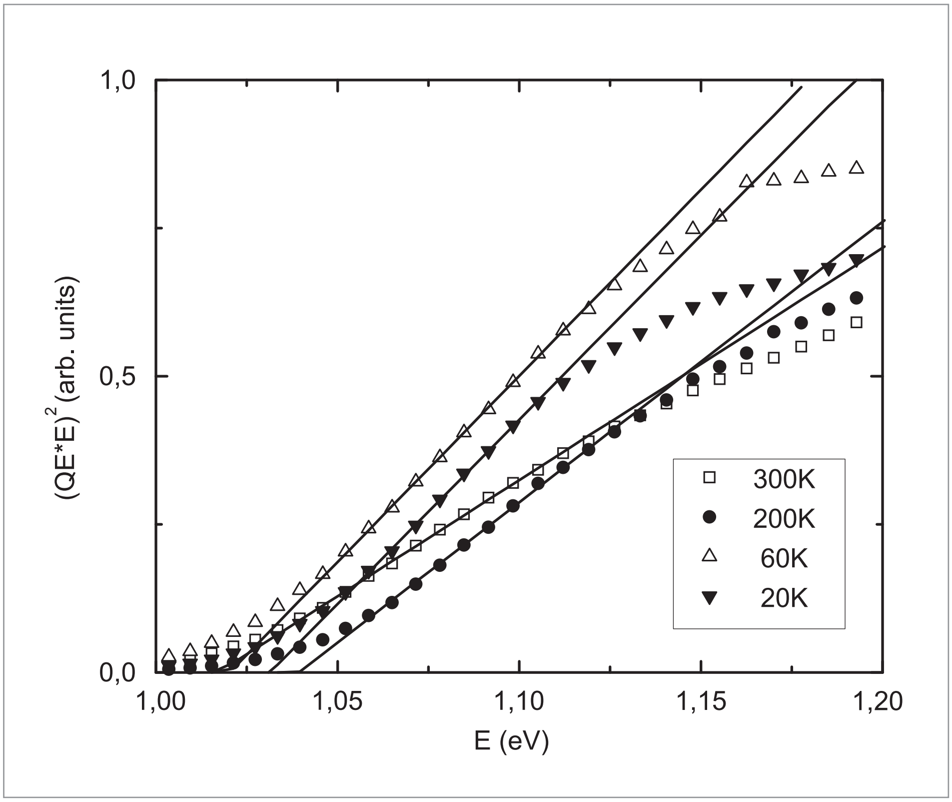


Figure 1.
Normalised QE curves versus photon energy measured at different temperatures (Krustok et al. 2010).

Last, but not least, modulation spectroscopy methods (photoreflectance, electroreflectance) were also used to gain information about the energy band structure of the developed materials.

Monograin powders and layers

The structural, morphological, electrical and optical properties of $Cu_2ZnSn(S,Se)_4$ materials depend strongly on their composition and on the additional thermal and chemical treatments. The composition of product materials depends on the initial Zn to Sn ratio. At the same time the content of chalcogens (S + Se) in the materials does not depend on the initial Zn/Sn ratio (Fig. 2) (Mellikov et al. 2010). Varying the S/Se ratio in the material allows one to control the quantum efficiency of the $Cu_2ZnSn(S,Se)_4$ SC-s. Current–voltage dependences of $Cu_2ZnSn(S,Se)_4$ MGL SC-s show V_{oc} values of over 700 mV, fill factors of up to 65%, and short circuit currents of up to 20 mA/cm^2. Efficiencies of SC-s of up to 7.3% were determined.

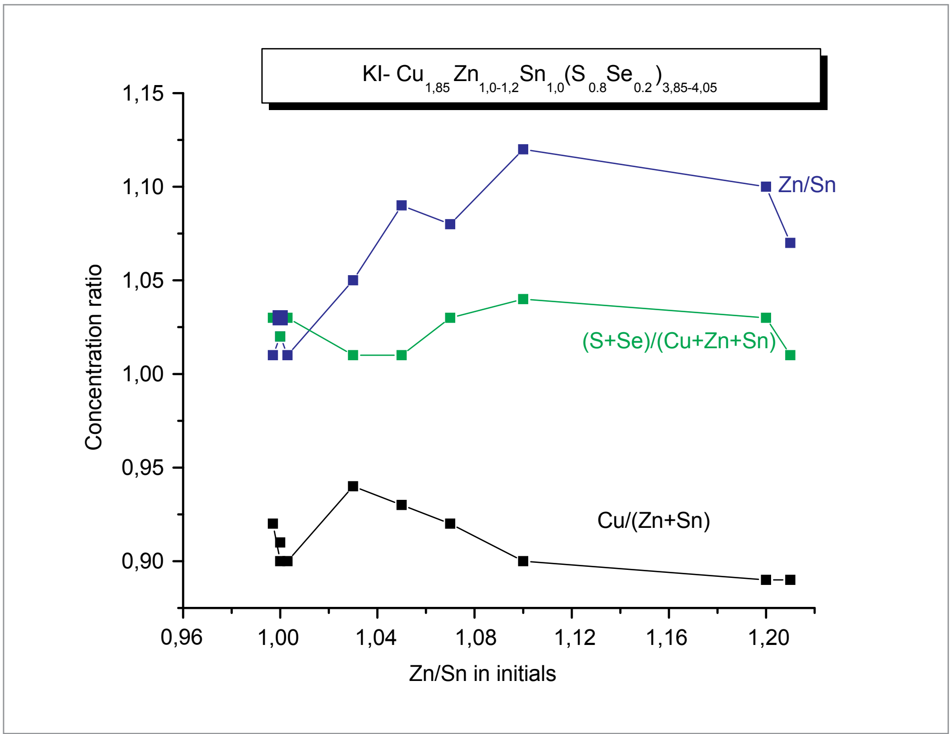


Figure 2.
The influence of initial Zn/Sn on the composition of materials (Mellikov et al. 2010).

Formation of thin films of $Cu_2ZnSn(Se)_4$

Precursor Sn-Zn-Cu films exhibit a well-formed 'mesa-like' structure of the surface in which larger crystals are located on a 'small-crystalline' valley (Mellikov et al. 2008b). For films with another sequence of metallic layers the mesa-like structure is not exposed, and well-formed flat precursor layers were produced replacing metallic layers with the Cu/Sn alloy layer. The use of stacked binaries as the precursors present us a challenge of avoiding Se treatment or to replace Se atmosphere with an inert one as the molecular ratio of the precursor layers was same as in stoichiometric $Cu_2ZnSnSe_4$. The sequence of binary layers ZnSe-SnSe-CuSe was used in our experiments. Our results indicate that the morphology of the deposited CuSe constituent films depends strongly on the speed of the deposition of CuSe and the substrate temperature and influences the path and kinetics of formation of CZTSe films in following selenisation process.

Regularities of selenisation of different metallic and binary precursor films appear to be quite similar. Our results indicate that in use of metallic precursors the selenisation of Cu is dominating in any sequence of layers and binary Cu selenides in different compositions are formed at low temperatures on the surface of films (Volobujeva et al. 2009b; Volobujeva 2010). The films selenised at low temperatures contain CuSe and agglomerated particles of a mixture of Cu_2SnSe_3 and $Cu_2ZnSnSe_4$. At the same time selenisation of metallic and binary stacked layers at temperatures higher than 420 °C results in multi-phase films consisting of $Cu_2ZnSnSe_4$ crystals and of a separate phase of ZnSe. The content of ZnSe diminishes with the rise of the selenisation

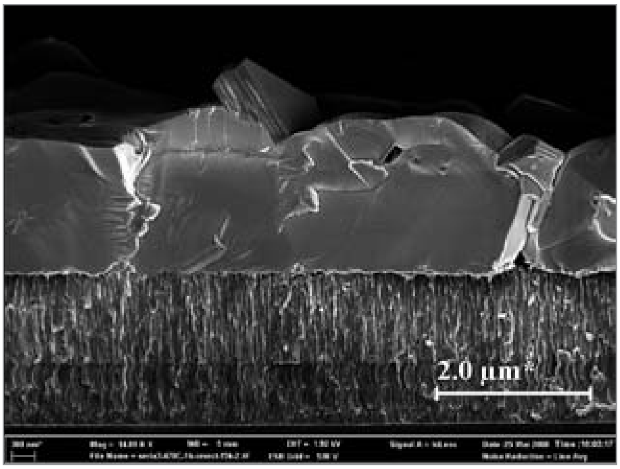


Figure 3.
SEM image of the cross section of the Cu-rich Sn-Zn-Cu film selenised at 470 °C for 1h (Volobujeva 2010).

temperature, but the selenised films stayed always multi-phased. Cu-rich precursors result in more dense films with larger crystals but contain a separate CuSe phase (Fig. 3).

The thermal treatment of ZnSe-SnSe-CuSe precursor films in the inert atmosphere already leads to the formation of homogeneous, but in Se deficit ([Se]=43-45 at %) $Cu_2ZnSnSe_4$ films. The regularities of selenisation of electrodeposited films were similar (Ganchev et al. 2010). The films appeared to be multiphase and contain phases of $Cu_2ZnSnSe_4$, $Cu_{1.8}Se$, $SnSe_2$ and SnSe and ZnSe (Fig. 4).

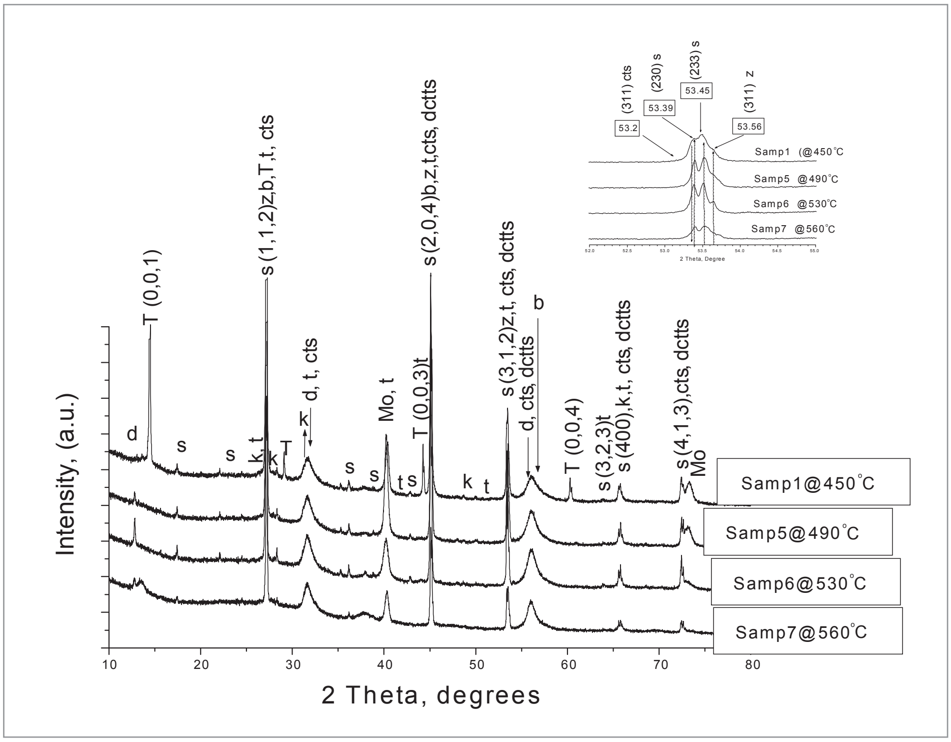


Figure 4.
XRD pattern of CZTS films formed in annealing of Cu-poor Cu-Zn/Sn metal layers in Se vapours at: 1-450 °C; 5-490 °C; 6-530 °C, 7-560 °C. Inset presents the pattern in 2Θ interval 52-55° to gain distinction of CZTS and ZnSe (Ganchev et al. 2010).

CdSe nanostructural films

CdSe nanostructural films were electrodeposited from aqueous solution on ITO coated glass substrates from aqueous electrolytic bath (Kois et al. 2011). The path of possible reactions of formation of stoichiometric CdSe thin films and nanowire structures was proposed. The formation of CdSe polycrystalline films is determined by the concentration of selenia species and cadmium ions in the solution. CdSe thin polycrystalline films with rod-like wurtzite crystals was formed from solutions with high Cd^{2+} concentration in solution. CdSe forms as nanowire and branched structures from solution with low Cd^{2+} content (Fig. 5).

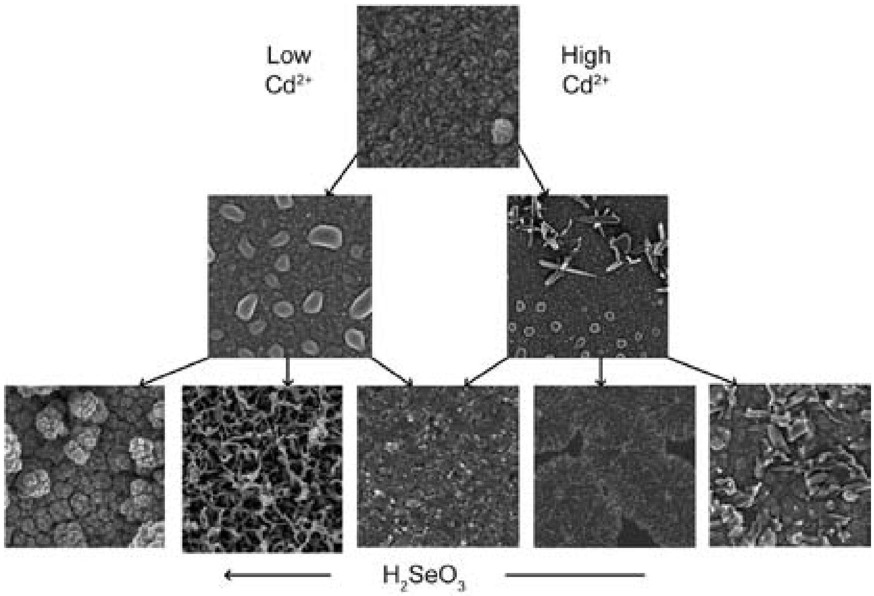


Figure 5.
The growth of thin film structures depending on Cd^{2+} and H_2SeO_3 concentration in electrolytic bath. SEM images inset shows the products at different stages and concentrations (Kois et al. 2011).

Development of hybrid structures

Among various types of SC-s, the organic-inorganic hybrid structures are among the most promising new types (Brabec et al. 2001). The organic materials should not create additional centres of trapping and recombination at the interface with an inorganic photoabsorber. Furthermore, the large dimensions of organic molecules should strongly limit the rate of their diffusion in inorganic layers and the formation of new impurity centres. These structures include well-known inorganic photoabsorbers (CdTe, $CuInS_2$, etc.) and electrically conducting polymers (PEDOT, polypyrol, …). Nanopowders of $CuInS_2$ (CISe) are promising candidates for photosensitive structures. A precursor solution containing copper and indium salts, thiourea and the conjugated polymer was prepared in pyridine, which was coated onto glass/ITO substrates followed by a heating step at 180 °C. By addition of small amounts of zinc salt to the precursor solution, zinc containing $CuInS_2$ (ZCIS) was formed. ZCIS-P3EBT active layers exhibited higher V_{OC} than $CuInS_2$-P3EBT layers and showed efficiencies of about 0.4% (Maiera et al. 2011). In another of our studies CISe powders were prepared by microwave assisted synthesis in triethylene glycol and polyethylene glycol. It was found that the average size of the prepared CISe powder grains depends on the average molecular weight of the solvent. Polyethylene glycol 600 used as the reaction solvent yields uniform polycrystalline CISe powder with an average size of crystals around 80 nm.

Highly-oriented uniform films of p-type $CuIn_3Se_5$ polycrystalline photoabsorber were prepared by high vacuum evaporation (HVE) technique from the stoichiometric precursor on glass/ITO substrates with the following annealing in argon at 500 °C. The values of built-in potentials and free charge carriers concentration profiles of the prepared structures (10^{15}-10^{16} cm^{-3}) that are appropriate for the SCs application were calculated by using the impedance spectroscopy (Laes et al. 2010).

Glass/ITO/CdS/CdTe photovoltaic structures were prepared and the high working-function back contact to CdTe, the conductive polymer layer (PEDOT doped with polystyrenesulfonate) was deposited onto CdTe layers. It was found that deposited conductive polymer gives a nearly ohmic back contact to the p-CdTe layers and improves around three times the photocurrent transfer between CdTe and graphite layers in complete SC structure (Jarkov et al. 2011).

Molecularly imprinted electrically conductive polymers

The overall objective of these investigations is to explore molecular imprinting of synthetic polymers (MIPs) focusing on the development of highly selective and robust antibody mimicking receptors as well as environmental sensors for detection of organic contaminants. Thin films of polypyrrole (PPy), electropolymerised on the surface of electrochemical quartz crystal microbalance (EQCM) electrode in the presence of L-aspartic acid (L-Asp) as template molecules with subsequent overoxidation to create shape complementary cavities, were evaluated as molecularly imprinted polymers (MIP). It was found that synthesis parameters such as electrolyte composition and pH value strongly influence the enantioselectivity of the resulting PPy/L-Asp films. We showed that PPy/L-Asp films prepared in alkaline media have the ca twenty-fold selectivity to the enantiomer used as template. The results suggest the feasibility of preparing molecularly imprinted films by electropolymerisation for the enantioselective recognition of amino acids and the suitability of EQCM for both monitoring the selective recognition as well as to electrochemically modulate the binding process (Ivaništšev et al. 2010). We proposed a new method to generate electrically conducting polymer-based

Surface Imprinted Polymer (SIP) micro- and nanorods for the selective recognition of proteins. The synthesis is based on the use of a microporous polycarbonate sacrificial membrane modified with a target protein. In the pores on this membrane, displaced on the surface of gold electrode, conducting polymer rods of PEDOT were generated electrochemically. The removal of the sacrificial membrane resulted in the formation of surface imprinted microrods localised on the gold surface (Menaker et al. 2009).

In cooperation with Hahn Meitner institute we developed a procedure for electrodeposition on ultrathin (8-120 nm) PPy films from aqueous solutions at pH 5.5 by use of potential pulse sequences on the Si surfaces. It was shown that these PPy films are excellent passivation layers for the Si surfaces. We have proved a new concept of preparing SIPs for protein recognition. The main advantage of the technology is that the SIPs are directly fabricated on SPR chips using standard photolithographic technology, which enables label-free detection of the rebinding effects and provides the premises for their routine large-scale production. Despite the structural complexity of the PEDOT/PSS micropatterns, their interaction with proteins can be determined in a straightforward manner with SPR imaging at sensitivities that outperform those of fluorescence imaging. Imprinting factors of approximately 10 and submicromolar dissociation equilibrium constants for specific bindings were obtained, which places the protein-imprinted PEDOT/PSS among the best SIP-s reported so far (Lautner et al. 2011).

FUTURE PERSPECTIVE

In order to reduce the application of the fossil energy sources and for the practical realisation of advantages of solar energetics in big scale in the future new solar cell materials that are abundant but much cheaper to produce than materials used in industry at the moment are needed. As CZTS is one of the most perspective new adsorber materials, we will continue our research in this field. In our research we will try to turn attention to the chemical side of materials developed and technologies used. As a single phase kesterite materials could be produced in very small chemical potential region, for the practical use of materials we need more precise investigations aimed at determination of experimental boundaries of single-phase formation of these quaternary materials. Several new technologies will be used in material preparation with the aim of realisation of potential of these materials as adsorber for solar cells. The main attention in this direction will be given to the powder and electrochemical technologies, as these are simple technologies and due to this their potential of making solar energetics cheaper is the biggest. In parallel we will continue our research in field of conductive polymers as back contacts for CdTe solar cells as the question of cheap back contacts is the most 'burning' in technology of CdTe solar cells.

Highly selective and robust receptors as well as environmental sensors in the innovative imprinting formats for production of MIP thin films, micro- and nanostructures integrated with various sensor transducers will be elaborated by using advanced technologies such as photo- and nanosphere lithography. Preliminary computational simulations for prediction of selected template molecules interactions with polymer matrices will be carried out in order to optimise the MIP composition. Special attention will be accorded to implementing of label-free sensing strategies for target molecules detection.

Development of electrical double layer capacitors, hybrid supercapacitors and solid oxide fuel cells

Solid oxide fuel cells (SOFC) (Schlapbach, Zutter 2001; Lust et al. 2005abc, 2007, 2010; Küngas et al. 2009ab) are very attractive and environmentally friendly chemical energy (i.e. fuel) conversion systems with noticeably higher energetic efficiency than that for internal combustion engines. Depending on the working temperature, construction and fuel used, the efficiency (electricity and heat together) varies from 55% to 85%. The high efficiency of fuel cells (FC) arises from the electrochemical oxidation of fuel, opposed to thermal burning. The electrical reversibility and efficiency of electric double layer capacitors (EDLC) and hybrid supercapacitors (HSC) is noticeably higher (from 96% to 98%), depending on the construction, electrolyte, electrode materials, i.e. system components used (Scherson, Palenscar 2006; Thomberg et al. 2010; Jänes et al. 2007, 2009, 2010; Laheäär et al. 2009, 2011; Tõnurist et al. 2009; Eikerling et al. 2005; Eskusson et al. 2011; Kurig et al. 2010). Somewhat lower efficiency has been obtained for Li-ion batteries (LIB) (Scherson, Palenscar 2006). Besides, the power densities of EDLC and HSC are very high, however the energy densities of EDLC and HSC are noticeably lower compared with LIB. The technologies for production of SOFC, EDLC and HSC are a bit complicated and therefore the prices of devices are comparatively high. Thus, less expensive systems (electrode materials, electrolytes, separators, construction elements etc.) have to be developed for wider exploitation of EDLC, HSC and SOFC.

Micromesoporous carbons (MMPC) and carbon nanotubes are the main building blocks for EDLC, LIB and HSC as well as for room-temperature (so-called polymer electrolyte) fuel cells (PEM) (Scherson, Palenscar 2006; Hallik et al. 2007, Laborde et al. 1994; Antolini 2009). Micropores are pores with diameters smaller than two nanometres, named according to IUPAC classification. Taking into account the very large specific surface area of MMPCs (1000-2700 m^2g^{-1}) (Thomberg et al. 2010; Jänes et al. 2007, 2009, 2010; Laheäär et al. 2009, 2011; Tõnurist et al. 2009; Eikerling et al. 2005; Eskusson et al. 2011; Kurig et al. 2010), EDLC and HSC with very high energy and power density can be developed.

Decreasing the amount of precious metal in the electrode catalyst layer is a major challenge in the development of low-temperature PEM (Schlapbach, Zutter 2001; Laborde et al. 1994; Antolini 2009). A reduction in the amount of catalyst can be achieved, in most cases, by increasing the active area of metal catalyst (as nanoclusters) that is actually utilised on the MMPC electrode surface. In addition, electrically conductive polymers (ECP) like polypyrrole (pPy) and polyaniline (pAni) as catalyst supports are another hot topic due to the sufficient permeability to fuel and protons as well as good electronic conductivity (Hallik et al. 2007; Laborde et al. 1994). One rather efficient synthesis route is the electrochemical incorporation of Pt(II) or other Pt-metal cations into ECP during electrosynthesis process and their subsequent electroreduction, immediately before completing PEM single cell or PEM FC. However, the price of Pt-metal catalyst layer is still high and less expensive catalysts like rare-earth metal oxides activated with various d-metal atoms (Mn, Co, Ni, Fe, etc.) and Ni-cermet or Cu-cermet anodes catalysing effectively oxygen reduction and fuel oxidation at $T > 500$ °C are very promising materials for SOFC.

Synthesis, study and application of micromesoporous carbons

Focused ion beam-scanning electron microscopy (FIB-SEM), EDX, X-ray diffraction (XRD), X-ray fluorescence (XRF), high-resolution transmission electron microscopy (HRTEM) and electron energy loss spectroscopy (EELS) methods have been used for the structural and chemical analysis of MMPC. Influence of the chemical composition of raw material, chlorination and deactivation temperature T_{synt} (Thomberg et al. 2010; Jänes 2007, 2009, 2010; Laheäär et al. 2009, 2011; Tõnurist et al. 2009; Eikerling et al. 2005; Eskusson et al. 2011; Kurig et al. 2010) on the MMPC parameters have been demonstrated. FIB-SEM (Fig. 6) demonstrates that the 3D hierarchical structure has been elaborated.

According to the studies (Thomberg et al. 2010; Jänes 2007, 2009, 2010; Laheäär et al. 2009, 2011; Tõnurist et al. 2009; Eikerling et al. 2005; Eskusson et al. 2011; Kurig et al. 2010), chaotically distributed linear formations of carbon show neither 2D nor 3D ordering of the crystal lattice, and therefore

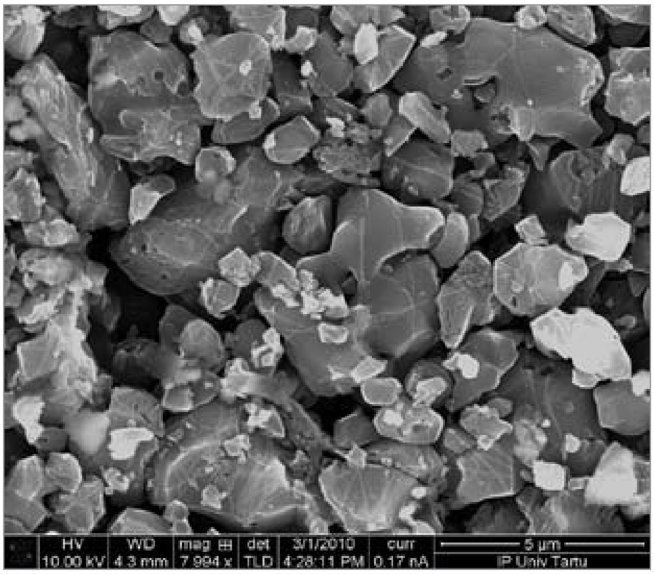


Figure 6.
SEM image of TiC derived carbon electrode.

are recorded as an amorphous phase in XRD patterns. The amount of carbyne (sp-hybrid state with so-called carbyne, acetylene or cumulene type bonds) formed is directly proportional to the percentage of amorphous carbon in MMPC, evaluated by the slope value of the XRD line at front diffraction angle, as well as the number of hexagonal carbon mono/microlayers and mean distance *d* between the hexagonal carbon monolayers in the microlayers.

According to experimental results (Thomberg et al. 2010; Jänes 2007, 2009, 2010; Laheäär et al. 2009, 2011; Tõnurist et al. 2009; Eikerling et al. 2005; Eskusson et al. 2011; Kurig et al. 2010), at T_{synt} > 1200 °C most of carbon atoms are in the sp^2-hybridisation state, forming hexagonal monolayers and thus, a well-developed graphitic microlayer structure that is inevitable for quick Li-ion intercalation into the negatively charged electrodes. Small angle X-ray scattering (SAXS) and small angle neutron scattering (SANS) methods will be applied for the comparative analysis of pore shape (curvature and structure of pores) in MMPC, and the data will be compared with the porous media 3D modelling data calculated using gas adsorption (H_2, Ar, etc.) density functional theory (DFT) and electrochemical impedance data

Raman-scattering experiments (Fig. 7) performed at ambient conditions using a back-scattering geometry for the fixed laser excitation lines show that there is a noticeable influence of T_{synt} and chemical composition of the raw material on the full width at half maximum (FWHM) of the so-called graphite G peak (at 1582 cm^{-1}) and disorder induced D (at ~1350 cm^{-1}) and G` peaks (at 1600-1620 cm^{-1}).

It was demonstrated that the ratio of the integrated intensities of D and G bands (I_D/I_G) depends on T_{synt} and the crystallite size of graphite particles, L_a, calculated using the Tuinstra and Koenig's equation depends noticeably on T_{synt} and raw material used (Thomberg et al. 2010; Jänes 2007, 2009, 2010; Laheäär et al. 2009, 2011; Tõnurist et al. 2009; Eikerling et al. 2005; Eskusson et al. 2011; Kurig et al. 2010).

Various test gases (N_2, Ar) have been used to determine the specific surface area (SSA), pore size distribution (PSD), ratio of nano- and mesopore surface area (S_n/S_m) and volume (V_n/V_m) of MMPC, mixed conducting metal oxide cathodes and metal-cermet anodes, prepared from various raw materials at different T_{synt} (Lust et al. 2005abc, 2007, 2010; Küngas 2009ab; Scherson, Palenscar 2006; Thomberg et al. 2010; Jänes 2007, 2009, 2010; Laheäär et al. 2009, 2011; Tõnurist et al. 2009; Eikerling et al. 2005; Eskusson et al. 2011;

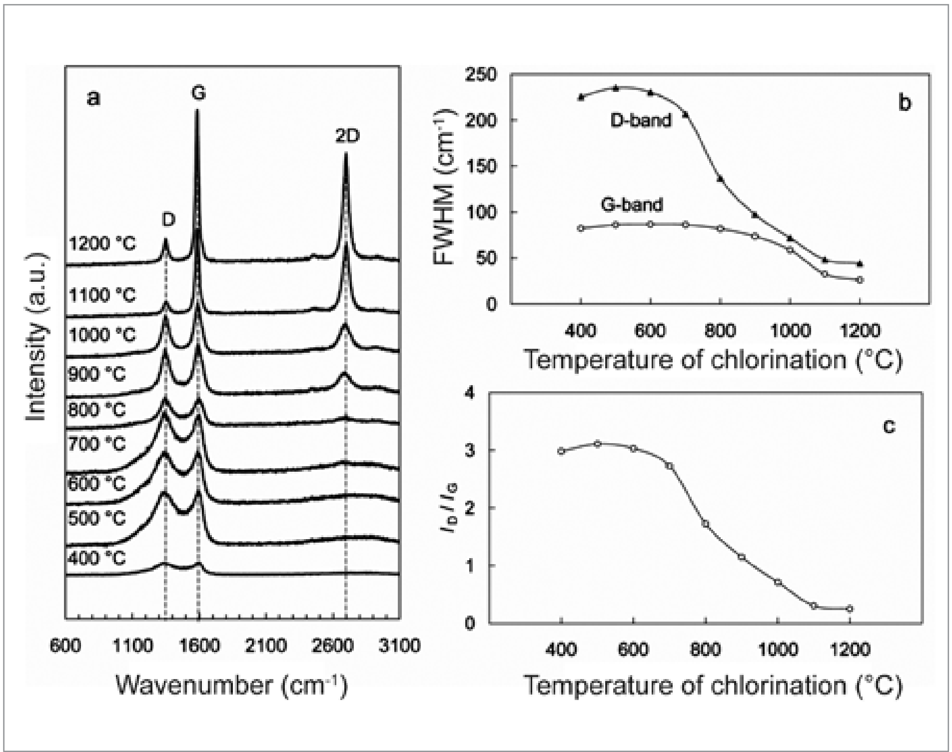


Figure 7.
Raman analysis data for CDC prepared from Mo_2C: (a) spectra for powders prepared at various chlorination temperatures (noted in the figure), (b) full width at half maximum (FWHM) of the D-band and G-band peaks vs. chlorination temperature, (c) ratio of integrated intensities of the D- and G-bands (I_G/I_D) vs. chlorination temperature.

Kurig et al. 2010). The SSA values have been calculated from adsorption isotherm (Fig. 8) using BET equation as well as other models. The influence of T_{synt} on PSD is remarkable.

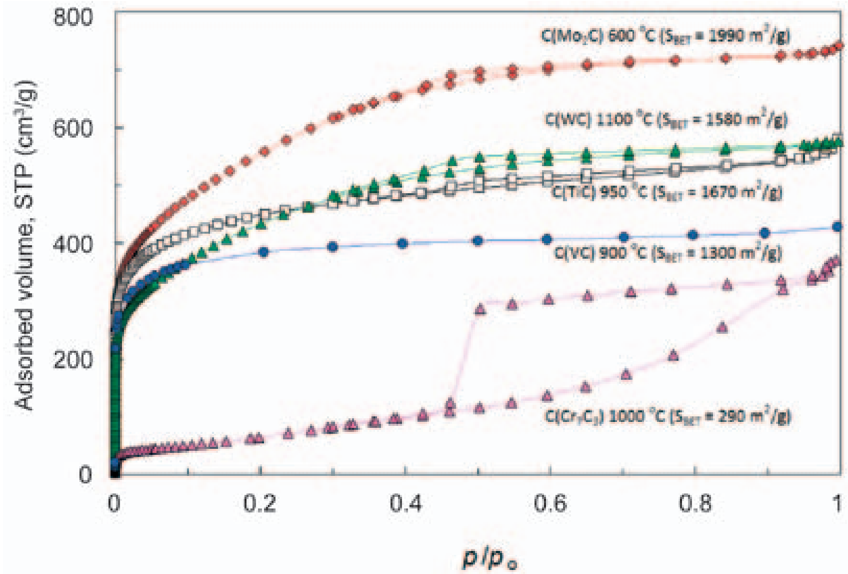


Figure 8.
The N_2 sorption isotherms measured at 77 K and specific surface area calculated according to Brunauer-Emmett-Teller theory (bracketed in figure) for carbon materials derived from different carbides at temperatures noted in figure.

For more detailed analysis various theoretical approximations (Horvath-Kawazoe, Saito-Foley, Barrett-Joyner-Halenda, DFT, etc.) have been used and will be analysed simultaneously with the SAXS and SANS data.
Cyclic voltammetry (Fig. 9) and electrochemical impedance spectroscopy (EIS), simultaneously with BET, FIB-SEM, EDX, EDS, XRD, XPS and SAXS methods have been used to characterise the influence of the chemical composition, crystallographic and porous structures on the electrochemical parameters. It has been found that the shape of Nyquist and Bode plots depends noticeably on the ratio of micro- and mesopores in MMP materials (Thomberg et al. 2010; Jänes 2007, 2009, 2010; Laheäär et al. 2009, 2011; Tõnurist et al. 2009; Eikerling et al. 2005; Eskusson et al. 2011; Kurig et al. 2010). Based on the Eikerling-Kornyshev-Lust model (Eskusson et al. 2011), the influence of SSA, PSD, S_n/S_m and V_n/V_m on the parameters of CDC as well as nanofibre

polymers, electrically conductive polymers and other carbon containing materials, has been demonstrated.

CDC with increased amount of graphite microlayers, synthesised at $T > 1100$ °C, have been used for the Li-ion battery tests in various binary, ternary and quaternary solvent systems with addition of various electrolytes ($LiClO_4$, $LiBF_4$, $LiPF_6$ and lithium bis(oxalato)borate (LiBOB)). Influence of the graphitisation degree of SDC on the kinetic and thermodynamic parameters of Li^+ ion intercalation, power and energy densities has been and will be tested.

Adsorption kinetics of various cations (Li^+, Na^+, tetraalkylammonium cation, tetrakis(diethylamino)phosphonium, tetrakis(dimethylamino)phosphonium, N,N-dimethyl-1,4-diazabicyclo[2,2,2]octanediium) and anions (ClO_4^-, BF_4^-, PF_6^-, BOB^-) (Jänes et al. 2010; Laheäär et al. 2009, 2011) on|inside MMPC from the electrolyte solutions have been compared with the corresponding Raman, XRD, surface normalised interfacial Fourier transform infrared radiation (SNIFTIR) and X-ray photoelectron spectroscopy (XPS) data. The rate determining steps, rate constants, pseudocapacitance, and electrical specific capacity (mAh g^{-1}) values have been calculated. The inter-

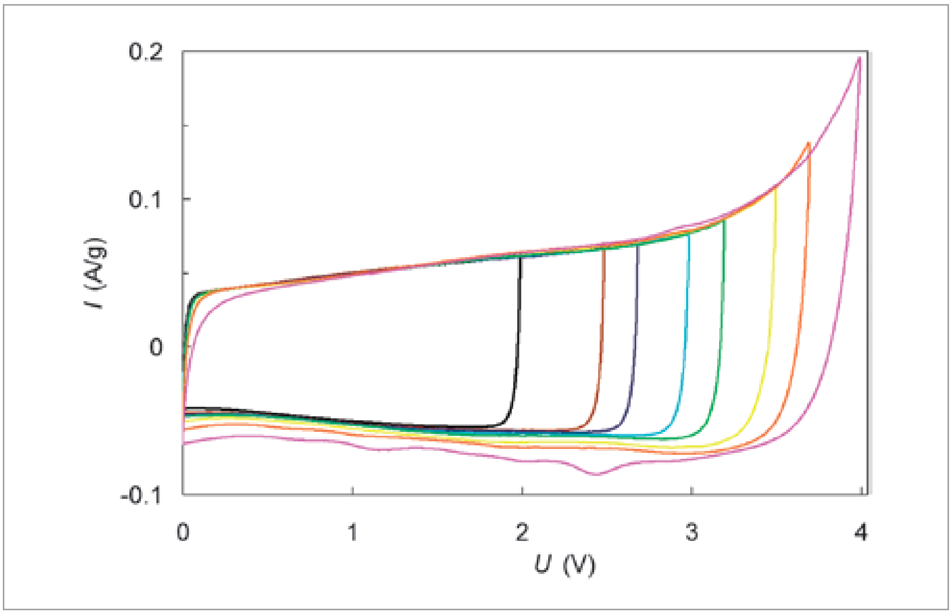


Figure 9.
The cyclic voltammograms for supercapacitor based on carbon cloth (AUVM) electrode and 1-ethyl-3-methylimidazolium tetrafluoroborate measured at voltage scan rate 1 mV/s

mediate products and surface components have been analysed using XPS, FIB- SEM, EDX, XRF (x-ray florescence) methods. According to impedance and constant current charge-discharge data, the very high specific electrical capacity can be achieved for moderately graphitised MMPC modified with some surface-active groups.

Influence of the electrolyte composition on the energy and power densities can be seen in Fig. 10 and 11. Extremely high power density can be achieved in the case of acetonitrile-based electrolytes. The specific surface area, pore size and volume characteristics of a carbon electrode have a very well expressed influence on the power and energy densities.

Room temperature ionic liquids (RTIL) are attractive electrolytes for EDLCs and HSC (Kurig et al. 2010; Siinor et al. 2010) due to the very wide region of ideal polarisation (Fig. 9), low vapour pressure, high dielectric constant and high concentration of charge carriers in fixed volume. Our measurements at Bi(111)|RTIL or nonaqueous electrolyte interface show that differential capacitance depends strongly on chemical composition of an anion. It was verified by detailed investigations of EMIm(CN_4) adsorption at the microporous carbon interface. Therefore in the following years the electrochemical properties of the MMPC|RTIL interface will be studied systematically varying RTIL anions as well as cations molar volume, chemical composition, viscosity, dielectric constant and, thus, polarisability. Influence of temperature and electrode potential on the electrochemical parameters (energy density, power density, characteristic time constant, etc.) will be established. New EDLCs single cells will be completed and tested under current pulsation conditions.

It should be noted that the adsorption/desorption kinetics of tetraalkylammonium, phosphonium and other cations at negatively charged MMPC electrodes determines the characteristic time constant for EDLC and HSC, as the adsorption/desorption of PF_6^- BF_4^- or ClO_4^- anions is comparatively quick (Thomberg et al. 2010; Jänes 2007, 2009, 2010; Laheäär et al. 2009; Tõnurist et al. 2009; Eikerling et al. 2005; Eskusson et al. 2011; Kurig et al. 2010).

Comparative quantum chemical calculations have been made based on the theoretical approximations at the self consistent field (SCF) and density functional theory (DFT) levels completed with Newns-Anderson approximations and molecular dynamics conception (Ivaništšev et al. 2010). These data are very important for future development of modern adiabatic, diabatic and non-adiabatic charge transfer theories as well as for design of HSC.

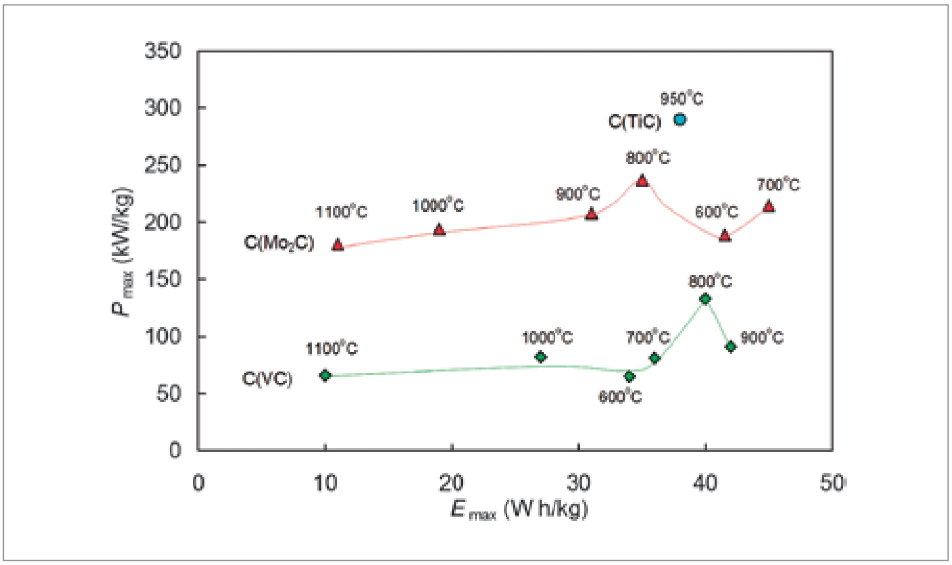


Figure 10.
The maximal specific energy and power densities for supercapacitors based on different carbide derived carbon electrodes (chlorination temperatures are noted in figure) and operational voltages 3.0 V in 1M Et_3MeNBF_4 acetonitrile solution.

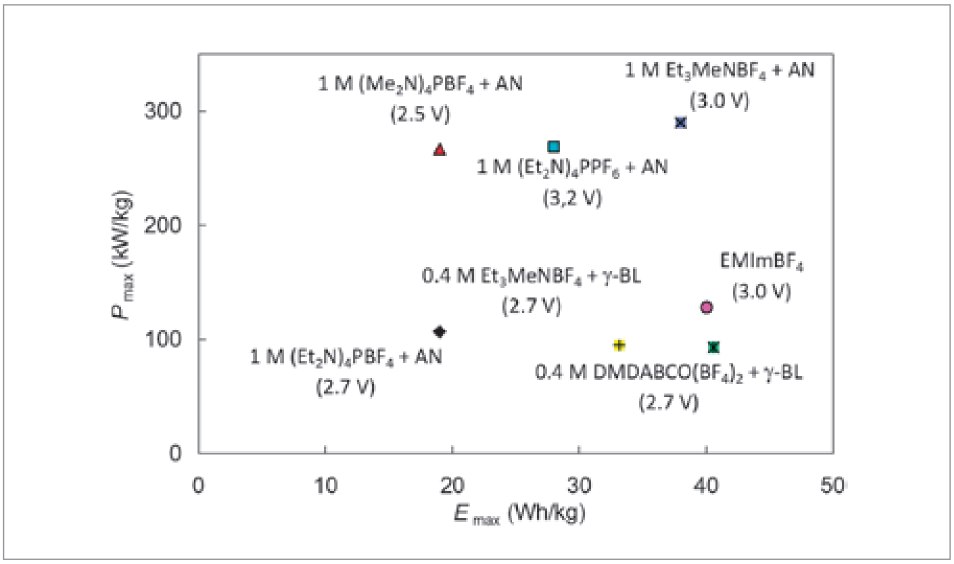


Figure 11.
The maximal specific energy and power densities for supercapacitors based on TiC derived carbon electrodes and different electrolytes at their maximal operational voltages (noted in figure).

Medium temperature solid oxide fuel cells

The chemical energy conversion efficiency increases with the rise in working temperature of FC, and for medium temperature solid oxide fuel cells (MT SOFC) (600 °C < T < 800 °C) it is higher than 60-70% (Schlapbach, Zutter 2011; Lust et al. 2005abc, 2010; Küngas et al. 2009ab). The MT SOFC are very attractive energy conversion systems as various hydrogen containing compounds (natural gas, diesel and biodiesel, alkanes, alcohols, esters, H_2S, NH_3, etc.) can be used as fuels. For MT SOFC Sm_2O_3-stabilised CeO_2 (CSO) or Gd_2O_3-stabilised CeO_2 (CGO), Sc_2O_3 stabilised ZrO_2 (ZSO) can be used as electrolytes and anode is made mainly from Ni/CSO cermet. However, the rate of electroreduction of oxygen from air is a very slow process at $La_{1-x}Sr_xMnO_{3-\delta}$ and one possibility to increase the catalytic activity is to use the electrochemically more active $La_{1-x}Sr_xCoO_{3-\delta}$ (LSCO) or $Pr_{1-x}Sr_xCoO_{3-\delta}$ (PSCO) cathodes (where x is the molar ratio of Sr^{2+} ions (Lust et al. 2005abc, 2007, 2010; Küngas et al. 2009ab). Our studies indicate that a more effective solution is to increase the reaction volume through the development of MMP structure (Fig. 12). The XRD, FIB-SEM, EDX, EDS, XPS, HRTEM, BET, AFM and STM methods have been used for analysis of the materials prepared. Based on the FIB-SEM 3D analysis (Fig. 13), the total porosity 62±2% and specific surface area 0.6±0.05 m^2 g^{-1} have been calculated, in a good correlation with BET data.

The EIS (Fig. 14), CV, CA and CP methods have been used to study the electrochemical properties of cathodes and anodes. The low total polarisation resistance R_p, apparent activation energy (E_{act}) and high power density values

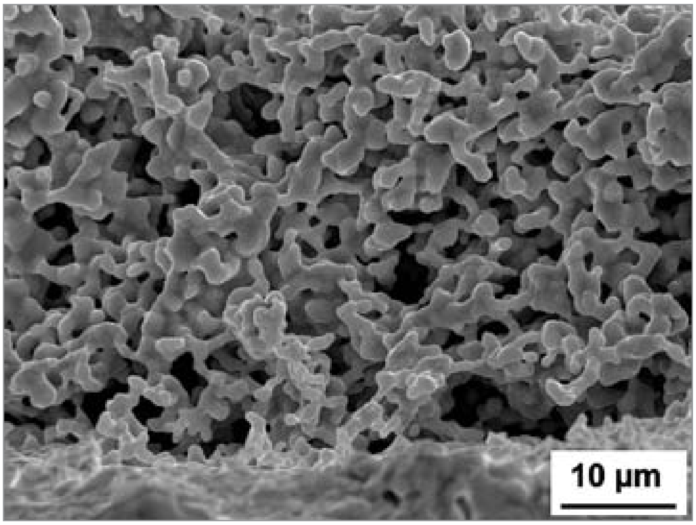


Figure 12. SEM images of the $Ce_{0.9}Gd_{0.1}O_{2-\delta}$| $La_{0.6}Sr_{0.4}CoO_{3-\delta}$ interface.

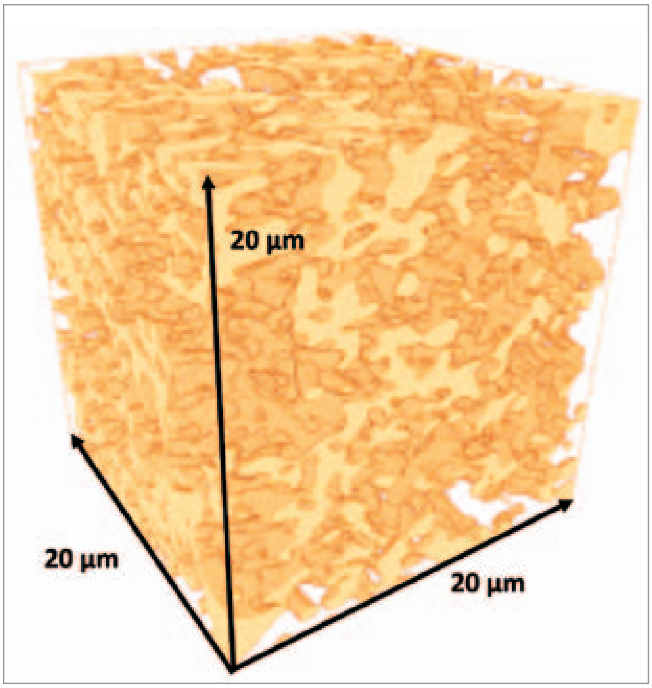


Figure 13.
3D model of $La_{0.4}Sr_{0.6}CoO_{3-\delta}$ cathode reconstructed from FIB-SEM slices.

(Fig. 15) have been obtained, depending noticeably on SSA, PSD, gas phase composition and electrode polarisation (Schlapbach, Zutter 2001; Lust et al. 2005abc, 2010; Küngas et al. 2009ab).

After long-lasting experiments (at least 5000 h) the interface between cathode and electrolyte will be analysed using FIB-SEM, EDX, XPS, XRF and other methods. The ^{18}O penetration profiles of variously pretreated cathode samples will be determined by FIB-SIMS (secondary ion mass-spectroscopy), and the ^{18}O concentration in samples at different penetration lengths will be obtained. The influence of various probable rate limiting steps (diffusion into and inside micropores, adsorption, charge transfer, dissociation and incorporation of O^{2-} anions into the solid phase, mass transport in solid phase) on the total rate of O_2 electroreduction process has been analysed using cathodes with different thicknesses, SSA, PSD, S_n/S_m and V_n/V_m values.

For good time stability of SOFC single cells, bi-layered $Zr_{1-x}Y_xO_{2-\delta}|Ce_{1-x}Gd_xO_{2-\delta}$ electrolytes have been prepared and tested under electrochemical polarisation and H_2 feeding conditions. It was found that these bilayered electrolytes have a good time-stability and thermal cyclability.

MT SOFC requires more active anodes as with decreasing the working T the catalytic activity exponentially decreases. In addition, the traditional Ni-cermet anodes for SOFC do not have good sulphur and carbon tolerance,

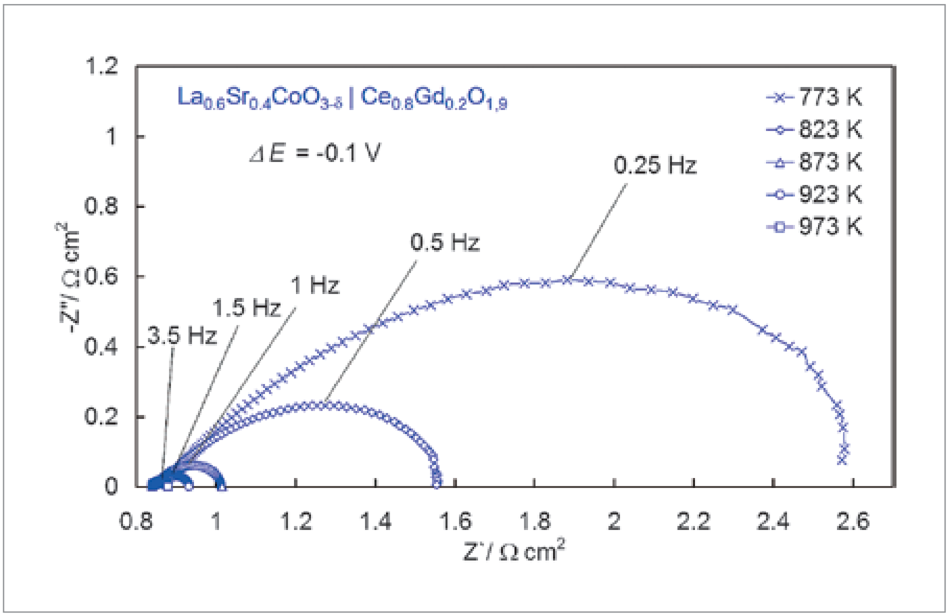


Figure 14.
Nyquist plot for $Ce_{0.9}Gd_{0.1}O_{2-\delta}$| $La_{0.6}Sr_{0.4}CoO_{3-\delta}$ half-cell at *T* and Δ*E*, given in figure.

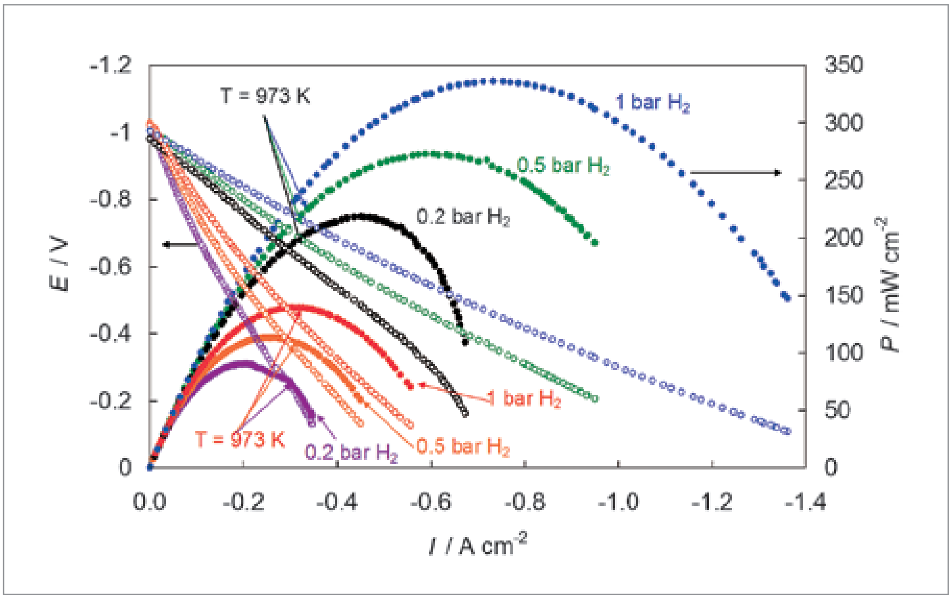


Figure 15.
Cyclic voltammogramms and power density plots at different temperatures and hydrogen partial pressures (given in figure). Single cell: Ni-CGO | ZSO, CGO | LSCO.

because the carbon deposition on|into nanoporous Ni is possible at lower $T \leq 650$ °C. Therefore, new synthesis methods for sulphur-tolerant highly effective MMP Ni-Fe-MgO/YSZ and Cu-CeO_2-Gd_2O_3 will be developed. Cu-CeO_2-cermet is one of the most promising alternative anodes for MT SOFC since Cu is an excellent electronic conductor and CeO_2 is a well-known oxidation catalyst having high ionic conductivity. Another alternative is to use various redox stable mixed oxides like Sc-Y-$ZrTiO_{3-\delta}$, (La, Sr) (Cr, Mn, Ni, Ru)$O_{3-\delta}$ or (Ba, Sr, Ca) (Ti, Nb)$O_{3-\delta}$. However, these systems need very extensive theoretical and experimental studies.

Future directions of studies

The influence of PSD, SSA, ratios of V_n/V_m and S_n/S_m on the catalytic activity for compiling of various fuels (H_2, CH_4, CH_3OH, C_2H_5OH, biodiesel, etc.) will be established, and various theoretical models will be tested and developed in future. SAXS, EXAFS and XANES methods for electrochemical *in situ* studies will be worked out and applied under real electrochemical conditions (fuel and H_2O feeding, electrode polarisation).

As a next step, the time resolved SAXS and SANS with picosecond resolution, element specific X-ray spectroscopy (XAS, X-ray Raman) with 3D depth resolution will be developed for the analysis of nanoparticles properties. Influence of synthesis conditions on the materials physical and chemical properties will be established. Extended X-ray absorption fine structure (EXAFS) method will be developed in future to determine the coordination numbers, bond distances and mean-square disorder for non-Pt metal catalyst surfaces under real electrochemical conditions. Surface structure of intermediates (i.e. adsorbed/absorbed species) will be established in order to analyse the role of the reaction centre properties and influence of electrode potential on the electrochemical kinetics (i.e. on the reaction rate and mechanism). The X-ray absorption near edge structure (XANES) method will be applied for the analysis of methanol, ethanol or methane electrooxidation kinetics at C|PtRu; C|Ni; C|CoN_4; C|FeN_4, etc. catalysts to obtain the d-metal atom oxidation state, local symmetry and electronic structure under the real electrochemical conditions. Electrocatalysis processes at Li-ion battery d-metal complex oxide cathodes and complex anodes, hybrid supercapacitor pseudocapacitive electrodes and solid oxide fuel cell anodes will be studied using XANES, and the data obtained will be correlated with electrochemical data to analyse the electrode surface blocking effect on the mass and charge transfer kinetics. Thus, as a result the more efficient energy conversion and storage systems will be developed.

Acknowledgement

This work was supported by the target financing by HTM (Estonia) projects No-s. SF0140099s08, SF0180002s08, and SF0142714s06, by the Estonian Science Foundation grants 6455, 6467, 7606, 7791, 8147, 8655, 8172, and 8357 and by AS Elcogen.

REFERENCES

Altosaar, M., et al. 2008. $Cu_2Zn_{1-x}Cd_xSn(Se_{1-y}S_y)_4$ solid solutions as absorber materials for solar cells. Physica Status Solidi A , 205, 167-170.

Altosaar, M., Mellikov, E. 2000. $CuInSe_2$ Monograin Growth in CuSe-Se Liquid Phase. Japanese Journal of Applied Physics, 39, 65-66.

Antolini, E. 2009. Carbon supports for low-temperature fuel cell catalysts. Applied Catalysis B: Environmental, 88, 1–24.

Brabec, C. J., Sariciftci, N. S., Hummelen, J. C. 2001. Plastic solar cells. Advanced Functional Materials, 11, 1, 15-26.

Eikerling, M., Kornyshev, A. A., Lust, E. 2005. Hierarchical model of high energy density double layer capacitor. Toward optimisation of the electrical double layer capacitor characteristics. Journal of the Electrochemical Society, 152, E24-E33.

Eskusson, J., et al. 2011. Physical and electrochemical characteristics of supercapacitors based on carbide derived carbon electrodes in aqueous electrolytes. Journal of Power Sources, 196, 4109-4116.

Feltrin, A., Freindlich, A. 2008. Material considerations for terawatt level deployment of photovoltaics. Renewable Energy, 33, 180-185.

Ganchev, M., et al. 2010. Formation of $Cu_2ZnSnSe_4$ thin films by selenization of electrodeposited stacked binary alloy layers. Energy Procedia, 2, 1, 65-70.

Grossberg, M., et al. 2009. Compositional dependence of Raman scattering and photoluminescence emission in Cu-Ga-Se films grown by MOCVD. Physica B, 404, 1984-988.

Hallik, A. et al. 2007. On the porosity of polypyrrole films. Synthetic Metals, 157, 1085-1090.

Ito, K., Nakazawa, J. 1988. Electrical and optical properties of stannite-type quaternary semiconductor thin film. Japanese Journal of Applied Physics, 27, 2094-2097.

Ivaništšev, V., Nazmutdinov, R. R., Lust, E. 2010. Density functional theory study of the water adsorption at $Bi_{(111)}$ electrode surface. Surface Science, 604, 1919-1927.

Jarkov A., et al. 2011. Conductive polymer PEDOT: PSS back contact for CdTe solar cell. Solar Energy Materials and Solar Cells, (in press).

Jänes, A., et al. 2009. Nanoscale fine-tuning of porosity of carbide-derived carbon prepared from molybdenum carbide. Carbon, 47, 23-29.

Jänes, A., et al. 2010. Novel doubly charged cation based electrolytes for non-aqueous supercapacitors. Electrochemistry Communications, 12, 535539.

Jänes, A., Thomberg, T., Lust, E. 2007. Synthesis and characterisation of nano-porous carbide-derived carbon by chlorination of vanadium carbide. Carbon, 45, 27172722.

Kois, J., et al. 2011. Electrocrystallization of CdSe from aqueous electrolytes: Structural arrangement from thin films to self-assembled nanowires. Journal of Crystal Growth, 320, 1, 9-12.

Krustok, J., Collan, H., Hjelt, K. 1997. Does the low temperature Arrhenius plot of the photoluminescence intensity in CdTe point towards an erroneous activation energy? Journal of Applied Physics, 81, 3, 1442-1445.

Krustok, J., et al. 2010. Potential fluctuations in $Cu_2ZnSnSe_4$ solar cells studied by temperature dependence of quantum efficiency curves. Physica B, 405, 3186-3189.

Kurig, H., Jänes, A., Lust, E. 2010. Electrochemical characteristics of carbide derived carbon I 1-ethyl-3-methylimidazolium tetrafluoroborate supercapacitor cells. Journal of the Electrochemical Society, 157, A272-A279.

Küngas, R., et al. 2009a. Statistical method to optimize the medium temperature solid oxide fuel cell electrode materials. Journal of Electroanalytical Chemistry, 629, 94-101.

Küngas, R., Kivi, I., Lust, E. 2009b. Effect of cell geometry on the electrochemical parameters of solid oxide fuel cell cathodes. Journal of the Electrochemical Society, 156, B345-B352.

Laborde, H., Leger, J. M., Lamy, C. 1994. Electrocatalytic oxidation of methanol and C1 molecules on highly dispersed electrodes Part 1: Platinum in polyaniline. Journal of Applied Electrochemistry, 24, 219-226.

Laes, K., et al. 2010. The impedance spectroscopy of $CuIn_3Se_5$ photoabsorber films prepared by high vacuum evaporation technique. Energy Procedia, 2, 119-131.

Laheäär, A., et al. 2009. $LiPF_6$ based ethylene carbonate – dimethyl carbonate electrolyte for high power density electrical double layer capacitor. Electrochimica Acta, 54, 4587-4594.

Laheäär, A., Jänes, A., Lust, E. 2011. Electrochemical behavior of carbide derived carbons in $LiPF_6$ and $LiCF_3SO_3$ nonaqueous electrolytes. Electrochimica Acta, (in press).

Lautner G., et al. 2011. Selective artificial receptors based on micropatterned surface-imprinted polymers for label-free detection of proteins by SPR imaging. Advanced Functional Materials, 21, 591-597.

Luque, A., Hegedus, S. 2006. Handbook of Photovoltaic Science and Engineering, Wiley.

Lust, E., et al. 2005a. Electrochemical characteristics of Ce0.8Gd0.2O1.9| La0.6Sr0.4CoO3-δ + Ce0.8Gd0.2O1.9 half-cell. Journal of Solid State Electrochemistry, 9, 674-683.

Lust, E., et al. 2005b. Electrochemical characteristics of La0.6Sr0.4CoO3-delta, Pr0.6Sr0.4CoO3-delta and Gd0.6Sr0.4CoO3-delta on Ce0.85Sm0.15O1.925 electrolyte. Journal of Solid State Electrochemistry, 9, 882-889.

Lust, E., et al. 2005c. Optimization of the cathode composition for the intermediate-temperature SOFC. Journal of the Electrochemical Society, 152, A2306-A2308.

Lust, E., et al. 2007. Meetod tahke oksiid kütuselemendi ühikraku valmistamiseks ja nimetatud ühikrakk. PCT/EE2007/000010 (31.05.2007).

Lust, E., et al. 2010. Electrochemical and gas phase parameters of cathodes for intermediate temperature solid oxide fuel cells. Electrochimica Acta, 55, 7669-7678.

Maiera, E., et al. 2011. $CuInS_2$–Poly(3-(ethyl-4-butanoate)thiophene) nanocomposite solar cells: Preparation by an in situ formation route, performance and stability issues. Solar Energy Materials and Solar Cells, 95, 5, 1354-1361.

Mellikov, E., Hiie, J., Altosaar, M. 2007. Powder materials and technologies for solar cells. International Journal of Materials & Product Technology, 28, 3-4, 291–311.

Mellikov, E., et al. 2008a. Research in solar cell technologies at Tallinn University of Technology. Thin Solid Films, 516, 7125-7134.

Mellikov, E., et al. 2008b. Monograin materials for solar cells. Solar Energy Materials and Solar Cells, 93, 65-68.

Mellikov, E., et al. 2010. $Cu_2(ZnxSn_2\text{-}x)(SySe1\text{-}y)4$ monograin materials for photovoltaics. Wicks, G. G., et al (eds.) Materials Challenges in Alternative and Renewable Energy. The American Ceramic Society, Wiley-Blackwel, 137-142.

Menaker A., et al. 2009. Electrosynthesized surface-imprinted conducting polymer microrods for selective protein recognition. Advanced Materials, 21, 2271–2275.

Mitzi, D. et al. 2008. A high-efficiency solution-deposited CIGS device. Advanced Materials, 20, 3657-3662.

Scherson, D. A., Palenscar, A. 2006. Batteries and electrochemical capacitors. The Electrochemical Society Interface, 15, 17-19.

Schlapbach, L., Zutter, A. 2001. Hydrogen-storage materials for mobile applications. Nature, 414, 353-358.

Siinor, L., Lust, K., Lust, E., 2010. Influence of anion composition and size on the double layer capacitance for $Bi_{(111)}$|room temperature ionic liquid interface. Electrochemistry Communications, 12, 1058-1061.

Thomberg, T., Jänes, A., Lust, E. 2010. Energy and power performance of electrochemical double-layer capacitors based on molybdenum carbide derived carbon. Electrochimica Acta, 55, 3138-3143.

Tõnurist, K., et al. 2009. Influence of mesoporus separator properties on the parameters of electrical double-layer capacitor single cells. Journal of the Electrochemical Society, 156, A334-A342.

Volobujeva, O. 2010. HRSEM study of $Cu_2ZnSnSe_4$ film formation. Imaging & Microscopy, 12, 2, 27-29.

Volobujeva, O., et al. 2009a. SEM analysis and selenization of Cu–Zn-Sn sequential films produced by evaporation of metals. Proceedings: Conference on Optoelectronic and Microelectronic Materials and Devices. IEEE Publishing, 257-260.

Volobujeva, O., et al. 2009b. $Cu_2ZnSnSe_4$ films by selenization of Sn–Zn-Cu sequential films. Journal of Physics and Chemistry of Solids, 70, 567-570.

Wibowo, R. A., et al. 2007. Single step preparation of quaternary $Cu_2ZnSnSe_4$ thin films by RF magnetron sputtering from binary chalcogenide targets. Journal of Physics and Chemistry of Solids, 68, 1908-1913.

Vision 2004 = A Vision for Photovoltaic Technology for 2030 and Beyond. European Commission.

ADVANCED MULTIFUNCTIONAL MATERIALS AND THEIR APPLICATIONS

M. Antonov, I. Hussainova, J. Kers, P. Kulu, J. Kübarsepp, R. Veinthal

INTRODUCTION

Basic documents of the European Technology Platform for Advanced Engineering Materials and Technologies (EuMaT) (2011) have emphasised that the materials (including but not limited to construction and structural engineering materials) play a critical role in competitiveness of companies, operational and workplace safety, and are a key factor in environmental compliance. The term 'advanced engineering materials and technologies' refers in EuMaT to the three pillars (Roadmap 2011):

- multifunctional materials;
- materials for extreme conditions;
- hybrid and multimaterials.

It was pointed out that in the field of materials science EU should respond to challenges described in Lund Declaration by generation of new advanced materials able to support the development of totally new products and goods, by radical improvement in the characteristics of widely used conventional materials, by substitution of traditional materials with most eco-efficient ones, by replacement and/or recycling of rare and/or scarce materials with less critical or expensive alternatives (Roadmap 2011).

However, novel advanced materials are still challenged by our less than complete knowledge of materials properties tailoring and the fundamental processes of materials degradation, as well as inadequate technologies for production, monitoring and predicting materials performance, environmental and/or health impacts.

Multifunctional materials are excellent examples of the systems that combine several functions. The focus is on novel concepts that effectively combine materials science, mechanics and physics to produce multifunctional systems for the specific needs.

Areas of research interest of Department of Materials Engineering of Tallinn University of Technology (TUT) are closely oriented to the trends of EuMaT and may be outlined as follows:

- ceramics and ceramic-based composite tribomaterials;

- coatings and surface engineering;
- tribology;
- recycling and reuse.

MAIN AREAS OF CURRENT RESEACH

Multiphase tribomaterials

The research programme related to multifunctional materials has been carrying out under the targeted financing project "Design and technology of multiphase tribomaterials" in 2008–2011.

The research has embraced several major problems of composites technology and performance, such as:

- design, development and modelling of multiphase materials;
- macro-, micro- and nanoscale characterisation;
- development of advanced production technology of multiphase materials.

Special focus has been placed upon the formation of multiphase and multicomponent ceramic-based materials, such as cermets, nanocomposites, and hybrid materials.

Research in the field of design, development and modelling of multiphase materials is aimed to design materials with tailored properties; assessment of their performance and materials selection under specific conditions of application. Recent research has resulted in the following:

- Influence of the interface surfaces in ceramic and metal composites on thermo-physical properties was found (Hussainova et al. 2009b). Thermo-residual stresses in cermets were studied both theoretically (including the method of stress analysis by modelling of the superposition of the elastic stresses around defects) and experimentally. A model of stress calculations considering core-rim structured inclusions embedded in a metal matrix was developed. Results allow prediction of geometrical parameters of a composites structure to prevent crack formation and obtain the cermets of high tribological performance. As an example cermets possessing lowest residual stresses reveal the highest erosion resistance (Hussainova et al. 2009a).
- The results for the calculated effective elastic moduli are in a very good agreement with experimentally measured effective elastic moduli. The analytical and experimental results coincide within few percent. The results obtained using the Mori-Tanaka theory fall be-

tween the Hashin-Shrikman bounds and are very close to the upper limit (Hussainova et al. 2011a).

- The new experimental equipment and related procedure were developed to study parameters of particle-target collision process. It was shown that energy absorbed by materials may serve as an appropriate guide for materials selection (Hussainova, Schade 2008).
- WC-based and zirconia doped cermets of high hardness and good fracture toughness (> 19.5 MPa·$m^{0,5}$) were produced by pressure assisted sintering. The addition of yttria stabilised tetragonal zirconia (~13 vol%) allows exploitation of the mechanism of stress-induced phase transformation of ZrO_2. Combined with WC grain refinement and, therefore, an increase in hardness, this ensures at least twofold increase in erosion and abrasion resistance, especially at high temperatures, as compared with conventional WC-Co hardmetals (Hussainova et al. 2011b).

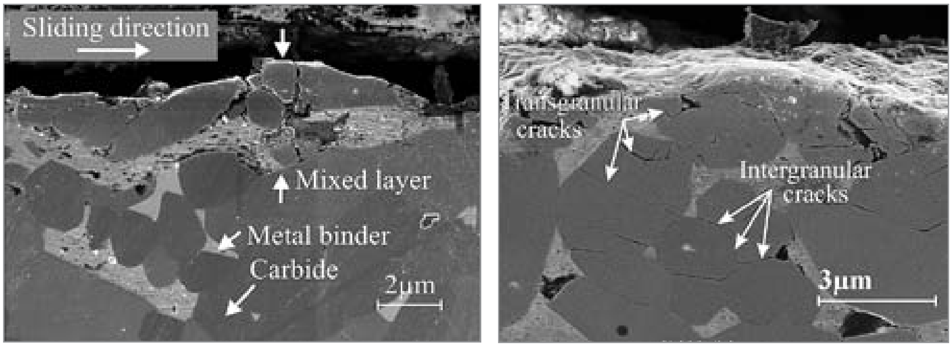
Research in the field of macro-, micro- and nanoscale characterisation of multiphase materials was aimed to establish generic relationships between structure, properties (mechanical, tribological) and mechanism of degradation. The main results in tribological characterisation of composites are the following:

- Tribological performance and wear mechanism of cermets with different composition (WC-, TiC-, Cr_3C_2-, B_4C- and c-BN-based) in different wear conditions (erosion, two-body and three-body abrasion, sliding and adhesion) has been studied. Abrasive and erosive wear performance was evaluated both at room and elevated (up to 900 °C) temperatures. Some of the most important conclusions are given below:
    - abrasive wear resistance of WC-, TiC- and Cr_3C_2-based cermets depends on the type of alloys and carbide/binder ratio. Abrasive wear mechanisms are similar for different cermets depending mainly on the microstructural features and hardness of an alloy and on the ratio H_a/H_m (hardness of abrasive and material) (Hussainova et al. 2009ab; Pirso et al. 2010a);
    - at equal carbide volume fraction WC-based composites are at advantage over TiC- and Cr_3C_2-based cermets in sliding, abrasive and erosive wear conditions (Pirso et al. 2009, 2010b; Klaasen et al. 2010a). In adhesive wear conditions TiC-based and steel bonded cermets demonstrate wear performance approximate to that of conventional WC-based hardmetals (Klaasen et al. 2010a);

- elevated temperature (up to 650 °C) erosion rate for the cermets of low binder content does not significantly increase due to an increase in plasticity and development of thin oxide layer. The cermets of high binder content are less erosion resistant due to softening and melting under impacts (Pirso et al. 2009);
- new design of a high-temperature erosive wear test rig (up to 1000 °C) to study synergy of wear and oxidation (Antonov, Hussainova 2009);
- mapping of the materials performance under erosive, high-temperature erosive, corrosive and oxidative conditions allows cermets selection for specific applications (Antonov, Hussainova 2009, etc.).

➢ An indentation method was used to reveal surface structural change under different wear conditions. Formation of subsurface mechanically mixed layers (MML) (Fig. 1) was found to be an essential feature of materials response to applied loading (Hussainova et al. 2009a; Antonov, Hussainova 2009). The structure and mechanical properties of such layer may influence tribological performance of materials.

➢ The morphology of surface failure in different wear conditions was studied. Capability of nominally 'brittle' phase (WC, TiC) to local plastic strain during loading (wear as well as fatigue) was shown. Plasticity of the carbide phase depends on the loading mode and is higher under prevalence of shear stresses (Klaasen et al. 2010a; Klaasen, Kübarsepp 2006).



a b

Figure 1.
Cross section of Cr_3C_2-15 wt% Ni cermet after abrasion (a) and erosion (b) (Antonov, Hussainova 2009).

- The morphology of surface failure in different wear conditions was studied. Capability of nominally 'brittle' phase (WC, TiC) to local plastic strain during loading (wear as well as fatigue) was shown. Plasticity of the carbide phase depends on the loading mode and is higher under prevalence of shear stresses (Klaasen et al. 2010a; Klaasen, Kübarsepp 2006).

Results in the field of mechanical and fatigue characterisation of composites are the following:

- The characterisation of mechanical properties (E and H) of cermets at micro- and nanoscales was carried out. Measuring of the intrinsic micro- and nanoscale properties of each phase of a composite separately gave the information on spatial heterogeneity in local material properties. The measured local elastic properties are used as inputs for modelling effective elastic response of composites and serve as a guide to process engineering and advanced materials design (Hussainova et al. 2011a).
- The data analysis and recalculation of the projected area are of primary importance for evaluating the mechanical characteristics (E and H) extracted from nanoindentation tests. The semi-ellipse method for accounting of the pileups has been applied to the multiphase material, and the hardness and elastic moduli of its phases showed a good agreement with published data. The differences in the values may point to the formation of new phases during sintering and, therefore, to a possible difference in mechanical behaviour of the bulk composite as compared to the expected one at the design stage (Hussainova et al. 2011a).
- Fracture and fatigue behaviour of WC- and TiC-based composites under cyclic three-point bending loading was studied in (Sergejev et al. 2008). Fatigue of heterogeneous materials is based on the fact that both phases are involved in the fracture process although failure starts preferably with plastic strain in the binder. The WC-hardmetals have higher strength properties (insert strength, fatigue limit) than TiC-cermets but a wider scatter in strength and higher fatigue sensitivity (Fig. 2). Therefore TiC-based cermets may be prospective candidates to replace the hardmetals in metal forming operations (Klaasen et al. 2011).
- Technique for surface fatigue testing was elaborated (Sergejev et al. 2007). Surface fatigue testing results showed excellent convergence with the experimental data received from three-point fatigue tests.

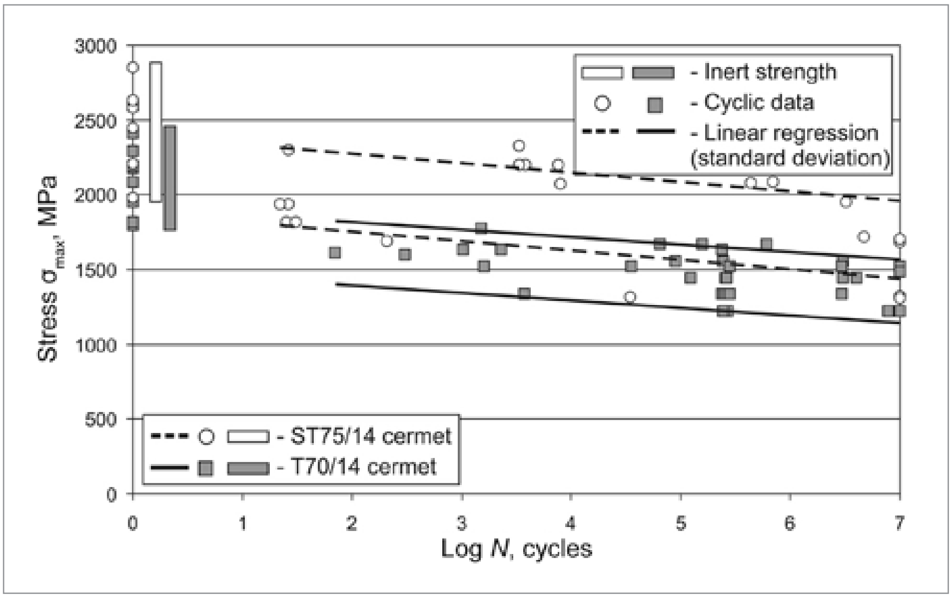


Figure 2.
Wöhler plot of the cycles to failure (N) for different cyclic stress amplitudes (σ_{max}) for TiC-based cermets.

Research in the field of development of advanced production technology is aimed to modify the structure (grain size and morphology, phase composition, structural homogeneity, etc.) of multiphase tribomaterials to improve their performance. Research has resulted in the following:

- The sol-gel process was developed for the production of the new powders for cermets production. Advanced Sinter-HIP technology was utilised in production of such cermets with enhanced tribological properties.
- The reactive sintering technology of Cr_3C_2-, TiC- and WC-based cermets was developed. The advantage of the new technology combining attritor milling with reaction sintering is that composites of ultrafine or nanograined structures are coupled with a cost-effective process of powder production (Pirso et al. 2010b).
- Self-propagating high-temperature synthesis (SHS-process) was utilised for production of lightweight cubic boron nitride (c-BN) and boron carbide (B_4C) composites.
- The technology of ultra-fine grained (nanostructured) Cu, Nb and Ta was developed by applying severe plastic deformation utilising equal-channel angular processing (ECAP) method (Kommel 2010).

- A new technology for manufacturing of nano-silicon carbide (nano-SiC) reinforced metal-matrix composites was developed. The advanced planetary milling and hot compaction (hot extrusion) technologies were used. Nano-SiC reinforcement increases remarkably hardness and tensile strength of metal matrix composites (Kollo et al. 2010).
- Influence of sintering technology (sintering modes and parameters) on the performance of WC-, TiC- and Cr_3C_2-based cermets was investigated. It was found, that gas compression during sintering (Sinter/HIP process) has positive effect on the performance of both WC- and TiC-based composites.
- Reliability of joints 'carbide composite+steel' produced by diffusion welding was investigated. The strength and reliability of the joint is higher when diffusion banding is used instead of brazing (Klaasen et al. 2010b).

Coatings and surface engineering

Research in the field of coatings in 2008–2011 has been carried out in the framework of targeted financing project "Coatings and surface engineering".

According to objectives of the research and planned activities R&D was carried out at four subtopics:

- thermal sprayed and PVD deposited coatings;
- testing and characterisation of tribosystems;
- determination of residual stresses in coatings;
- reliability of machines and products.

Research in the field of wear resistant thermal sprayed coatings is oriented to production of composite coatings based on recycled hardmetal powders and commercial Fe-based self-fluxing alloy powders and oriented to applications in such cost-sensitive areas as mining, energy production, etc.

To produce the powders from used hardmetal/cermets parts, the high energy mechanical milling with the help of a semi-industrial disintegrator system DS-350 (for preliminary milling) and a laboratory disintegrator system DS-175 (for final milling to produce powder with particle size 20–100 μm) was used. Grindability, chemical composition, particle granularity and angularity were studied by sieving analysis, laser granulometry, SEM analysis and analytical methods.

To produce thermal sprayed coatings, the composite powders on the base of iron self fluxing alloy powder (FeCrSiB) with WC-Co reinforcement (20-50 vol.%) and from the deposition methods plasma, high velocity oxyfuel (HVOF) and plasma transferred arc (PTA) spraying methods and corresponding devices were used. A post treatment (laser cladding) was applied to improve the properties of sprayed coating (experiments were performed at the Institut für Fügetechnik und Werkstoffprüfung, Jena and Tampere University of Technology).

The microstructural analysis, mechanical characterization and tribotesting (at abrasion, erosion and impact wear conditions) of the spray fused coatings were performed. For comparison the nickel self-fluxing alloy based coatings were studied and advantages of Fe-based coatings reinforced by WC-Co hardmetal was demonstrated *(*Fig. 3a and b*)* from the point of optimal structure formation.

In the field of the thin hard coatings physical vapour deposited (PVD) coatings (mono-, multilayer and composite coatings of TiN, TiAlN, AlTiN; nanocoating nACo) and different coating systems (hardmetal-coating, high speed steel (HSS)-coating, nitrided steel – coating-duplex coatings) were investigated.

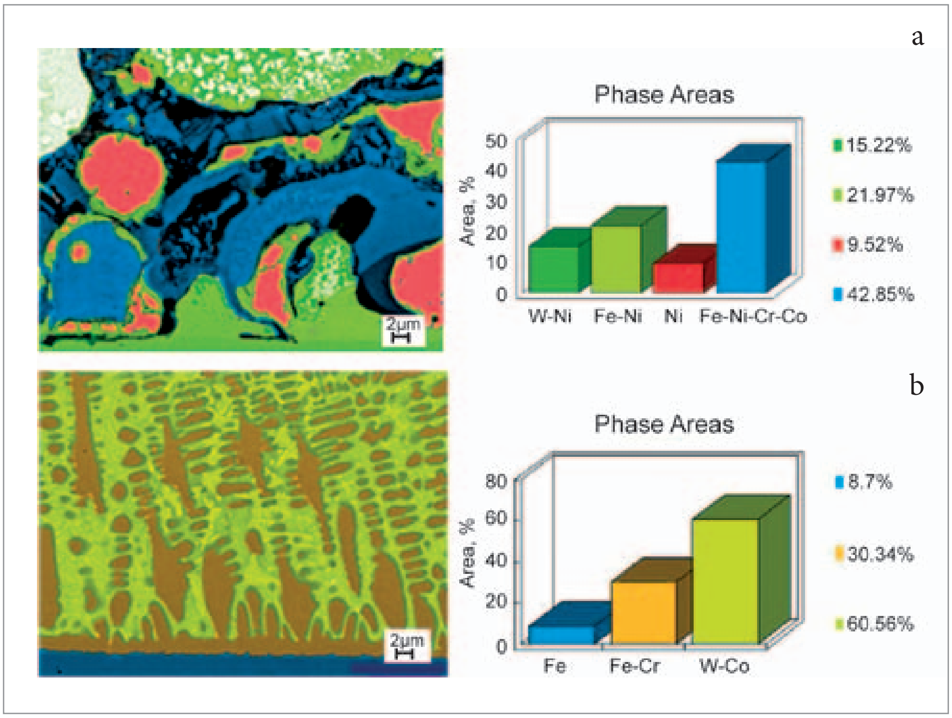


Figure 3.
Microstructure and composition of plasma sprayed-laser remelted self-fluxing alloy based coatings: a – NiCrSiB-20 vol% (WC-Co); b – FeCrSiB-20 vol% WC-Co.

The different commercial coatings (TiN, TiCN, Ti/AlN and nACo) on WC-Co substrate with different surface roughness were studied (Fig. 4). The relationship between surface roughness, coefficient of friction (CoF) and wear resistance was clarified. To decrease the droplet phase on surface and increase the bonding strength of coatings the research for optimisation of deposition parameters was performed. Besides the classical WC-Co hardmetal substrates the PVD coatings on TiC-NiMo cermets were studied. The properties of coatings (adhesion, nanohardness, cracking resistance and surface fatigue) using indentation methods (Rockwell hardness test, cyclic indendation and impact wear) were studied and recommendation for the use of coatings for tooling were proposed.

To extend the application areas of thin hard coatings the duplex coatings and duplex treatments (PVD coatings on plasma nitrided steel, laser hardening of PVD coated surfaces) were under consideration. As the result of studies the principles for coatings selection under different operation conditions were offered.

The main results in the field of research connected to the thin coatings are the following:

- the architecture of different coatings systems for different loading conditions has been optimised;
- the technological parameters of deposition and other technological means for improving coatings quality and performance have been proposed;
- the principles for coatings design and selection for users – toolmakers have been offered.

For characterisation of the thin films a new nanoindentation device was supplied and original tribometer – rototribometer was developed. A new area of research is coatings based on carbon – diamond coatings (DC) as well as diamond-like coatings (DLC) along with carbon nanofibers (CNF) and/or carbon nanotubes (CNT) reinforced coatings.

The research in the field of testing and characterisation of tribosystems was oriented to increase the wear performance of materials and coatings for applications in erosive and impact wear conditions.

The following issues were in the focus of the study:

- wear performance of novel composite bulk materials and coatings;
- mechanisms of wear degradation of the coatings;
- wear processes related to surface fatigue;
- analytical and computational modelling of wear.

The aims of the research were:

- to analyse and provide new analytical models and physical test methods for wear prediction in different wear conditions;

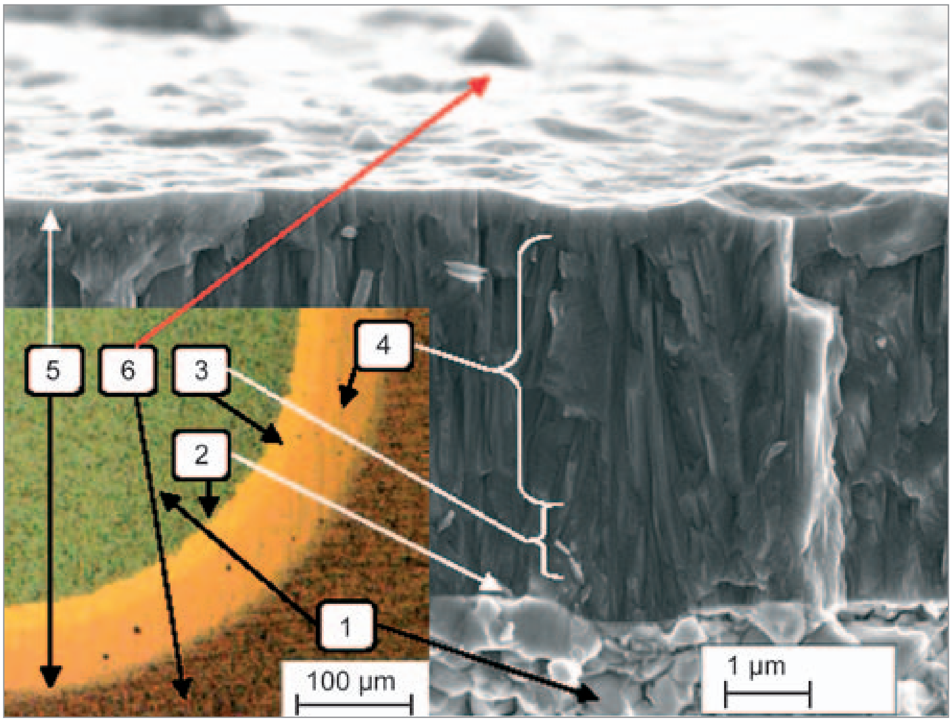


Figure 4.
Optical microscope and SEM images of fractured surface of hard PVD TiCN-coating on WC-Co hardmetal: 1 – WC-Co substrate; 2 – Ti adhesion layer; 3 – TiN coating; 4 – TiN gradient coating; 5 – C-rich TiN top layer and 6 – coating surface with microdroplets.

- to develop novel materials and coatings with unique combination of mechanical properties;
- to develop the model of materials degradation;
- to evaluate the perspective and applicability of double-reinforced materials;
- to estimate the intensity of fatigue wear in different wear modes and to develop counter-measures against this mechanism.

Main results in the field of modelling and prediction of wear are the following:

- New methods for evaluation of wear performance were developed (a contact-fatigue test apparatus; a high-energy impact wear tester) and some test set-ups were modernised: a centrifugal particle accelerator for erosion testing, a disintegrator-type apparatus for impact wear testing, some tribometers for adhesive wear testing, a drum-type tribometer for fretting wear studying, a high-temperature erosive wear tester at protective gas environment.

- Further development of the analytical models of Beckmann and Gotmann (Kleis, Kulu 2008) to predict the erosive and impact wear considering materials hardness distribution and failure probability and taking into account material's response to dynamic loading (dynamic hardness). Based on experimental and analytical results mentioned above the new mathematical model for prediction of erosive wear performance of the bulk materials and coatings has been proposed (Kleis, Kulu 2008).

Research in the field of residual stresses in materials and coatings was oriented to development of:

- thermoelastic theory of non-homogeneous plates with multilayers;
- material layer removal method for determination of residual stresses;
- method of residual stresses determination in the multilayered coatings on cylindrical specimens from isotropic material (so-called force method).

The residual stresses in PVD coatings on nickel and steel specimens were determined experimentally. The residual stresses in other type of coatings (brush-plated Ag- and Au-coatings) on thin ring substrates were estimated. Methodology for uncertainty evaluation was proposed. The methodology for determination of the residual stresses in thick non-homogeneous coatings using a hole-drilling method was elaborated.

Research in the field of reliability of machines and products was oriented to the study of reliability of transport chains and sliding surfaces.
The main achievements are:

- the methodology for testing and an experimental device designed for simultaneous testing of 3 chains in the same controlled conditions;
- determination of wear resistance using the Wazau SVT500 tribometer working either in dry or lubricated 'pin-on-flat' sliding conditions ;
- the repeat coating technology of tools restoration by PVD and strengthening of worn parts by thermal spray, etc.

Recycling and recovery of materials

European waste recycling policy has the following hierarchy of waste management: prevention, reuse, recovery, energy recovery, incineration, and landfill. An integrated environmentally viable solution for recycling of printed wiring boards (PWB) is needed.

Laboratory of Recycling Technologies of the Department of Materials Engineering with a set of disintegrator milling equipment is unique in Estonia and the Baltic countries. This equipment is used in research related to recycling of different materials – metal and hard-metal scrap, ceramics, as well products – PWB composite plastics, etc. Different types of disintegrator milling systems were used to produce materials in the form of powders with different particle size (from some µm up to 5-10 mm) for materials recovery in different applications (powder metallurgy, pyrolysis, fillers in composites, etc.).

Most of the research in 2008–2011 was concentrated in PWB scrap recycling and materials recovery and oriented to the following problems:

- size reduction of PWB-s in disintegrator mills to liberate different materials;
- production of PWB powders with different particle size;
- materials separation (e.g. ferrous metals removal by magnets; non-metallic content removal by inertial-air classificatory);
- investigation of pyrolysis process of milled PWB powders in small TG-DTA scale;
- design and modelling of new composite materials with enhanced mechanical and tribological properties.

During the last years the research resulted in the following:

- The reprocessing technology for recycling PWB-s scrap in disintegrator mills was developed.
- Inertial air-classifier was developed for separation of metallic and organic materials from PWB powder. Advantages of air classification equipment are its basically 'dry' mode of operation without the use of any liquids to be cleaned from contaminants. It is essential that the feed material is in a narrow size-range to guarantee effective separation. The effective sizes of milled PWB powders for air separation of metals (Cu, Fe and Al) were determined.
- EDS analysis showed that most of the separated organic particles are composite materials, with plastic matrix and metallic, ceramic or fibre reinforcements.
- Pyrolysis experiments of PWB powders and 'PWB+additives' demonstrated a success in use of chemical additives (Fe_2O_3, $CaCO_3$) to control toxic substances (HBr and C_6H_6) evolving from the pyrolysis process, and to enhance the pyrolysis process. The mass of the solid products depended on the size of PWB powders and the additives. The pyrolysis liquid products mainly consisted of aromatic compounds with phenol and substituted phenol, but few alkane chains.

The recovered metals from PWB are in pure grades and thus more economic to recover than refining from ores. Mostly the economic initiative is oriented to recover precious metals (Au, Ag, Pd) and non-ferrous metals (Cu, Al, Pb, Sn).

Tribology & Tribosystems. Development of tribotesters

Wear and friction are essential processes for bodies in contact experiencing relative motion. Tribological tests can be performed in an almost endless number of ways. As the outcome of a tribotest is strongly related not only to the characteristics of the materials couple, but also to the whole mechanical system and its environment, the process of selecting the most appropriate test for a specific purpose is fundamental to making meaningful interpretations.

Laboratory of Tribology of the Department of Materials Engineering has several test devices for studying tribological performance of materials in a wide variety of conditions (dry and slurry erosion, two- and three-body abrasion, sliding wear in dry and lubricated conditions – continuous and reciprocating, surface fatigue, impact wear, etc.). There are also combined methods where two or more processes are taking place simultaneously (erosion and oxidation, abrasion and oxidation, sliding and oxidation, slurry erosion and corrosion, etc.). Other auxiliary devices such as optical microscope, scanning electron microscope equipped with energy-dispersive X-ray spectroscopy module, atomic force microscope, X-ray diffraction analyser, etc. are used to improve our understanding of the wear phenomena.

There is a trend of growing interest toward (1) wear testing at high temperature, (2) testing of coatings and surface layers at micro- and nanoscale, (3) development of laboratory test methods enabling testing materials or components in conditions close to working ones, etc.

A device for studying the erosion at temperatures up to 700 °C is efficiently used. A new device with possibility to introduce protective gases minimising oxidation has been developed and is in production at the moment. It allows studying the effect of oxidation on high temperature erosion rates. Another device has been developed to study the three-body abrasion at temperatures up to 1000°C with possibility to run experiment for several days and test 36 samples simultaneously in two abrasives. A device enabling sliding and abrasive testing at temperatures up to 450°C has been equipped with sensors for measuring coefficient of friction to reveal effect of additives providing self-lubrication properties.

Testing of coatings (especially thin ones) requires additional precise measurements. In situations of mild wear, the mass loss is often very small in relation to the total mass of the worn component. A precision balance typically has a resolution of 10^{-6} of the maximum load (e.g., 0.1 mg resolution at 100 g load), which naturally sets a limit to the minimum load possible to quantify in relation to the total weight of the component. In order to monitor the wear rate an analytical balance with repeatability of 0.02 mg, tactile and optical surface measuring station should be used. A new 3-dimensional optical system with improved precision has been ordered. A new device for surface fatigue testing has been applied for assessment of coating resistance to repeatable impacts (usually 10^3-10^7 impacts). A ball cratering device combined with micro-abrasion tester has been used for measuring the coating thickness, structure analysis and estimation of wear resistance. A new device for sliding test in various configurations with controlled environment (temperature, humidity) has been ordered. The technique allows single and repeated indentations; single and multi-pass scratching of material subjected to as low load as 500 μN has been used for measuring surface layer properties.

Device enabling adjustment of inertia and rigidity of loading system while keeping the load at the same level has been designed and successfully applied for testing of rough coatings. It also allows measuring vibration parameters (velocity, acceleration, amplitude) in required directions. Full scale device has been designed and used for bench tests of chains to study the effect of chain material, design and test conditions (abrasive, wood type and content, humidity, type of oil, etc.). High-energy disintegrator type impact wear tester has been modified. Centrifugal accelerator has been updated to study low angle (lower than 15°) erosion that is characteristic for straight ducts. Drum-type device has been built for high speed sliding wear conditions with extended length of wear track providing time for restoration of the surface layer between the subsequent sliding events.

Industry oriented research

To implement research and technological development (RTD) programmes and results the Department of Materials Engineering of TUT has been promoting cooperation with and between Estonian enterprises, research centres and universities. Since 2005 a long-term efficient collaboration has been conducted with Norma Inc and a contract has been extended until 2016. The cooperation is aimed at elaboration of the system and procedures to evaluate and check the quality of raw materials. The production of Norma includes safety components for several well-known car companies like Volvo, Lada, Mercedes, BMW etc. Since safety belt components meet extremely high requirements, the quality control is of paramount importance. In order to ensure the requirements the production process has to be smooth from fineblanking to galvanic coating and final assembly of the product. The main objective of the collaboration has been to find relations between material characteristics (mechanical properties, microstructural features and composition), cold forming process, thermal treatment and coating technology. In co-operation with the steel producers the best materials for production of the safety belts steel components have been developed. The project also incorporates fundamental research related to development of thin hard coatings for tooling and their tribological behaviour. The research is focused on tools working in extreme conditions used in fineblanking. Following the purchase of Norma by Autoliv Inc more research has been conducted and the results have been introduced worldwide.

In the field of cermets technology and wear the most active collaboration has been established with the international company Metso Minerals OY (Finland). In 2003 a research concerning wear resistant materials was launched. Metso is one of the world's biggest corporations offering jobs for 25,000 people in about 50 countries. Our collaboration has been a great success partly due to Metso's long experience in co-operating with Finnish universities (Aalto University, Tampere University of Technology). During the past years several PhD theses (abrasion and erosion resistance of metal-matrix composites, fatigue strength of powder steels in aggressive media, etc.) initiated by Metso have been defended and the results have been published.

The aim of the biggest project "Cermets for Wear Parts" has been to develop new cermets and their manufacturing technologies. Those materials can be used in mining industry for producing grinding and milling machinery and in facilities handling mineral materials.

Traditionally manganese steel and white cast iron are used in grinding equipment. Both materials have several disadvantages. Metal-matrix composites and ceramic metal composites have proved competitive in many appliances in abrasive and erosive wear conditions. Reliability characteristics of different cermets and metal matrix composites have been found out in course of the project running, their implementation possibilities have been analysed, and fabrication technology developed and tested. A technology of WC-based cermets produced by means of reactive sintering has been elaborated. The new cermets can be implemented as alternatives for new generation milling devices or they can upgrade materials used until now.

Moreover cermets technology has been of great interest to the Estonian Ministry of Defence. A research ordered by the Ministry for developing new ballistic armour materials has been conducted at the Department of Materials Engineering. Since the development of military technology has brought along speed increase of hand weapon bullets and implementation of hard core (hardened steel or hard alloy) armour penetrating ammunition bullet proofness cannot be ensured by polymer matrix composites (aramide, diolene, UHMWPE, Twintex, S-2 glass, etc.). The aim of the research was to provide an overview on implementation of light armour in military technology, introduce materials and technologies for producing light armour and requirements and experimenting standards and possibilities to produce light armour panels in Estonia. It is known that in order to stop a hard core (hard alloy or hardened steel) bullet the armour has to contain materials the surface hardness of which is close to that of the bullet. During the past 15 years the usage of ceramics in armour materials has considerably increased. Ceramic materials with the best characteristics are hot pressed boron carbide B_4C, silicon carbide SiC and beryllium oxide. These materials are characterised by extreme hardness on the one hand and light weight on the other. The biggest disadvantage is the high price of the material. Due to cost efficiency sintered ceramic materials are preferred in armouring vehicles. Light weight and excellent ballistic defence characteristics are advantages of ceramics compared to steel. That enables to use cheap glass fibre reinforced composites or metal sheets as basic materials. As a result of the research it was revealed that due to ultra-high hardness, corrosion resistance and low density TiC-based composites can be considered perspective materials for light armour panels and wear resistant details. Prototypes were produced and conducted pre-experiments proved to be promising. Hopefully, technology development in the nearest future enables to start producing armour panels in Estonia.

FUTURE ORIENTED R&D

According to the trends in materials technologies in Europe the new research projects will be initiated in the field of key technologies. The main focus is advanced multiphase materials and production technologies (composite materials and coatings including carbon based materials; ultra-fine and nanostructural materials and related technologies; rapid manufacturing and prototyping technologies, e-manufacturing, etc.).

Nanomaterials and nanotechnologies

"There is plenty of room at the bottom" – this bold and promising statement from Nobel laureate R. Feynman launched the Nano age. Nanomaterials and nanotechnologies have been developed and currently remain under development as a consequence of truly significant recent advances in the materials science. Hopes exist for being able to make more reliable materials of needed characteristics while cost-effective and environmentally friendly.

Nanocomposite materials as an enormously wide class of composites is thought to possess unique properties for serving many industries especially under severe conditions of exploitation. The general idea behind the addition of the nanoscale second phase particles is to create a synergy between the various constituents, such that novel properties capable of meeting or exceeding design expectations can be achieved. Among the various nanocomposites types, fine-grained ceramic-metal composites (cermets) reinforced with nanoscale structures of the third phase are expected to have superior mechanical properties as compared to conventional composites. However, many technological advances are limited by the impossibility to combine high mechanical performances with critical functional or structural reliability. The reasons include inadequate dispersion and alignment of the reinforcement, poor bonding and load transfer at interfaces. One of the challenges is design and development of the pore-free composites with strong interfaces between constituents resulting in high toughness, increased wear resistance and damage tolerance, which will enhance their functionality and utility. The main drawback to be overcome is grain growth during processing. Using sinter/HIP, HIP, high energy mechanical attrition techniques and sol-gel processing as well as incorporation of specific sintering additives (grain growth inhibitors) allow producing novel composites of the tailored properties.

The influence of the grain boundaries or interfaces for composite properties evaluation is no doubtful. Significant fraction of atoms at or near grain boundaries results in alteration of mechanical properties leading to a substantial increase in strength and hardness. To play with the fracture toughness, the additional mechanisms should be considered for composite design. One of the current research grants in the Department of Materials Engineering exploits the idea that carbide based structures can be dispersion-strengthening by nano-crystalline oxides/carbides of transition metals and toughening by either stabilised zirconia through the mechanism of phase transformation or by fibres through mechanisms of cracks deflection during loading. The study covers a design, production and characterisation of novel composites reinforced by third phase inclusions such as ZrO_2, ZrC, NfC and the newly developed tubular structures of TiO_2 and ZrO_2.

Another great challenge in novel composite structures development is obtaining the nano-scaled ceramic-based material exhibiting superplasticity at relatively low temperatures.

Interest in superplasticity is extremely high. Superplasticity, the ability of some materials to undergo very large tensile strains, is a subject of intensive research because of its clear commercial applications especially for the shape-forming of engineering materials resulting in simplification of the production process and its cost. The goal is to produce near-net-shape parts to avoid the expensive machining of ceramics. Superplasticity is also an important field of scientific research both because it presents significant challenges in the areas of flow and fracture and because it forms the underlying basis for the commercial superplastic forming industry in which complex shapes and curved parts are formed from superplastic sheet metals.

The most important point for the superplastic ceramics development is the establishment of the interaction between intragranular slip and grain boundary sliding in the process of deformation of the materials that are expected to possess the superplasticity. The interaction between lattice dislocations and grain boundaries is significant in explanation of the dominant contribution of sliding along grain boundaries during deformation. On the basis of the latest concept in the physics of high-angle boundaries, it is shown that grain boundary sliding and migration are mechanisms responsible for superplasticity and the rate of this process determines the macroscopic strain rate and possible elongation of the material. To make these mechanisms capable for functioning, the ultra-fine and/or nanostructured materials should be taken into consideration. It is because cavitations due to grain-boundary sliding must be accommodated by diffusion and/or dislocation processes for successive deformation, a short accommodation length, meaning a small grain size, is essential for attaining high-strain-rate superplasticity. For the same reason, stability of the small grain size is also important. If grain growth occurs actively during deformation, the accommodation length increases and retards facile grain-boundary sliding. This causes an increase in the level of stress necessary for successive deformation, that is, strain-hardening. Strain-hardening enhances the extent of stress concentration on the sliding grain boundaries or grain corners, resulting in the formation of intergranular cavities that leads to premature failure.

Nanocarbon and related technologies_

Diamond is an ultra hard material with unique combinations of properties including high mechanical strength and thermal conductivity, outstanding optical and electronic properties, exceptional chemical inertness.

There has been an increasing interest in smooth diamond films produced by plasma enhanced chemical vapour deposition (PECVD) technology on metallic and ceramic substrates for use in a variety of engineering applications. The highest hardness, stiffness, thermal conductivity, low friction coefficients, as well as chemical inertness to acidic and saline media far exceed those of any other known material. The unique combination of these properties provides an ideal solution for tribological applications.

Most of applications require an optimisation of diamond properties to fulfil simultaneously different functions: hardness, surface smoothness, optical transparency, electrical conductivity, field emission, etc. The objective assumes investigation of the properties of diamond films doped with different impurities for development of technology for all-diamond atomic force microscope (AFM) probes for long-lifetime scanning with high force constant to monitor electrical properties of the surfaces possessing high hardness.

The application of AFM probes for SSRM technique is a primary aim of the project. SSRM technique assumes testing electrical properties of the materials in a contact mode. Other possible applications include microelectromechanical system (MEMS), Oxidation Nanolithography, Electro-Chemistry AFM, Piezo-Force Microscopy.

To achieve a good contact between the tip and the surface a high force should be applied on the tip. Strong contact between the tip and surface means stability of mechanical and electrical properties of the tip under high contact pressure. Contact electric techniques generally require the absence of a gap between the tip and the sample and sufficient area for ohmic or capacitive contact, which is related to higher tip-sample forces and cantilever spring constants. During the scanning of the surface, the tip should be as sharp as possible. Therefore, to reduce the wear of the tip, the friction coefficient should be as low as possible. Materials of high chemical inertness are suitable for this purpose. To summarise, most important properties include chemical inertness, low wear rate under high load, high conductivity and low coefficient of friction of the tip material.

Diamond Like Carbon (DLC)

Tribological properties of DLC are affected by structure and mechanical properties of both DLC and counterbody and parameters of a tribological test. Several mechanisms were suggested to explain the observed results. For non-hydrogenated DLCs, it is generally agreed that low coefficient of friction (CoF) in ambient air is caused by adsorbed gaseous species, which passivate carbon dangling bonds. Wear and/or elevated temperature induce micrographitisation of DLC resulting in increase in CoF during test running. A number of authors have claimed formation of a transfer layer reducing friction.

The main focus of the research is evaluation and optimisation of tribological properties (CoF, wear) of DLC coatings. Behaviour of tribological system during steady state may depend on the behaviour during running-in. On the other hand, the tribological properties of tribopair depend on wear test conditions. Understanding of influence of the wear test conditions on the behaviour during running in and subsequently steady state allows optimising DLC coatings properties for particular applications.

Carbon Nanotubes (CNT)/Nanofibers (CNF) growth on top of hard coatings

Nanotubes can hardly interact with other materials. Low reaction temperature and the high-purity of alcohol catalytic chemical vapour deposition (ACCVD) technique guarantee easy and cost-effective scale-up production. Furthermore, the reaction temperatures lower than 600 °C ensures that this technique can be applied to the direct growth of CNT/CNF on semiconductor devices previously patterned with aluminium.

The project "Self-lubricating carbon nanofibers tribological coatings" is focused on growth of CNT/CNF on the surface of hard PVD coatings. The results obtained definitely show a decrease in the CoF measured on the samples coated by CNF as compared to the uncoated ones. Instabilities of the CoF value observed during the tribological test are likely due to the morphology of the nanocomposite surface. The droplets are formed on a surface of hard coating (nACo) during arc deposition. The size of defects can be up to some micrometres, while the thickness of CNF is only a few layers of nanofibres. Therefore during the fretting test, thin nanofibres layer can be suddenly removed.

Further development of the ACCVD based technology includes the deposition of CNT/CNF on smooth hard coating surfaces, for instance Co-based hard coatings. In addition, the deposition of SWCNT (Single-Walled Carbon Nanotubes) and MWCNT (Multi-Walled Carbon Nanotubes) is also scheduled.

Acknowledgements

This research was supported by target-financed projects SF0142082s03, SF014505s03, SF0140062s08 and SF01400091s08 of the Estonian Ministry of Education and Research

REFERENCES

Antonov, M., Hussainova, I. 2009. Experimental setup for testing and mapping of high temperature abrasion and oxidation synergy. Wear, 267, 1798-1803.

European Technology Platforms on Advanced Engineering Materials and Technologies (EuMaT). 2011. http://eumat.eu/

Hussainova, I., et al. 2009a. Micromechanical properties and erosive wear performance of chromium carbide based cermets. Wear, 267, 152-159.

Hussainova, I., et al. 2009b. Effect of stressed state on erosive performance of cermets with core – rim structured ceramic grains. Wear, 267, 177-185.

Hussainova, I., Hamed, E., Jasiuk, I. 2011a. Nanoindentation testing and modeling of chromium carbide based composites. Mechanics of Composite Materials, 46, 6, 667-678.

Hussainova, I., Schade, K. P. 2008. Correlation between solid particle erosion of cermets and particle impact dynamics. Tribology International, 41, 323-330.

Hussainova, I., Smironov, A., Antonov, M. 2011b. Mechanical characterization and wear performance of WC-ZrO2-Ni cermets produced by hot isostatic pressing. Advanced Materials Research, 214, 344-348.

Klaasen, H. et al. 2010a. Adhesive wear performance of hardmetals and cermets. Wear, 268, 1122-1128.

Klaasen, H., et al. 2010b. Reliability of dual compounds "carbide composite + steel" produced by diffusion welding. International Journal of Refractory Metals and Hard Materials, 28, 580-586.

Klaasen, H. et al. 2011. Performance of carbide composites in cyclic loading wear conditions. Wear, 271, 837-841.

Klaasen, H., Kübarsepp, J. 2006. Abrasive wear performance of carbide composites. Wear, 261, 520-526.

Kleis, I., Kulu, P. 2008. Solid Particle Erosion. Occurrence, Prediction and Control. Springer, London.

Kollo, L., et. al. 2010. Investigation of planetary milling for nano-silicon carbide reinforced aluminium metal matrix composites. Journal of Alloys and Compounds, 489, 394-400.

Kommel, L. 2010. Properties development of ultrafine-grained copper under hard cyclic viscoplastic deformation. Materials Letters, 64, 1580-1582.

Pirso, J., et al. 2009. High temperature erosion of Ti(Mo)C-Ni cermets. Wear, 267, 1894-1899

Pirso, J., et al. 2010a. Three-body abrasive wear of TiC-NiMo cermets. Tribology International, 43, 340-346.

Pirso, J., Letunovitš, S., Viljus, M. 2010b. Method for making chromium carbide-nickel composites. Pat. No. EE05371B1 (15.12.2010).

Roadmap of the European Technology Platform for Advanced Engineering Materials and Technologies. 2011.

http://eumat.risk-technologies.com/filedown.aspx?file=3229

Sergejev, F., et al. 2007. Correlation between surface fatigue and microstructural defects of cemented carbides. Wear, 264, 770-774.

Sergejev, F., et al. 2008. Fatigue mechanics of TiC-based cemented carbides. Proceedings of the Institution of Mechanical Engineers. Part J, Journal of Engineering Tribology, 222, 3, 201-209.

RESEARCH INTO COGNITIVE ARTIFICIAL SYSTEMS IN ESTONIA

L. Mõtus

The paper surveys the state-of-the-art of research in cognitive artificial systems, developments in pragmatic interests toward cognitive artificial systems, the existing research groups, their cooperation and plans for the future in Estonia. The presentation is illustrated by samples of existing research results, basic unsolved problems, and concise discussion on future research trends in cognitive artificial systems. Since cognitive artificial systems are novel in Estonia, a number of descriptive examples of operating cognitive systems have been included. This paper focuses on research issues in conventional cognitive artificial systems (software intensive artefacts, networks of those artefacts, interacting colonies of biological creatures enhanced by embedded computer systems, etc.), and avoids speculations related to cognitive artefacts potentially emerging from applying the results of synthetic biology.

BACKGROUND OF COGNITIVE ARTIFICIAL SYSTEMS

Cognitive artificial systems (see for details Vernon et al. 2007) are further in this paper interpreted as dynamically collaborating communities that exhibit adaptive, anticipatory and proactive goal-directed behaviour, comprise software-intensive artefacts and (may be) biological creatures. The operation of cognitive artificial systems is essentially based on the use of feedback (Rosenblueth et al. 1943). In other words, cognitive artificial systems are teleological systems composed of interacting and collaborating, autonomous components of different origins, whereas majority of those components are sufficiently autonomous in their decisions, and capable of:

- acquiring goal-oriented knowledge (a.k.a. SITUATION AWARENESS) from their ambience by the use of perception, reasoning, and 'intuition' based on the previously acquired experience and knowledge;
- coordinating and harmonising their knowledge with the knowledge obtained by their partners in the system (a.k.a. FORMING TEAM SITUATION AWARENESS), considering also their own personal goal functions and the system's goal functions; this means that the components do not only share their individual knowledge with the others,

but they are also ready to ameliorate their own knowledge, or even their short-time goals if this increases the probability of achieving the system's goal.

Cognitive science builds upon many research disciplines, e.g. psychology, philosophy, neuroscience, linguistics, anthropology, sociology, education, its methods are often combined with those of artificial intelligence and computational science. Historically the first cognitive systems were noticed among biological creatures; hence the research in cognitive science is essentially biased towards naturally evolving systems of biological origin. The attribute 'cognitive' assigned to an artificial object can be vaguely justified since the 1960s due to the first rise of interest to developing and application of artificial intelligence methods in adaptation, self-learning and self-organisation related experiments. Since the 1960s the sensory, reasoning, and actuating capabilities of computers have been persistently improved in the process of building and intensive application of embedded computer systems. With mild reservations one can claim that some cognitive research was applied to individual computer-based applications (e.g. to embedded systems) and have become visible on abstract levels – as systems' developing methods and operational decision-making methods in the artificial world (e.g. cognitive engineering, see Hall et al. 2006).

Cognitive artificial systems emerged into the practice in mid-1990s, due to increasing networking of autonomous computer systems (e.g. building systems of systems) from components that so far had been operating in standalone, closed-loop mode within natural and/or artificial systems. This process is named in the academic world as 'building networked pervasive computing systems', also known as 'cyber-physical systems'. Pervasive computing systems (a.k.a. ubiquitous computing systems) are usually interpreted as processors embedded in physical objects with inputs and outputs connected to those objects, exchange information with other processors and/or objects, and are constantly operational. Each node of such network could be an autonomous computing system, with its own sensory system, its own executive system, its own reasoning system that includes memory for storing knowledge and past experience, and with its own communication system for information exchange with other nodes in the network, and with the environment. At the same time each node in such network can be considered as an entity that is capable of cognition-based behaviour and may have some similarity with the 'intelligent and proactive' behaviour of social biological creatures (e.g. humans).

As part of the network theory (see for details Cohen, Havlin 2010) social networks are used to study properties of social structures emerging in the community of interacting individuals (mostly of human origin) and their organisations. Social networks are often used as test cases for studying hypotheses and theories regarding epidemiology, rumour spreading, diffusion of innovations, and other similar domains. Networked autonomous pervasive computing systems exhibit many properties characteristic to conventional social networks – e.g. the process of forming team situation awareness, spreading information, diffusion of innovations, susceptibility to viruses, etc. The interest towards studying social behaviour in cognitive artificial systems and impact of cognitive artificial systems on social behaviour of humans is increasing proportionally to the level of autonomy granted to autonomous cognitive components of artificial systems, and to the level of mastering the self-X functionalities in components and systems of cognitive artificial systems.

Synthetic biology has opened new potential perspectives for building cognitive artificial systems (see a survey EASAC 2010 for more information) but these perspectives are not discussed in this paper.

THE IMPACT OF COGNITIVE ARTIFICIAL SYSTEMS ON DAILY LIFE

The cognitive capability in artificial systems springs up gradually. Embedded systems, operating in a stand-alone mode (e.g. for partial control of technological processes, monitoring and partial control of environment) in the 20th century operated under strict supervision of human operators. Those systems were not believed to be sufficiently dependable to grant them autonomy in decisions that might have system-wide influence. Hence their already existing cognitive capabilities (e.g. sensors, actuators, and embryonic reasoning) were used only in well-defined situations, within heavily restrictive safety envelope, and only under human supervision. The human-supervised and/or -controlled embedded systems are still being applied for solving comparatively regular and not too complex tasks. Consider, for instance, subsystems that provide feedback-based solutions to tasks in conventional robotics, in conventional flexible manufacturing systems, in common telecommunication equipment, in software-intensive devices for laboratory experiments (e.g. in genetics), in monitoring and assisting tools used in medicine. Such systems are usually not called cognitive artificial systems, although they do possess embryonic cognitive capability.

However, the complexity of all the above listed applications increases rapidly, at the same time the level of incompleteness of accessible information increases. The stand-alone autonomous embedded systems are being rapidly networked, thus increasing their interdependence and fostering more sophisticated and distributed reasoning. The collaborating, networked embedded systems are more capable to operate in dynamically changing environments in spite of the incomplete information. Flexible in-network collaboration supports the system's ability to implement self-X features, such as coping with non-lethal failures by dynamic reconfiguration of the network structure, and/or functionality in nodes (self-healing, self-organisation, etc.). Due to humans' physical (e.g. strength and response time) and mental (e.g. in-depth, rapid reasoning ability) limitations, the response to environmental changes and modification of the systems behaviour needs to be provided automatically – e.g. in order to get shorter response time to changing requirements, and/or to survive sufficiently long in a hostile environment. The artificial systems that are capable to overtake the responsibilities from human-operator in order to improve the probability of achieving the goal function – by (may be temporarily) substituting human-operator in order to provide better response time, better endurance in hostile environmental conditions, or better performance under heavy mental stress – are usually called cognitive artificial systems (a.k.a. proactive computing systems).

Some illustrative (and at least partially operating) sample systems, where the role of humans has been severely reduced by enhancing cognitive capabilities of the respective software-intensive system and shifting (a remarkable part of) decision-making from humans to artificial cognitive system, are discussed in the following. Those systems have been built by means of classical 'error and trial' method, since a number of essential theoretical foundations, behavioural subtleties, and methods for verifying behavioural features of respective artificial cognitive systems still need further basic research. Some of those essentials will be discussed later in this paper.

One of the eldest practical attempts in building cognitive artificial systems is the parking assistant – its development started already in the 1970s. The task was theoretically solved for various types of cars and trucks, and the solution worked perfectly in simulators. However, the commercial applications appeared only in 21st century. Today automatic parking feature is provided, for instance by BMW, Ford, Lexus, Mercedes, Toyota, and Volkswagen. The autonomous parking system covers parallel and back-in parking, nevertheless implementation details and the required human's participation may differ from manufacturer to manufacturer.

Air traffic control system is still under general supervision of humans. Nevertheless, many aspects of decision-making procedures have been fully automated and are operating autonomously – for instance, aircraft trajectory's envelope calculation and analysis of aircraft collisions probability, pilot's intention assessment, and pilot's experience and health assessment systems. All those automatic subsystems provide information to human supervisor for the final decision/action. Additional features, legally not covered today are those emerging from the joint use of airspace by aircrafts piloted by humans, and by a variety of unmanned aerial vehicles. The ongoing development of the advanced air defence systems promotes the architecture of networked highly autonomous systems in collaboration with the conventional civil and military decision-making bodies (see more details in Motus 2010).

Separate subclass of similar tools have been developed and applied in defence against terrorism for detection of suspicious behaviour in public places, for remote monitoring of traffic on the motorways in order to detect drunken or overtired drivers, detection of stolen vehicles, etc.

The impact of components' autonomy and topological properties of their interactions on the behaviour of integrated systems have become a serious issue in systems of systems. System of system is a second order system where components used to build the system may include autonomous systems with their own goals and rules. Such architecture is wide-spread in systems formed by Nature but is comparatively new in the artificial world. System of systems usually exhibit emergent behaviour which is one of the characteristics of complex systems. Emergent behaviour is generated dynamically and cannot be deduced from the specifications of components, or from stationary structure of components' interactions. The hypothesis is that emergent behaviour is enabled by autonomous behaviour of components, dynamically changing topology of interactions, and incompletely known impact of the environment upon the system of systems. Widely accepted methods for detecting and influencing the emergent behaviour are still absent today and are one of the research topics in the theory of cognitive artificial systems – those features cannot be reasonably handled in natural cognitive systems either (unless we accept the existence of God, who is capable of controlling the emergent behaviour).

Intuitively the concept of system of systems can be illustrated by analysing the activities carried out during military operations. The efficiency of warriors can be increased by artificial system that builds and supports exploiting networked enabled capabilities (NEC) that foster well-coordinated actions of groups of warriors, fusion of information acquired from a variety of sources, validation and fusion of that information into a more abstract knowledge about the actual situation, sharing that knowledge, decision-making processes for modifying that situation, and executing those decisions.

Any activity that requires coordinated action of several entities, e.g. playing football, designing and building a complex device, planning the economical development of a country and others can be formulated as a NEC problem. However, the militaries form the largest community that has adopted the term NEC to denote explicit development of (semi-)automatic coordination and synchronisation capabilities for improving their winning probability. Encouraging evidence about the efficiency of NEC can be found, for example, from the Stryker brigade report (by RAND 2005). Some aspects of cognitive artificial systems as applied in the tactical NEC context have been discussed in Motus et al. (2008) and Motus et al. (2006). The NEC development poses a multitude of yet unsolved theoretical problems, often on the borderline of artificial and natural world – for instance, situation aware computing as a tool for reducing complexity, shared situation-awareness between autonomous interacting partners, activity coordination in multiple times and space, and others.

The NEC problems and other similar applications lead us to research problems that hitherto have been studied only in anthropocentric systems – e.g. monitoring, forecasting, and influencing the behaviour in social networks. In our case the social networks may include humans and cognitive artificial systems as equal partners. Interaction networks have been studied for some years already in the case of pure human based (e.g. conventional social networks, where computer is just a tool for communication and archiving information) and in pure artificial systems, where some computers have the capability of cognitive agents. To illustrate the latter case, imagine the not too sophisticated flexible manufacturing, or advanced logistics systems where the main body of functionality related to remote monitoring, diagnostics, self-healing, the decision-making and information-sharing, and other activities, is carried out by networked computers – usually without the interference from humans. The research of joint functioning of mixed interaction networks comprising humans and cognitive artificial systems as equal partners is still at its infancy.

All the cognitive systems (natural and artificial) are essentially complex systems. Multi-agent systems are often accepted as models for describing and analysing by simulation of large subclass of complex systems that are not susceptible to conventional mathematical methods, e.g. differential equations. The most successful approach to using multi-agent models for handling practical-sized problems is known under the name 'synthetic environment for analysis and simulations' – the idea is to build and maintain a continuously running, continually updated mirror model of the real world that is more efficient in predicting and evaluating the future than the conventional methods of extrapolation. This approach has been applied to simulate the future evolution in diplomatic, economic, political infrastructure and social issues.

SOME RESEARCH ISSUES IN COGNITIVE ARTIFICIAL SYSTEMS

Cognitive artificial systems are typically software-intensive, inevitably comprising computers or computer networks, immediately connected to sensors, and in many cases also to effectors (a.k.a. actuators). Those systems can interact with their environment and operate without human intervention for remarkable time periods. This means that cognitive artificial systems need to be self-sustainable – able to adapt to potential changes in the environment, able to self-learn and self-heal, accumulate experience, reason to find a way of achieving its goals, etc. In order to understand the operation subtleties in artificial cognitive systems, one needs to merge knowledge from many research disciplines, which usually are handled separately from each other. In the following I survey several aspects of cognitive artificial system that are being studied, or are scheduled to be studied in Estonia in cooperation with the rest of the world.

Computational issues

The application of computers has evolved from transformational data processing systems (from 1939 forward) to reactive systems (starting from the end of 1950s and maturing in 1990s), and since the end of 1990s is moving towards cognitive (a.k.a. proactive) computing systems. This evolution is discussed in more details in research report of the Lab for Proactive Technologies (ProLab report 2009). At the same time the conventional computational model forming the foundation of theories, methods, and tools widely used in practice still stems from the Church-Turing thesis, suggesting the use of Turing computable functions, and assumes satisfaction of axioms that hold for Turing computable functions. The frequent violation of those axioms in contemporary computer applications reduces the dependability of theories built on those assumptions, and in some cases makes these theories not applicable for behaviour analysis and verification. Since the appearance of reactive computing systems several axioms related to Turing computable functions are regularly violated, for instance (Motus et al. 2005):

- non-terminating (on-going) computation has become a rule, not an exception as in conventional data processing; reactive and proactive computing systems often need to process streams of indefinite length, not strings;
- non-interference assumption that is mandatory in simultaneous execution of several Turing machines (e.g. in case of algorithmically parallel computations) cannot be satisfied in forced concurrent

mode (usually imposed by the environment) that is a characteristic operation mode for reactive and proactive systems;

- Turing computable functions (and Turing machines) have no historical memory, whereas many computing processes in reactive and proactive systems are essentially situation-aware (a.k.a. context-aware) which requires historical memory.

Since 1980s many researchers have been working on relaxing the axioms that are persistently violated in practice, and thus are seriously reducing the applicability of commercially available verification tools. One perspective direction is the development of interaction-based models of computation which, in a long run, leads to multi-stream interaction machine to be used in building and formal analysis of reactive and proactive systems, in addition to Turing machine. As a by-product, interaction-based model of computation enables the explicit use of stream processing and also enables the expansion of the stream variables' values with the on-line added metadata – for instance, linking the variables to specific time intervals, to specific locations, or to some other non-functional characteristics (e.g. stress level of the human user) – this linkage will support on-line validation of the variable values, or demonstrate the consistency and coherence of the streams and/or their elements (Motus et al. 2005; Dosch et al. 2007). This class of problems attracts persistent attention of researchers working on formal behaviour analysis of cognitive artificial systems and on more sensitive cases of pervasive computing systems.

Systems integration and dynamic evolution of systems

Cognitive artificial systems are built from autonomous software-intensive components that interact with each other and with their environment, but usually do not form a fixed structure of interactions. As a rule, the systems are capable to modify the interaction pattern of their components on-line, and even add or remove some components during their operation. All the components may have passed pre-integration verification and testing, nevertheless the integration process may reveal inconsistencies that lead to integration errors. Usually those subtleties cannot be detected by the conventional verification and testing methods. The amount and seriousness of errors revealed at integration increase as autonomy, proactive/cognitive behaviour, and dynamic variability in the interaction pattern of components increase. Many of those undetected errors are caused by emergent behaviour that is inevitable in complex systems – the commercially available verification methods and tools, as well as the mainstream research results do not readily support the analysis of systems' properties emerging in the process of integration, or due to dynamical evolution of the integrated system (or its environment) during the system's operation.

For instance, in typical multi-agent systems the emergent behaviour is so conspicuous that it cannot be neglected any more. Such observations have lead to joint studies of the theories for computer systems and for complex systems. The insufficient understanding of the essence of emergent behaviour and insufficient ability to (at least partially) control the emergent behaviour in systems, hinders the behavioural management of cognitive artificial systems and is also one of the central issues in managing complex systems. It has become obvious that the theories in computer science cannot forever address the ideal, completely known (e.g. Turing computing based) world. The denial of the influence of incomplete information on computing, amplified by hard-to-predict decisions from autonomous and proactive components on the system's behaviour has become unreasonable and dangerous for the surrounding real world. Some progress in handling those hitherto neglected properties has been achieved, for instance introduction of multiple times and several simultaneously used time concepts, explanation of seemingly random occurrence of (timing) errors, that have paved the way to interaction-centred and situation sensitive models of computation.

The research of cognitive artificial systems is further complicated by necessity of permitting the dynamic on-line evolution of those systems (i.e. change its own composition, change its interaction pattern, change its guiding goal-functions, etc.) as suggested by their cognition-based reasoning subsystem, and/or required by their environment. Ideally one should verify the system's behaviour after each evolutionary step – but on-line dynamic verification of behaviour is, in general, a hitherto unsolved task. In order to become able to verify the systems' behaviour on-line, one needs to develop a suitable super-Turing model of computation (see section "Computational issues" of this paper, and suggestions in Motus et al. 2005).

In Motus (1995) a formal framework has been suggested for detecting a number of timing errors and some other inconsistencies invoked by integration of autonomous components and not detectable by conventional testing. Motus (2010) discusses the impact of essential features that are to be handled by cognitive artificial systems – for instance, incomplete information about the environment, autonomy of components, types of interactions (e.g. direct, indirect, mediated) actually applied in the system, and proactivity of the components/system – on system's properties and behaviour. Research in true self-organising systems and analysis of other self-X properties is being prepared in the Research Lab for Proactive Technologies – e.g. ability to on-line behaviour verification and other supporting tools is an essential precondition – we are still looking for partners in that area.

Individual and team situation-awareness

Managing the complexity of modern world requires collecting, organising and reasoning about a vast amount of information from a variety of sources. For instance, think of the applications in managing production chains in manufacture, business intelligence systems in networked enterprises, network enabled capabilities and the resulting systems of systems in the military community. Cognitive artificial systems have been, not so seldom used to help humans to manage the complex non-linear reasoning problems. Back in the 1970s the theory of automatic control reduced the complexity of non-linear control by introducing sliding mode control – altering the dynamics of a continuous time non-linear system by applying a discontinuous control law – e.g. the control law changes at automatically detected switching points and providing thus a reasonably good and less complex control law. Situation-based control and management of complex systems is an abstraction of sliding mode control, the difficulty of decision-making is partly shifted to defining and detecting of situations, whereas within each situation the control rules are less complex to compute. In behavioural sciences, the (not formalised) notion of situation has been pivotal for long time already.

Majority of researchers still view situation-awareness and situation management from the perspective of human requirements and human processing capabilities. Conventionally a situation is the aggregate of biological, psychological, socio-cultural, and environmental factors acting on an individual or a team of agents to condition their behavioural patterns. In the context of artificial systems we redistribute the focal points in a definition of situation – we slightly reduce the importance of biological, psychological, and socio-cultural characteristics of a situation and pay more attention to the details of environmental factors. We partition the environmental factors into, for instance:

- time issues – e.g. how to cope with a multitude of time counting systems and with a variety of simultaneously used time concepts;
- space issues – e.g. position of system's agents in physical space, processing information from agents moving in a physical space, plus more abstract problems related to conformity of decision-making spaces, and security issues in cyber space;
- ontological issues – e.g. matching domain ontologies; in the case of humans this is believed mistakenly to be a minor problem, however it needs serious attention in a network of computers and software-intensive devices;
- rights of system's agents to access and process situational information; also their capability to process situational information, e.g. automatic data validation, and data consistency verification before data fusion.

Individual situation-awareness – each cognitive agent (a component of cognitive artificial system) has its own sensors, effectors (actuators), and may use data provided by fellow-agents; it also has a limited processing capability and its own dedicated knowledge base. A cognitive agent is able to detect and validate situations with respect to its interests and requirements. The agent interprets situation's impact on its behaviour and on the probability of achieving its goals – it uses the situational information for proactive decision-making, it may attempt to change or modify the situation via its effectors, and it may communicate its reasoning results to other interested agents in the system in order to launch collaborative actions. The validation of situations is not a trivial task – in anthropocentric systems the validation is often neglected as too laborious, relying implicitly on human intuition. In computer-centric systems validation of any detected situation is crucial.

Validation starts by analysing variables that characterise situation one by one – e.g. from the acquisition of reading raw sensor data, tagging it with time instant (in the agent's time) and with estimated sensor position in physical space, calculating the reading's validity interval (pending on the agent and its goal), checking conformity with the other known evidences and agent's requirements. The following step is to analyse coherence and cohesion of the reading's value with other readings in the stream, and with the values of other variables characterising the situation. The valid situation provides data that can be trusted in further, more abstract, processing and decision-making. Similar checks are to be repeated when one wants to merge several, already validated situations into a more abstract situation.

Our research in situation-awareness is illustrated in (by Preden et al. 2011), and a more thorough coverage of the topic is provided in PhD thesis by J.-S.Preden (2010).

Team situation-awareness – it may happen that a collection of agents with individually validated situations performing actions, which follow from individual reasoning about those situations, never reach the system's goal – although each component of a system may reach its own respective goals. The system's goals do not necessarily match with the goals of individual agents that form the system. In real life the system's goal has higher priority than any goal of a single agent (component of a system). In order to guarantee the achievement of the system's goal one needs to harmonise the interpretation rules of situations' impact, and the reasoning processes about decisions to be used by individual agents so as to increase the probability of reaching the system's goal – which may require neglecting some of the interests of an individual agent (components of the system). Harmonisation of component's goals is a dynamic process; its outcome is comparable with the emergent behaviour in a complex system and cannot be prefixed during the system's design. In many cases some components in a system need to be substituted, or their goal functions need to be modified in order to reach useful team situation awareness.

The harmonisation process leads us to negotiations between system components – usually it requires compromise between the components' interests. These negotiations should, in the case of pure artificial cognitive system, be carried out without humans. There is still a long way to generally accepted methods of achieving group situation-awareness.

Sensors, proactive middleware, and unified interface

Any cognitive system needs perception that collects information from the environment by interpreting sensory signals that result from physical stimulation of sensors. Hence sensors are an important part of cognitive artificial system. Sensor is a device that converts a physical quantity into a proportional signal whose strength can be measured, and communicated for further processing. Sensors are designed to have as small as possible effect on what is measured. Today many sensors are manufactured in microscopic scale – e.g. using micro electromechanical systems (MEMS) technology. Cognitive artificial systems prefer sensors that have been integrated into a package comprising sensor, converter to a standard digital signal, processing device for tagging and pre-processing the signal, communication device for distributing the sensor readings, and a power source. Such packages are called smart sensors, or smart dust, etc. Smart sensors used to monitor the overall status of the environment form intelligent wireless sensor networks. The latter is a subdomain of mobile *ad hoc* networks (MANET). Today, mobile *ad hoc* network is the most popular architecture used for building systems from autonomous, proactive, software-intensive devices. Cognitive artificial systems are a special case of those systems.

The Lab for Proactive Technologies runs experiments with the smart dust type of sensors, with mobile ad hoc networks, with navigating mobile agents in an unknown environment by using coordinate network generated by beacons; the Lab also develops methods for positioning mobile and stationary agents, and develops and tests signal processing methods on a variety of applications. The results are being applied in real-world technological processes (see some hints in the last section of this paper).

All the components – e.g. sensors, communication devices, processing devices, reasoning subsystems, effectors (actuators) – used for building cognitive artificial systems can be considered as agents and often form nodes of MANET. Roughly speaking, large part of the system's dynamic functionality (e.g. defining, detecting, and interpreting situations) is defined by the interactions between the agents; the stationary part of functionality (e.g. acquiring or calculating values of individual variables, reasoning about situations, etc.) is defined by in-agent algorithms. By analysing and moderating the information flow between agents one can control, validate, and verify on-line the system's behaviour.

The information exchange pattern of agents is rather complex. First, the amount of situational information required in a complex system of practical size involves several thousands of agents, is distributed on a large area, and information is distributed selectively – each agent may subscribe to specific information, and each agent is assigned specific personal access rights partly depending on situation-dependent time and space constraints. Second, the configuration and topology of interaction channels in a multi-agent system is not stationary and may change at runtime due to self-X properties of the system. This means that producers and consumers of situational information are established dynamically at runtime. Each agent that subscribes to data can also set specific requirements to the validity of received data.

The Lab is developing a content-centric type of proactive middleware (see Preden et al. 2011) that supports subscriptions to data, dynamic validation of that data, checks the access rights of the subscriber, and contributes to dynamic change of the topology of interaction channels – the proactive middleware maintains and controls mediated interactions. The mediated interaction system and its control methods have been elaborated from the ideas described in (Motus 1995).

The agents exchange messages that have a uniquely identified and verified content and do not worry about where or how the message content is obtained. The content of a message is constructed based on the paradigm of a digital map, called 'interactive map' – the information is stored in layers, separate layers are for stationary information and for dynamically changing information (e.g. recent changes, position of a mobile agent). In some (e.g. military) applications the interactive map can be built on top of a regular digital map. This provides a user friendly interface that can be processed equally well by humans and machines, and facilitates transfer of a longer message by layers, depending on the available bandwidth and available processing power of the consumer (see for details Motus et al. 2008).

Organisation of the research in cognitive artificial systems domain in Estonia

Cognitive artificial systems and networked pervasive proactive systems, as a second generation of embedded systems, is clearly the mainstream of computer applications in 21st century – already in 2010 more than 96% of all processors operating in the world were embedded into natural and artificial systems and operated without direct intervention of humans. This new situation has invoked revision of the theoretical foundations in several research disciplines, e.g. computer science, automatic control theory, signal processing, robotics, manufacturing, and indirectly also in cognitive sciences. As a by-product, a new time counting system has been introduced into some of the theories – in addition to continuous (tense) time (modelled by real numbers) and discrete time (modelled by integers), a random subset of integers is used to model time as counted in computer systems.

High level decision-makers in Estonia are cautious and rather distrustful towards developing technology based solutions. Still, today Estonia has quietly fading image of a country that has been pretty successful in e-governance, has some good usages of cross-linked databases to support management, and has had some other innovative computing-related ideas. It seems likely that the quietly fading image of e-Estonia can be improved if cognitive artificial systems and networked proactive computer systems would be taken into use.

The reasonable body of researches (10 PhD level researchers, 7 researchers working on their PhD thesis, and a background team) involved in research of cognitive artificial systems is today concentrated at the Research Laboratory for Proactive Technologies (Department of Computer Control, Tallinn University of Technology), and there are several other research groups in Estonia involved in this domain. For instance, the Lab has had long term collaboration with Institute of Technology (University of Tartu), Lab of Socio-Technical Systems (Faculty of Information Technology, Tallinn University of Technology), with IB Krates OÜ, and Defendec OÜ. The Lab is a partner in Estonian Competence Centre in Innovative Manufacturing Engineering Systems (IMECC) – this is our source of industrial expertise in applying software-intensive cognitive methods.

The international research cooperation of the Lab is carried out mainly through NATO RTO task groups, European Defence Agency's projects, ARTEMIS Joint Undertaking projects, ITEA (Information Technology for European Advancement) projects, also collaboration with other European research managing bodies, and direct contacts with several universities are important.

Five Estonian research groups from two universities, looking forward to prosperous research future in Europe, have signed a Memorandum of Understanding on future collaboration in cognitive artificial systems, provided that there will be political or economic commission to applying cognitive artificial systems supported by adequate finances. The undersigned research groups from University of Tartu are: Institute of Technology, Development Centre for Information Technology; Institute of Physics, Department of Materials Science, Laboratory for Thin Film Technology; and Faculty of Science and Technology, WG on Space and Military Technology. The undersigned research groups from Tallinn University of Technology are: Institute of Informatics, Laboratory for Socio-Technical Systems; and Department of Computer Control, Laboratory for Proactive Technology.

REFERENCES

Cohen, R., Havlin, S. 2010. Complex Networks: Structure, Robustness and Function. Cambridge University Press.

Dosch, W., Meriste, M., Motus, L. 2007. Enriching interactive components with again commands. Proceedings IEEE International Conference on Electro/Information Technology, IEEE Computer Society, Los Alamitos, 200-205.

EASAC 2010 Realising European potential in synthetic biology: scientific opportunities and good governance. European Academies Science Advisory Council (EASAC) policy report no. 13, www.easac.eu

Hall C. M., McMullen, S. A. H., Hall, D. L. 2006. Cognitive engineering research methodology: A proposed study of visualisation analysis techniques. NATO RTO-MP-IST-063, 1-10, http://www.rto.nato.int

Motus, L. 1995. Timing problems and their handling at system integration. Tsafestas, S. G., Verbruggen, H. B. (eds.) Artificial Intelligence in Industrial Decision Making and Automation. Kluwer, 67-87.

Motus, L. 2010. Chapter 5, The impact of system integration on design, analysis and verification. Ramage, J., et al. (eds.) Design Considerations and Technologies for Air Defence Systems, 1-26. (NATO RTO Publications; SCI-181).

Motus, L., Meriste, M., Dosch, W. 2005. Time-awareness and Proactivity in Models of Interactive Computation. Electronic Notes in Theoretical Computer Science, 141, 69-95.

Motus, L., Meriste, M., Preden, J.-S. 2006. Network Enabled Capabilities – grass-roots perspectives. IST Symposium on Dynamic Communication Management, 1-13. (NATO RTO).

Motus, L., Meriste, M., Preden, J.-S. 2008. NNEC technologies focused on use by semi-autonomous groups. Proceedings SCI-187 Symposium on Agility, Resilience and Control in NEC, 1-16. (NATO RTO).

Preden, J.-S. 2010. Enhancing Situation-awareness, Cognition and Reasoning of *ad hoc* Network Agents. Tallinn University of Technology.

Preden, J. et al. 2011. Situation awareness for networked systems. Proceedings IEEE 1st International Multi-disciplinary Conference on Cognitive Methods in Situational Awareness and Decision Support, 123–130.

ProLab report 2009 = Research Laboratory for Proactive Technologies in 2009. Tallinn University of Technology.

RAND 2005. Gonzales, D. et al. Network-Centric Operations Case Study: The Stryker Brigade Combat Team. National Defense Research Institute.

Rosenblueth, A., Wiener, N., Bigelow, J. 1943. Behavior, purpose and teleology. Philosophy of Science, 10, 1, 18-24.

Vernon, D., Metta, G., Sandini, G. 2007. A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. IEEE Transactions on Evolutionary Computation, 11, 2, 151-180.

RESEARCH ON DIGITAL SYSTEM DESIGN AND TEST AT TALLINN UNIVERSITY OF TECHNOLOGY

R. Ubar, P. Ellervee, T. Hollstein, G. Jervan, A. Jutman, M. Kruus, J. Raik

INTRODUCTION

We are entering the era of pervasive computing. PCs have been dethroned by technology to embed computers in almost everything. 98% of computers in the world are embedded. Massive amounts of electronics and software are controlling our everyday life. Pacemakers, cell phones, home appliances, energy generation and distribution, antilock brakes in cars – this is what we are depending on at every moment. We do not even notice this dependence, except when something is going wrong because of the computer crash. Embedded systems (ES) create huge value for society, but also unprecedented risks because of possible errors and faults. We need dependability from the embedded computing world around us.

Silicon technology is continuously decreasing in size. According to the 2010 International Technology Roadmap for Semiconductors (ITRC), the industry should have the 22-nm technology by 2016. As critical dimensions reach the nanometre range, the devices will be subject to an increasing number of defects and intermittent faults during operation. Engineers are facing a paradox: they have to build reliable systems from unreliable components. From the development costs of systems, because of their complexity up to 70% is going only for design verification and removing errors.

To cope with design verification and reliability problems, design for test and test synthesis methods are constantly evolving. New fault models have to be developed to handle new emerging defect mechanisms and hierarchical approaches to synthesise automatically test programs is the way to deal with increasing complexity.

The research on digital hardware (HW) design and test in Estonia is concentrated mainly in the Department of Computer Engineering (DCE) at the Tallinn University of Technology (TUT), and is carried out in close cooperation with Estonian electronics industry especially with the companies like Ericsson, Elcoteq, National Semiconductor, Testonica Lab, and with R&D centres CEBE and ELIKO. In the following, an overview is given about the recent results in the field of design and test achieved in DCE and about the visions for future R&D in the field.

Our research has been concentrated on developing of new methods in design and test of digital HW regarding the topics in Fig. 1.

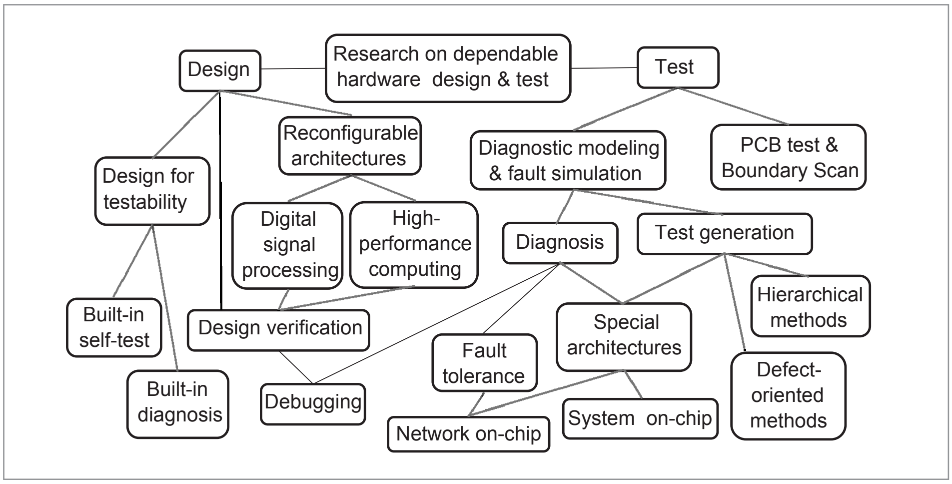


Figure1.
Research on dependable hardware design and test at TUT.

Diagnostic modelling of systems

Traditionally, the theory of design and test of Digital Systems (DS) has been developed on the abstraction level of logic gate networks. To cope with the increasing complexity of systems, new high-level and hierarchical methods for diagnostic modelling of systems are emerging. On the other hand, shrinking geometries in nano-scale technologies produce new failure mechanisms which have forced to search for more advanced fault models. We have been looking for a uniform way of modelling systems on different levels of abstraction. A solution has been found in developing a novel mixed-level graph-model in form of Decision Diagrams (DD).

Low-level modelling

Binary Decision Diagrams (BDD) have become state-of-the-art data structure in VLSI CAD. BDDs were introduced in 1959 (Lee 1959) for representing Boolean functions. For diagnostic modelling of digital circuits, a special type of circuit structure related BDDs were proposed at TUT in 1976 (Ubar 1976). The first test generator in the world, based on BDDs, was built in a cooperation of TUT and Institute of Cybernetics in 1982.

After Bryant showed in 1986 the simplicity of graph manipulations and proved the model canonicity (Bryant 1986), the BDDs started quickly to spread. However, traditional use of BDDs has been still functional, i.e. the target has been to represent the functionality, and not the structure behind the functions as in Structurally Synthesised BDDs (SSBDD) developed in (Ubar 1976, 1996, 1998).

SSBDDs provide mapping between the graph and the related circuit, which allows to model different test relations and properties of circuits like signal paths, delays, faults, the properties of faults like masking, equivalence, dominance, redundancy, etc. These issues cannot be simulated explicitly with 'classical' BDDs. The motivation to introduce SSBDDs was to improve the efficiency of test methods by reducing the complexity of the model.

High-level modelling

To overcome the difficulties of high-level diagnostic reasoning of complex DS, we have introduced High-Level DDs [HLDD] as an extension of SSBDDs (Ubar 1996; Karputkin et al. 2010). The model covers different levels of abstraction, and supports uniform fault analysis and diagnostic reasoning principles.

The most important impact of the HLDDs is the possibility of generalisation of the logic level methods for using on higher levels by the same formalism. The class of variables was extended from Boolean to integer variables, whereas the class of Boolean functions was extended to data manipulation operations typically used in high-level descriptions of DS.

We have developed the canonical presentation of HLDDs in the form of characteristic polynomials (Karputkin et al. 2010). The synthesis of HLDDs can be carried out either by symbolic execution of procedures, or by iterative superposition of HLDDs, similarly to the synthesis of SSBDDs (Ubar et al. 2011).
Both, SSBDDs and HLDDs have been the basis of a lot of research carried out in DCE to develop efficient methods and algorithms for automated fault diagnosis in digital systems.

Fault modelling and simulation

The DD-based fault modelling allowed us to improve the existing fault collapsing methods and to create a uniform fault formalism for different levels of abstraction (Ubar et al. 2010). It covers a broad class of faults like resistive shorts, opens and bridges (Raik et al. 2005). To deal with the non-determinism of faults, the general Byzantine fault and *X*-fault models have been implemented with great success using SSBDDs for fault simulation purposes (Ubar et al. 2010).

Fault simulation is the basic task in testing, and the fault simulators are used nearly in all CAD tools related to test. We invented a parallel critical path tracing algorithm based on SSBDDs which allowed to speed-up simulation several times compared to the available commercial tools (Ubar et al. 2011). The main innovation was in the parallelisation of analytical fault reasoning for many test patterns simultaneously. This idea was extended for general fault classes like conditional SAF (Raik et al. 2005) and X-faults (Ubar et al. 2010).

Test and verification

A general formulation of the problem

Research in digital test is carried out in two main directions: test synthesis and fault diagnosis. Both problems can be regarded as reverse mathematical tasks. Consider a digital system in a very general form as a vector function $Y=F(X)$, where Y and X are the function and argument vector variables, respectively. A test experiment can be described as a diagnostic equation $F^*(X,dX) = dY$, where $F^*(X,dX)$ is a full differential of the system function, describing jointly the correct and faulty behaviours, where the vector differential dX describes the possible faults, and dY describes the result of test experiment in the form of possible deviations from the correct behaviour of the system. In the Boolean space the equation $F^*(X,dX) = dY$ takes the form of Boolean differential equation.

Without going into details, we can formulate now the two main test related tasks. The first is the direct task – to find dX if X and dY are given, i.e. to locate the fault by the results of the given test experiment, and the second is the reverse task – to find X if dX and dY are given, i.e. to generate test for the selected faulty behaviour.

Representing the reverse and direct tasks as the solutions of the same differential equation helps to have a uniform view on a lot of problems in testing: fault modelling, simulation and masking, test synthesis and analysis, fault diagnosis. For example, the fault simulation can be regarded as a special case of the direct task of fault location. The main problem is how to process the sequential behaviour of systems and cope with the complexity issues. In our research we have used DDs for solving the direct and reverse tasks of the diagnostic equation on different levels of abstraction.

Test generation

The test generation efficiency is sensitive to the complexity of the diagnostic model. On the other hand, the quality of tests is highly dependable on the accuracy of fault modelling. We have been targeting both, the efficiency by developing hierarchical approaches (Raik, Ubar 2000), and accuracy, by using conditional SAF (Raik et al. 2005).

The cornerstone of the defect-oriented test generation method (Raik et al. 2005) was the mapping of faults from one hierarchical level to another. Each library component is represented by a set of constraints calculated on the exact layout level. On the logic level, the physical defects are modelled by the constraints i.e. by conditional SAFs.

Figure 2.
Debugging errors in a digital design.

In (Raik, Ubar 2000) we developed a novel hierarchical approach to test generation for systems with mixed-level DDs. The method handles data and control parts in a uniform way. Experiments showed higher speed of test generation and higher fault coverage compared to the known academic tools. Commercial tools are missing on the market. Based on the described method, a very efficient test generator DECIDER was implemented, and used for industrial designs at the Fraunhofer Institute Dresden in Germany.

New advanced methods for testing other classes of physical defects like shorts (Raik et al. 2005), crosstalks (Bengtsson et al. 2008), and high-level functional faults (Raik et al. 2008) have been developed at DCE recently.

Verification and design error diagnosis

SSBDDs and HLDDs have been used for improving the methods of design verification and error diagnosis at different system abstraction levels (Jenihhin et al. 2009; Ubar 2003). A novel approach for assertion coverage analysis targeted the quality assessment of verification stimuli and design error debug (Jenihhin et al. 2009). The approach considers HLDD-based design verification flow and relies on the extended HLDDs to support modelling of temporal properties.

Most research in design error and fault diagnosis has been done using specific fault and error models. But, the variety of design errors and hardware faults is practically infinite. A new method for debugging of errors, which does not use any error model, was proposed in (Ubar 2003). Based on the diagnostic reasoning of the circuit, a subcircuit suspected as erroneous is extracted. Opposite to known works, re-synthesis of the subcircuit need not be applied to the whole function of the erroneous internal signal in terms of primary inputs; it may stop at arbitrary nodes inside the circuit. As the result, the speed of repair will be significantly increased. The model free design error diagnosis method was extended also for using at higher levels of abstraction (Ubar et al. 2011). Since our method executes in polynomial time, the larger designs will be more efficiently handled by the proposed method than with the existing formal approaches.

Test methods

A new concept and a method for test and diagnosis was developed for NoC designs (Raik et al. 2009). The method is based on functional fault models and it implements packet address driven test configurations. The approach is well scalable. The dedicated DfT techniques were proposed for application of test patterns from external boundaries of NoC. The experiments showed near-100% test coverage at the expense of less than 4% of extra switch area.

Electronic systems are usually based on SoC such as microcontrollers or signal processors that communicate with many peripheral devices on the system board and beyond. We pointed out particular challenges in testing of the printed circuit boards (PCB) and proposed a general modelling methodology for test automation for microprocessor SoC based system boards (Devadze et al. 2009). The method is a considerable step forward in test automation, since today the test routines for PCBs are still programmed manually. We developed a new Boundary Scan (BS) based test access protocol for system-level testing of boards for manufacturing defects. The new technique dramatically extends the applicability of BS testing in complex on-board data buses and protocols. The side-effect of the technique is the possibility to increase the speed of in-system programming of flash memories where a speed-up of 50 times was achieved.

Hardware simulation

Meeting timing requirements is an important constraint imposed on highly integrated circuits (IC), and the verification of timing is one of the most critical tasks for CAD tools. The first time, a novel algorithm for multi-valued simulation based on Boolean differential algebra for detecting hazards was implemented with SSBDDs in (Ubar 1998). New algorithms were developed using SSBDDs, which allowed to speed-up timing simulation up to three times compared to gate level simulation (Jutman et al. 2001).

Built-in self-test

New design paradigms like SoC and NoC have made external testing of systems increasingly difficult. The speed of SoC is constantly increasing and the technology used in external testers is always one step behind. Therefore, Built-In Self-Test (BIST) is emerging as a promising solution to the VLSI and SoC testing. We have developed different approaches to improve the efficiency and quality of BIST architectures. For the hybrid BIST which combines pseudorandom and deterministic test, a method and algorithms were developed for fast calculation of the cost of BIST to speed-up the solutions exploration (Ubar et al. 2005).

A new approach to increase the speed of self diagnosis was developed, in which instead of the commonly used bisection of test patterns, the idea of bisectioning of faults was proposed (Ubar et al. 2008). To improve the diagnostic resolution, a novel approach of using multiple signature analyzers (SA) was developed. The key problem was to find an optimal interface between the circuit and SAs. An algorithm was created to find the optimal number of SAs for achieving the best diagnostic resolution.

Hardware based acceleration of fault simulation

Traditional SW-based simulators do not provide sufficient speed in case of complex systems. Especially time consuming is the BIST quality analysis in case of sequential circuits where test sequences are long and simulation speed is low. To achieve higher productivity in fault simulation, we developed a novel HW accelerator using reconfigurable logic on FPGAs (Ellervee et al. 2007). The proposed approach allowed speed-up of 40-500 times as compared to the SW-based fault simulation.

Research environment

A new modern embedded system research environment was recently established as the result of the project SARS, EU23626. Since 1995 our department has been a member of EUROPRACTICE – a European Commission initiative which supports the cooperation between the selected universities and the electronic industry in Europe. As the result of this membership, CAD software from the most prominent EDA vendors, such as Cadence, Synopsys, Mentor Graphics, Xilinx has been installed in DCE. The environment includes as well the tools that have been developed in DCE as a side effect of research:

- TURBO-TESTER (TT) – for logic-level test (Novak et al. 2005; http://www.pld.ttu.ee/tt/);
- DECIDER – for hierarchical test generation (Raik, Ubar 2000);
- DOT – for defect oriented test (Raik et al. 2005);
- xTractor – CAD software for high-level synthesis;
- APRICOT – verification framework for high-level designs (Jenihhin et al. 2009);
- DEFSIM – for research of the physical defects in integrated circuits (http://defsim.pld.ttu.ee:8080/defsim/);
- BIST Analyzer – a tool for research and analysis of the quality of BIST (http://www.pld.ttu.ee/applets/bista/);
- Trainer 1149 – for investigating the basic concepts of the Boundary Scan standard (http://www.testonica.com/news/2009/04/15/trainer-1149-11-version-released);
- Java applets – for e-learning of logic and RT-level test (http://www.pld.ttu.ee/applets/).

TT has gained a large popularity all over the world; it has been licensed from more than 100 institutions in about 40 countries. The tools of TT allow solving different test related tasks by alternative methods and algorithms in a wide scope of scenarios:

- test pattern generation by deterministic, random and genetic algorithms;
- test optimisation and compaction;
- fault simulation by different algorithms;
- multi-valued simulation for analysing dynamic behaviour of circuits.

The tools of TT are based on SSBDDs, there are converters to link TT with commercial CAD systems. The tools support gate- and macro-level modelling. The latter helps to reduce the complexity of the model and to improve the performance. DECIDER uses two inputs – RT level descriptions in VHDL and low gate-level descriptions for components in EDIF. DOT is an extension of the TT for defect-oriented test. DEFSIM allows carrying out the low-level research of real physical defects to map them on the logic level. It is an HW/SW environment for experimental study of CMOS physical defects. APRICOT is an extension of DECIDER, to connect two communities – the hardware test community with the design verification counterpart. Trainer 1149 is a multi-functional SW system, which provides a simulation, demonstration, and CAD environment for learning, research, and development related to IEEE 1149.1 BS standard. BISTA and Java applets are a selection of tools that have been developed specially for teaching purposes, and are integrated into e-learning environment to support university hands-on training.

The originality of the environment is in its multi-functionality (important for research and training), low-cost and ease of use (Fig. 3). The multi-

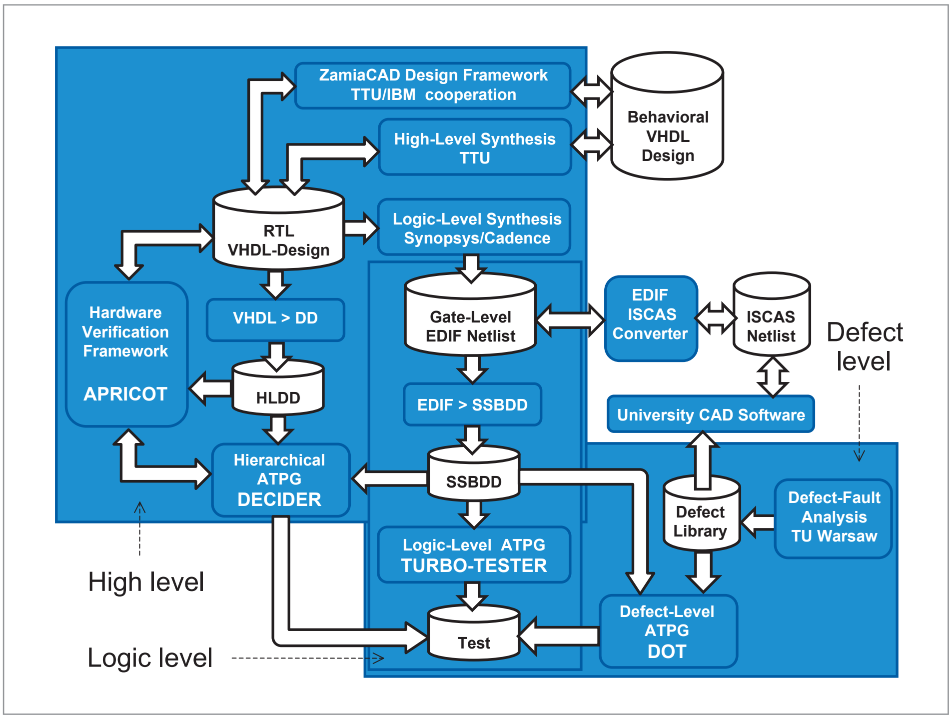


Figure 3.
CAD tool environment for research in HW design and test.

functionality means that different abstraction level models can be easily synthesised (to analyse the impact of the complexity of the model on efficiency of methods).

Alternative methods for the same task are implemented to compare different algorithms, and the fault models can be easily exchanged and updated to analyse the accuracy of testing. The multi-functionality allows setting up and modifying easily different experimental scenarios for investigating new ideas and methods. In traditional commercial design tools these purely research and teaching oriented possibilities are missing.

Cooperation in the excellence centre CEBE

In 2008, a Centre of Integrated Electronic Systems and Biomedical Engineering (CEBE) was established at TUT. It is one of the seven Estonian centres of research excellence, supported by EU structural funds, and it joins the research teams of DCE, Department of Electronics (DE), and Technomedicum (TM). The centre is carrying out interdisciplinary R&D in the fields of electronics, computer and biomedical engineering with applications in medicine, semiconductor and information technologies.

The research at DCE is greatly coordinated by the goals of CEBE, especially regarding applications. The partners of CEBE are also the founding members of the Competence Centre in Electronics, Info- and Communication Technologies ELIKO, which was established with the goal to develop innovative technologies and products, based on intelligent embedded systems, through strategic co-operation between the science and industry sectors.

The research cooperation in CEBE is organised in the form of 7 internal projects targeting joint applications involving the partners' competences. DE provides knowhow in signal processing whereas TM involves experts from medical and bioengineering fields. A remarkable synergy has been achieved already in CEBE regarding joint design and test of biosignal processors for medical applications (Min et al. 2010; Pilt et al. 2010). For example, based on the method invented in TM for detecting depressive disorders by measuring bioelectromagnetic signals of the brain (Hinrikus et al. 2009), a mobile EEG analyzer is being now implemented in CEBE (Fig. 4). Novel optical methods for estimating dialysis adequacy are being developed at TM (Lauri et al. 2010). Surprising synergy has been shown in CEBE by implementing the ideas and models of diagnosis from the field of electronics developed in DCE for analysing the quality of dialysis in the medical field. CEBE is well cooperating with industry and hospitals.

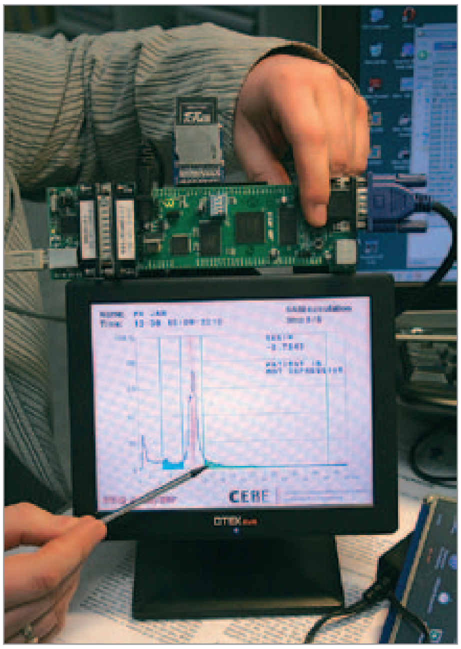


Figure 4.
EEG analyzer detects depression.

International cooperation

The research team of DCE is characterised by active international co-operation. It has been involved in 14 EU-level projects within the COPERNICUS, COST, ESPRIT, FP5, FP6 and FP7 framework. Currently, two FP7 projects DIAMOND and CREDES are running, in both of them DCE is the coordinator.

DIAMOND (Diagnosis, Error Modelling and Correction for Reliable Systems Design) aims at improving the productivity of electronic system design in Europe by providing a systematic methodology and an integrated environment for the diagnosis and correction of errors. The consortium includes top-level research groups from universities of Tallinn, Linköping, Bremen, Graz, and the industry partners IBM, a leading microprocessor manufacturer, and Ericsson, a leader in the telecom field. Supported by CAD vendors TransEDA Systems and Testonica Lab, the team covers a full spectrum of the manufacturing chain from the tool development to the systems design and technology.

CREDES aims at creating a Centre of Research Excellence in Dependable Embedded Systems, based on the research potential of DCE, Dept. of Computer Science (DCS), and infrastructure of the Laboratory ASSA at TUT. The ambition is to become one of Europe's leading institutions responsible for R&D in the areas of design, verification, test and diagnosis of ES. The

Centre will be created by developing TUT's scientific expertise and capacities in collaboration with research groups at Universities of Verona and York, TU Darmstadt, and TU Brandenburg in Cottbus.

Cooperation with Testonica Lab and Göpel Electronic GmbH (Germany) has already resulted in two products "Microprocessor models for testing" and "Embedded instrumentation IPs for testing of electronic boards and chips" which are marketed worldwide.

FUTURE RESEARCH DIRECTIONS

Our future research will target new breakthroughs in test, verification and dependability of embedded systems by developing new methods and CAD tools with design applications in health care. The research topics we are targeting belong to the forefront of scientific research and correspond to the priority research areas of roadmaps like ARTEMIS (http://www.artemis-sra.eu/; https://www.artemisia-association.org/), ITRS (http://www.itrs.net/), and HiPEAC (http://www.HiPEAC.net/roadmap).

Research perspectives for dependable embedded system design

Historical view on the vertical disintegration in IC Industry

Electronics industry has gone through major changes. Fig. 5 depicts the developments in vertical disintegration of the semiconductor and electronic equipment design industry with predictions until 2020. In the beginning, microchips were completely designed and manufactured by integrated electronics companies. Equipment manufacturers were assembling chips to systems, and equipping them with the firmware and SW based on own developments.

With increasing system complexity, the process of disintegration began on the ES market: semiconductor companies were offering their technologies to design companies by providing design libraries. Fabless semiconductor companies are offering own silicon products like FPGA. Currently, on the HW and SW side, many companies offer intellectual property (IP) components, which are used by system designers and system integrators to develop products efficiently in short development cycles. Such disintegration will continue and in 2020 we expect the situation, that fabless companies will provide large scale reconfigurable platforms, consisting of fixed and reprogrammable HW as well as embedded processors (SW platforms). The platforms will enable opportunities for other companies to develop different applications.

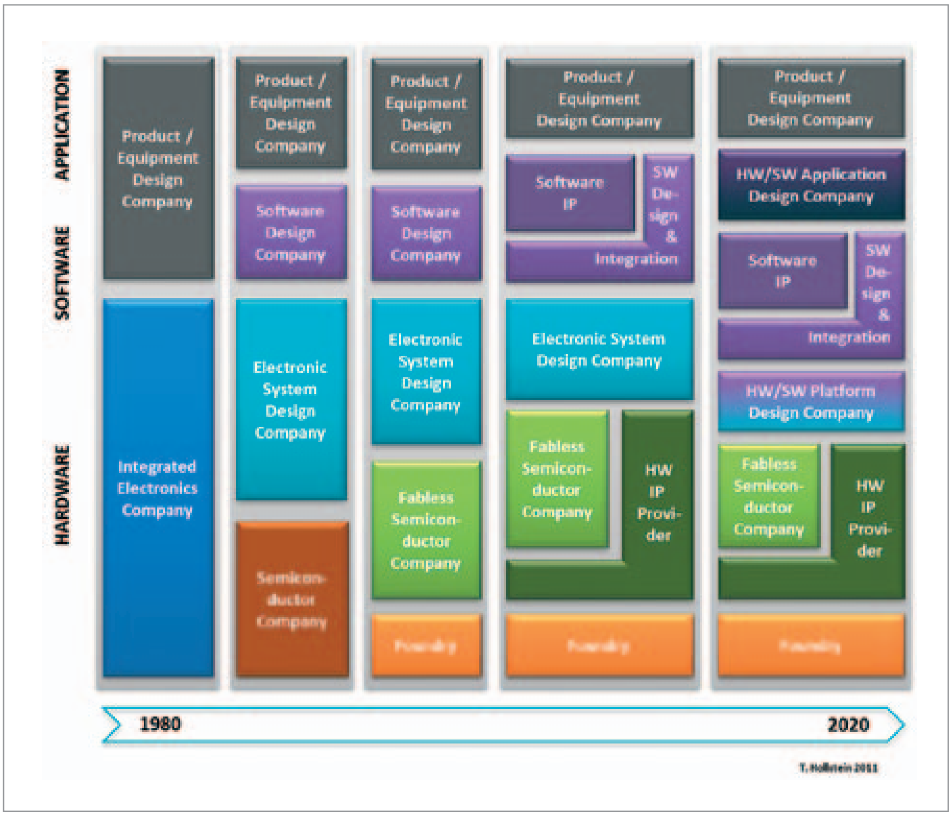


Figure 5.
Vertical disintegration in semiconductor and electronics equipment industry.

The described development implies several economical and technical consequences:

- increasing system complexity requires new methods for integrated HW/SW design, which especially would address system reliability and dependability;
- system test strategies are required on all levels of design and integration, and powerful debugging strategies are crucial.

Consequences for industrial development and education

The vertical disintegration in the IC and IT industry generates a huge market for SMEs (especially important for small countries like Estonia), to offer own solutions in different stages of the IT product value chain. Fabless concepts allow successful business ideas for companies to offer HW IP components. Complete HW/SW platforms can be offered as fabless designs for customers targeting solutions in application areas. Highly sophisticated products can be developed based on available reconfigurable HW/SW platforms in cooperation with other companies.

In order to fulfil the requirements of future markets, the strategic education in universities and other educational institutions needs an early orientation towards the expected needs with respect to technical and professional skills. A key issue is that while in the early stages of disintegration specific detailed know-how (HW or SW) was sufficient, future product development needs an increasing amount of system engineering. Highly skilled system designers who understand the components and methods coherences and who are capable to come up with system design decisions which match the constraints and requirements after the implementation and integration of all components, are needed.

Consequences for research in dependable embedded system design

The increased complexity of systems generates a demand for sustainable integrated HW/SW design methods with a strong emphasis on test, verification and dependability. Nano-scaled technologies with increased physical parameter variations have to be encountered with dependable circuit design methods in order to guarantee a sufficient yield and reliability of systems. Dependability concepts cannot be limited to IC design, but have to be deployed on different levels of abstraction. Large-scaled future embedded platforms will comprise a large number of processor cores, specific HW components and reconfigurable HW. These numerous system components will communicate via flexible and scalable on-chip communication architectures such as NoC.

In our future research on dependable ES, we will focus on the dependability of HW components and adaptive NoC architectures, which allow bypassing faulty communication links and components, and novel application mapping methods on NoC-based multiprocessor platforms, considering reliability aspects.

Design research on dependable Networks-on-Chips

In the NoC-based systems the communication is achieved by routing packets through the network infrastructure, rather than routing global wires. However, communication parameters (inter-task communication volume, link latency and bandwidth, buffer size) might have major impact on the performance of applications implemented on NoCs. To guarantee predictable behaviour and to satisfy performance constraints of real-time systems, careful selection of application partitioning, mapping and synthesis algorithms is required. In the present, no methods or tools for solving these tasks for NoC-based real-time embedded systems exist.

Our approach will operate on a resource minimised NoC without virtual channels which are extremely area consuming. We assume to have NoC architecture with best-effort service without packet interleaving. A new concept for communication synthesis will be developed, which would not be run off-chip as a CAD tool on a workstation, but on-chip and being activated whenever the multi-core system is re-configured or new applications are added dynamically. The emphasis is on simple and efficient modelling and communication synthesis for dynamically reconfigurable systems, and the main idea is to calculate exact communication deadlines that describe how the communication synthesis can guide the scheduling process such that network conflicts can be avoided.

As technologies advance, a high degree of sensitivity to defects begins to impact overall yield and quality. The ITRS (http://www.itrs.net/) states that relaxing the requirement of 100% correctness for devices and interconnects may dramatically reduce costs of manufacturing, verification, and test. Such a paradigm shift is likely forced by technology scaling that leads to more transient and permanent failures of signals, logic values, devices, and interconnects. It means that even in consumer electronics, for example, where the reliability has not been a major concern so far, the design process has to be changed. Otherwise, there will be a high loss in terms of faulty devices due to problems stemming from the nanometre manufacturing processes.

Due to the future design complexities and technology scaling it is infeasible to concentrate only to low level reliability analysis and improvement in system design. We have to target application level reliability analysis and improvement, i.e. we have to assume that the manufactured devices might contain faults and the application, running on the system, must be aware that the underlying hardware is not perfect. Moreover, design methods for dependable systems will have to cope with not only 'failed' devices right after the production, but also with further failures in the field.

Hence, our future research will consider the topics of BIST, built-in self repair and recovery mechanisms that work in the field of application. In present, there is no complete system-level design flow taking into account the NoC modelling and system-wide dependability issues and therefore our main research effort will focus on various modelling and optimisation methods that can be used for developing dependable, timing sensitive multi-core systems.

Application oriented design using reconfigurable logic

FPGA technologies have become a great potential to assist general-purpose processors in performance-critical tasks which is often the case in ES. There are a few reasons to use FPGAs instead of general purpose processors or graphical processors. The first one is the flexibility of the reconfigurable fabric allowing implementing heterogeneous parallel computing units inside the same chip. Another reason, which is getting more and more important nowadays, is the energy efficiency – FPGAs consume less power per operation than processors.

At DCE under coordination of CEBE, reconfigurable devices are planned to be used to accelerate calculations in different bio-medical measurement devices.

Measurement of electrical bioimpedance enables to characterise a state of tissues/organs, to get diagnostic images, to find hemodynamical parameters, etc. The main problem with bio-impedance measurements is the fact that the useful information is hidden under the background noise generated by the normal body activity. An example would be respiration generated noise when measuring heart activities. The impedance of tissues and organs is measured between the electrodes having different locations. The response voltage is digitised into a uniformly sampled sequence of data that is processed in a Digital Signal Processor (DSP). Synchronous sampling can be extended to the multi-frequency measurement with the help of time-multiplexed sampling. A problem to be solved in sampling is to avoid aliasing effects. Promising results have been achieved already by implementing a pre-processing unit for eight-channel bioimpedance measurement device with 80 MHz sampling frequency using Xilinx Spartan3 FPGA.

Other design research plans are related to applications in the human brain research field. The analysis of human brain activity bio-signals requires complicated calculations and their interpretation remains to be a challenge for science. Everyday stress has made depression and other mental disorders more frequent. The method for detection of depressive disorder is based on analysis of the electroencephalographic (EEG) frequency spectrum and is capable of determining depressive disorders. A spectral asymmetry index (SASI), developed in CEBE (Hinrikus et al. 2009) is used as a depression indicator, and can be calculated as a relative difference in power of two EEG special frequency bands selected higher and lower of the EEG spectrum maximum. The main challenge will be to develop a portable device which would offer real-time analysis and acceleration of the calculation procedure.

Similar devices based on DSP modules and implemented as reconfigurable architectures can be easily adjusted on-the-fly for solving other tasks in health monitoring.

The results of the research described will be used in future work to define a modular signal processing platform (Fig. 6) and its components for reconfigurable logic based high-performance computing devices. The platform will consist of communication network, sensors/actuators, and signal processing units. While the sensors and actuators can be seen as more-or-less standard devices, the other parts must be configurable to allow making trade-offs depending on the design constraints. The constraints can be performance, energy consumption, data format, etc. Some applications may need point-to-point fast cross-bar networks while other applications can work with slower but cheaper bus-like networks. The same applies to the processing units where a good example is Fast Fourier Transform that can be implemented in very many configurations depending on the performance and window size requirements, for instance. The research will focus on analysing the requirements of different signal processing applications to identify the needs for parameterisation and developing parameterisable network and processing modules.

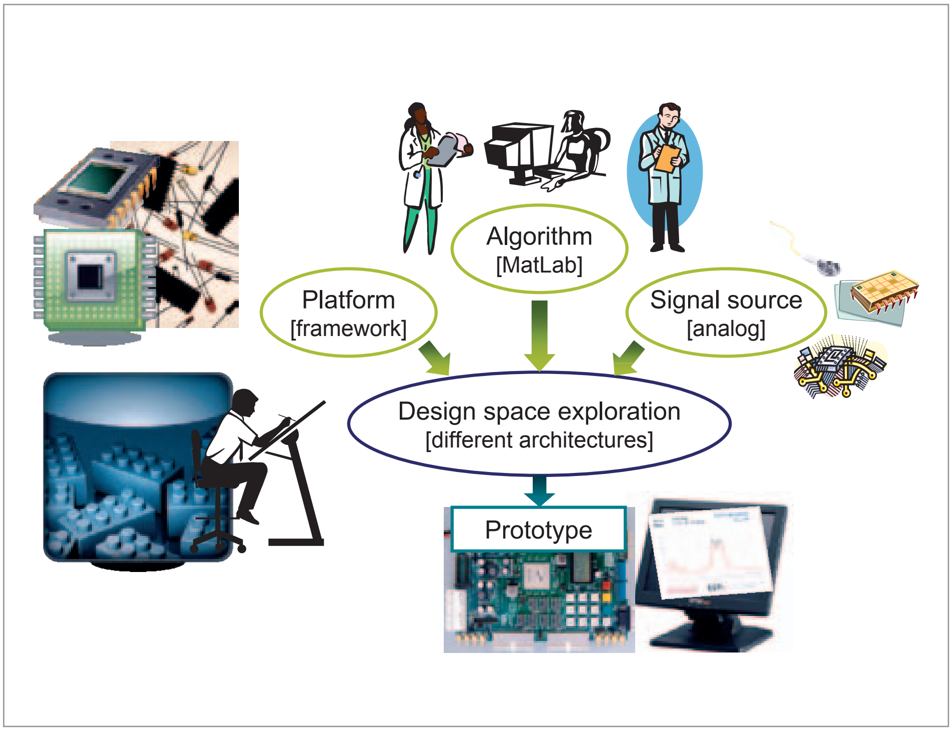


Figure 6.
Design platform for modular signal processing and reconfigurable logic based high-performance computing devices.

Future research challenges in system test and verification

The cost of designing nano-scale chips is rapidly increasing. Together with the growth of design complexity, the design methods as well as test and verification methods have to shift to higher abstraction levels. In particular, transaction-level modelling at Electronic System Level is gaining ground in modern design process.

There is a need for new verification methods that would operate on higher abstraction levels and provide readable and compact information for the design engineer. The feedback from existing verification solutions is far too detailed and contains too much data. This makes understanding the root causes of errors and debugging in complex systems an overwhelming task. Similarly, the test methods have to move to higher levels. At the same time there are challenges emerging from the lowest abstraction levels because of new effects in nanometre technologies like process variations, power consumption, or vulnerability to crosstalk effects. More emphasis should be put on functional and at-speed test as they tend to cover more defects than the conventional methods. New failure mechanisms are essentially dynamic, not static, and delay faults are gaining importance. The power dissipation has become a serious issue and power-aware testing methods are becoming inevitable.

According to Rent's Rule the number of input/output terminals of chips grows as a square root with respect to the number of transistors. This implies that the test access is going to be more and more difficult. BIST is required to provide test access to modern systems and to test them at-speed.

Soft errors caused by radiation effects are becoming a severe issue in nanometre processes. This means that it will not be possible to create working miniaturised electronic circuits anymore without implementing fault tolerant techniques.

Finally, adoption of 3D stacking process in chip manufacturing causes a paradigm shift in test. Test access will be a problem. When you are doing a wafer test you cannot really access the contacts without proper pads. It is not possible to probe the Through-Silicon Vias (TSVs) due to the small size and low fracture strength of micro-bump contacts. The community is forced to apply scan-path, BIST, test compression and other DfT solutions. Therefore, the re-use of DfT on different dies becomes essential. Serious problems remain with signal integrity and power integrity.

The considerations above will determine our future research in test, verification and fault diagnosis. To cope with the complexity of problems, we will use hierarchical approaches. To handle the problems of diagnosis at different levels of abstraction, we will use the theory of multi-level decision diagrams which, however, needs extensions and further developments. Fault simulation is the basic tool in all test related tasks. We have achieved good results in logic level fault simulation, however these results must be extended to support higher level verification and debugging to cope with the complexity.

Challenges in testing of complex electronic boards

The electronic industry is entering a new age, where static structural test technologies like Boundary Scan (BS) and In-Circuit Test are quickly losing their efficiency in terms of fault coverage, while no systematic alternatives to replace them currently exist. This brings up a lot of concerns in the industry and calls for due attention from academic researchers.

Various solutions exist for manufacturing defect testing on complex PCBs, but all of them have certain limitations (low fault coverage, low speed, and expensive equipment). As a result, numerous production test types/test phases have to be used, which turns testing into a time-consuming and expensive process. As modern systems become 'smarter' and more complex, traditional production tests are losing in quality and efficiency in terms of fault coverage. BS has shown good diagnostic capability, but is also essentially slow. We are aiming to extend BS applicability to high-speed testing, yet keeping the good diagnostic properties and low cost of equipment.

Despite of constantly improving test automation solutions, the new technological reality elevates the cost of testing. Due to usage of numerous production test types/test phases and a time-consuming testing process, new electronic products need a longer time to reach the market. The mentioned problems are getting even more critical taking into account the fact that system integration has taken the path of the emerging 3D chip technology.

Our future research will target the testability problems in emerging high-performance design architectures – 3D chips and advanced system boards. This will be the basis of future development of an automated and relatively low cost test solution that guarantees a good fault coverage and diagnostic resolution. Our goals are to develop embedded test instruments beyond the state-of-the-art targeting unsolved test challenges especially those applied to 3D architectures, and a new technology that would enable generating adaptable instrumentation cores.

CONCLUSIONS

The primary objective of our future research will be to develop methodologies for designing reliable applications out of non-reliable hardware. In the long term, the obtained solutions will be fundamental in the emerging paradigm of massively parallel computational devices.

The general impact of our research lies in the development of new efficient methods, algorithms and tools for design and test of dependable and mission-critical embedded systems with the goal to reduce time-to-market, cost and improve the quality of applications.

This research will be important for the development of Estonian economy and for the whole society. As a short-term importance, the outcomes can be used in electronics industry and improve its competitiveness in the international markets. The developing scientific excellence will help to strengthen the quality of university teaching for educating students with professional skills. As a long-term impact, new competitive spin-off SMEs are expected to appear in the Estonian electronics industry.

REFERENCES

ARTEMISIA – a legal body for communication with EC – the association for R6D actors in the field of ARTEMIS. https://www.artemisia-association.org/

Bengtsson, T., et al. 2008. Test methods for crosstalk induced delay and glitch faults in NoC interconnects. IET Computers & Digital Techniques, 2, 6, 445-460.

Bryant, R. E. 1986. Graph-based algorithms for Boolean function manipulation. IEEE Transactions on Computers, C-35, 8, 667-690.

Devadze, S., et al. 2009. Fast extended test access via JTAG and FPGAs. Proceedings of 40th International Test Conference – ITC'2009, Paper 2.3, 1-7.

Ellervee, P., et al. 2007. FPGA-based fault emulation of synchronous sequential circuits. IEE Proceedings on Computers & Digital Techniques, 1, 2, 70-76.

European Technology Platform for Embedded Intelligence and Systems ARTEMIS. http://www.artemis-sra.eu/

Hinrikus, H., et al. 2009. Method and device for determining depressive disorders by measuring bioelectromagnetic signals of the brain. US 2009/0054801.

HiPEAC Roadmap: High-Performance Embedded Architecture and Compilation. http://www.HiPEAC.net/roadmap

International Technology Roadmap for Semiconductors. http://www.itrs.net/

Jenihhin, M., et al. 2009. PSL assertion checking using temporally extended high-level decision diagrams. Journal of Electronic Testing: Theory and Applications, 25, 1-12.

Jutman, A., Ubar, R., Peng, Z. 2001. Algorithms for speeding-up timing simulation of digital circuits. DATE, Munich, March 13-16, 2001, 460-465.

Karputkin, A., et al. 2010. Canonical representations of high level decision diagrams. Estonian Journal of Engineering, 16, 1, 39-55.

Lauri, K., et al. 2010. A HPLC study of uremic fluids related to optical dialysis adequacy monitoring. The International Journal of Artificial Organs, 33, 2, 96-104.

Lee, C. Y. 1959. Representation of switching circuits by binary decision programs. The Bell System Technical Journal, 985-999.

Min, M., et al. 2010. Method and device for measurement of electrical bioimpedance. Owner: Tallinn University of Technology. US Patent US7706872B2.

Novak, O., Gramatova, E., Ubar, R. 2005. Handbook of Electronic Testing. CTU Printhouse, Prague.

Pilt, K., et al. 2010. An experimental measurement complex for probable estimation of arterial stiffness. Proceedings IEEE EMBS, 194-197.

Raik, J., et al. 2005. DOT: New deterministic defect-oriented ATPG tool. 10th IEEE European Test Symposium, Tallinn, 96-101.

Raik, J., et al. 2008. Mixed hierarchical-functional fault models for targeting sequential cores. Journal of Systems Architecture, 54, 465-477.

Raik, J., Govind, V., Ubar, R. 2009. Design-for-testability-based external test and diagnosis of mesh-like Network-on-a-Chips. IET Computers & Digital Techniques, 3, 5, 476-486.

Raik, J., Ubar, R. 2000. Fast test pattern generation for sequential circuits using decision diagram representations. Journal of Electronic Testing: Theory and Applications, 16, 3, 213-226.

Ubar, R. 1976. Test generation for digital circuits with alternative graphs. Proceedings of TTU, 409, 75-81 (in Russian).

Ubar, R. 1996. Test synthesis with alternative graphs. IEEE Design & Test of Computers, Spring, 48-57.

Ubar, R. 1998. Multi-valued simulation of digital circuits with structurally synthesized binary decision diagrams. Journal of Multiple Valued Logic, 4, 141-157.

Ubar, R. 2003. Design error diagnosis with resynthesis in combinational circuits. Journal of Electronic Testing: Theory and Applications, 19, 73-82.

Ubar, R., et al. 2005. Energy minimization for hybrid BIST in a system-on-chip test environment. 10th IEEE European Test Symposium, Tallinn, 2-7.

Ubar, R., et al. 2010. Parallel X-fault simulation with critical path tracing technique. Proceedings of Design, Automation & Test in Europe : Dresden, Germany March 8-12, 2010. IEEE Computer Society Press, 879-884.

Ubar, R., Kostin, S., Raik, J. 2008. Embedded fault diagnosis in digital systems with BIST. Journal of Microprocessors and Microsystems, 32, 279-287.

Ubar, R., Raik, J., Vierhaus, T. (eds.) 2011. Design and Test Technology for Dependable Systems-on-Chip. IGI Global, Hershey.

CONTRIBUTION OF FUNDAMENTAL RESEARCH TOWARDS SOLVING CHALLENGES OF CHANGING TIMES IN COASTAL SCIENCE AND MANAGEMENT

T. Soomere

THE MEANING OF WAVES IN MARINE AND COASTAL SCIENCE

The complexity of processes in marine, atmospheric and mainland environments drastically increases at the interfaces between different environments. The sea surface is the most dynamic interface that is commonly and permanently used for a range of activities that are vital for the entire society. Interfaces of similar kind, some of them perhaps less clearly visible for an unarmed eye, play a vital role in the functioning of the entire ecosystem (Emelyanov 2005). The most fascinating property of flexible interfaces is their ability to serve as almost perfect waveguides.

Wave motion is the fundamental phenomenon in Nature that carries energy through the medium. Not only gives it rise to the associated remote impact but also provides a natural way of concentration of the influence of spatially distributed forcing fields into massive energy packets. We experience this commonly when moderate wind blowing over large water bodies creates high waves at some coasts. Moreover, internal energy redistribution between wave components may lead not only to gradual changes in the wave properties but sometimes to extremely furious and dangerous events (Kharif et al. 2009)

The situation becomes even more complicated when three primeval forces – air, water and land – meet in the coastal zone. The complexity of the system, the extension of the interplay of different impacts and the variety of outcomes drastically increases in such regions. The industry-driven challenge to manage the processes affecting coasts and ports and the complementary need to ensure environmental health requires the joint efforts of experts from different fields to understand the nature of the changes and to derive effective solutions to the inevitable problems. Another dimension here is the life: coastal and shallow-water areas are the primary life reproduction zone. It definitely hurts when we lose control over some mainland or sea domain because of a major disaster, such as the Gulf of Mexico oil pollution or the

Fukushima event. However, if we lose a spawning or nursing shallow-water region, the chain of life reproduction will be broken, with eventually much wider consequences.

Ocean waves are the key players here. While the knowledge of wave properties is decisive for many offshore issues, equally important is the impact of waves in the nearshore and upon the coast. Wind waves carry substantial amounts of energy and create extensive loads on the seabed, natural coast and coastal engineering structures, serving in this way both as blessing (source of energy) and curse (coastal destruction, wave-induced flooding, danger to lives, etc.). Although the spatio-temporal scales of various wave motions and their appearance may vary a lot, the dynamic similarity of wave phenomena at different scales (Didenkulova et al. 2011) greatly simplifies their research.

The primary role of ocean waves suggests that a useful indication of the coastal changes can be found in the wave field. A change to wave properties may result in intricate spatial and temporal changes to coastal dynamics and these may be completely different in neighbouring areas (Soomere, Räämet 2011). This feature motivates to use the term 'coastal science' in a somewhat non-traditional manner below, to denote the pool of research areas (involving physical coastal oceanography, coastal engineering and management, coastal processes, studies into marine-induced natural hazards, etc.), in which the direct impact of waves plays a key role. This understanding embraces and emphasises the fact that the majority of the agents of marine-induced impacts have a wave nature, at least in a certain stage of their generation or propagation. A drawback of this definition is that several other important fields of physical oceanography (such as optical oceanography (Arst 2003) or various remote sensing techniques (Arst 2003; Kutser 2004)) are not considered in what follows.

COASTAL SCIENCE IN ESTONIA

The importance of studies into waves, coastal processes and related items is obvious for Estonia with its 3800 km long coastline hosting many harbours and a substantial part of economy relying on the geographical location of the country. Systematic studies of Estonian coasts started more than a half-century ago. The basic knowledge about the coasts and processes driving their development was formulated by the mid-1970s (Orviku 1974). The obtained knowledge is still quite fragmented, mostly descriptive, and has not been brought to the level accessible for experts in neighbouring areas; perhaps because of the change in the social system in 1991 and accompanying change in the language of scientific and applicational publications from Russian to English (Ojaveer et al. 2000).

Although systematic observations of hydrometeorological parameters were started in Estonian coastal waters already at the beginning of the 19th century, contemporary physical oceanography was launched in the 1970s. During the 20th century, oceanographic research in Estonia was mostly focused on offshore domains, except for attempts towards ecosystem modelling (Tamsalu, Myrberg 1995). The research in coastal science and oceanography was not only disconnected from each other but also almost separated from the fisheries research and marine biology (Ojaveer et al. 2000).

The growing maturity of Earth sciences in Estonia and the increasing needs of society favoured the interaction between marine and coastal science from the beginning of the 21st century. Requirements of industry (Elken et al. 2001), an apparent intensification of coastal processes (Orviku et al. 2003), the threat of sea level rise and associated coastal erosion (Kont et al. 2008) and the impact of extreme storms (Fig. 1) (Suursaar et al. 2006) called for the implementation of models and concepts able to not only post-factum describe what has happened but also to predict what might happen and how the society should react.

The scientific community responded by entering the operational oceanography system from about 2000, launching the teaching of port and coastal engineering at Tallinn University of Technology, implementation of real-time forecast of sea level and, last but not least, by rapid advancement of the fundamental research into the dynamics of various wave-driven phenomena in the marine and coastal environment.

This research provided extremely interesting information. For example, waves from high-speed ships are frequently considered with respect to their effects on the coastal structures, or with respect to safety (Soomere 2005). They also have great potential for use as a small-scale model for studies of extremely dangerous phenomena such as landslide-induced tsunamis or run-up and overtopping by storm waves (Didenkulova 2011). Their impact on beaches allows quantification of the reaction of coasts under changing natural forcing factors (Soomere et al. 2011c).

Another challenge is the use of semi-persistent surface current patterns (Soomere et al. 2011a). Unfavourable currents may bring oil spills to the beaches. However, through the smart placement of offshore activities, favourable current patterns can be used to provide protection to coasts and vulnerable sea areas (Soomere et al. 2011b).

Figure 1.
High waves attack the dune forests at Pirita Beach in the aftermath of an extreme storm of January 2009. Photo by A.Kask (Soomere et al. 2008).

WAVE CLIMATOLOGY AND VESSEL WAKES

Systematic studies into the properties of the Baltic Sea waves go back more than a half-century (Soomere 2008). The relevant research in Estonia focuses on the major aspects of wave dynamics and the impact of waves on coastal processes and coastal engineering structures, ranging from short wave generation and long-wave excitation, following wave propagation over the sea surface and ending up with wave-coastal zone and wave-structure interaction, including the related mathematical and computational background. The key challenge has been and still is the advancement in the understanding on the properties of waves of different kind and their spatio-temporal variations. Equally important aspects are the role of surface waves among marine-induced hazards and fundamental and generic aspects of wave science (Quak, Soomere 2009).

Early attempts to reconstruct the Baltic Sea wave climate have covered relatively short periods or have been concentrated on specific areas (Soomere 2008), mainly because reconstruction of the Baltic Sea wave fields is an extremely complicated task and the amount of wave data is limited. The situation was resolved by merging historical wave observations, available instrumental measurements and long-term numerical wave hindcast (Soomere, Räämet 2011). The changes in wave climate apparently were marginal in 1960-1985. A drastic increase in the annual mean wave heights occurred in the northern Baltic Proper in the 1980s and 1990s, followed by a rapid decrease since the mid-1990s. Simultaneously, wave heights along the Lithuanian coast show a remarkable minimum in the 1990s. There exists a complicated pattern of temporal changes to the wave properties over the Baltic Sea: some areas experience a decrease in average wave heights but increase in extreme wave heights (Soomere, Räämet 2011). The key reason for such patterns is the substantial turn in the predominant wind direction: an increase in the frequency of SW winds.

A fascinating object of studies has been nonlinear wake waves from high-speed ferries (Soomere 2007). High-speed ferries are interpreted as those vessels, for which the regular sailing regime contains sections in which the depth Froude number (the ratio of the ship's speed and the maximum phase speed of linear water waves for the given depth) exceeds 0.6. The damaging potential of vessel wakes has been widely recognised in archipelago areas, narrow straits and inland waterways (Soomere 2007). Their role as a qualitative new hydrodynamic forcing factor for many high-energy coastal regions, the importance of their remote impact (Soomere 2005), their far-field properties (Soomere 2007) and implications to the coast and the near-shore (Soomere et al. 2011c) have been recognised based on recent studies in Tallinn Bay.

WAVE MATHEMATICS AND WAVE-DRIVEN COASTAL PROCESSES

The recent years have brought rapid progress in understanding the fundamental aspects of wave science such as rogue waves (Didenkulova et al. 2006a; Kharif et al. 2009), interactions of solitons (Soomere 2009), wave run-up or general wave mathematics (Quak, Soomere 2009). The studies into interaction of solitons on water surface are one of the few fields of physical oceanography that has substantially contributed to cutting edge fundamental research. For example, surface soliton interactions in the deep ocean have initiated research into optical rogue waves (Solli et al. 2007). Nonlinear interaction of their sister phenomena – shallow-water solitons – may lead to a drastic increase in the local surface elevation (Peterson et al. 2003) accompanied by an even larger increase in the slope of wave fronts (Soomere 2010).

This effect is the core of a unique mechanism for the formation of long-living rogue waves (Kharif et al. 2009) (Fig. 2).

The problem of run-up of ocean waves along the beaches and seawalls has been extensively studied for a number of wave classes, mostly with idealised, symmetric shapes. A breakthrough in this field is the demonstration that even small deviations of the shape of the incident wave from the idealised one may drastically increase the run-up height (Didenkulova et al. 2006b). This effect may be even more drastic in non-plane geometry. Wave run-up along certain convex bottom profiles also shows fascinating features: effectively no wave reflection occurs along most of the profile, shoaling effects are very strong, and extremely strong reflection together with high water velocities may occur in the immediate vicinity of the shoreline (Didenkulova et al. 2009).

The Baltic Sea coasts develop in rare conditions of a large micro-tidal water body hosting a highly intermittent wave regime (Soomere, Healy 2005). Their evolution typically has a step-like manner: rapid changes take place when rough waves occur simultaneously with high water levels (Orviku et al. 2003). Especially the southern and eastern coasts of the Baltic Sea consist of

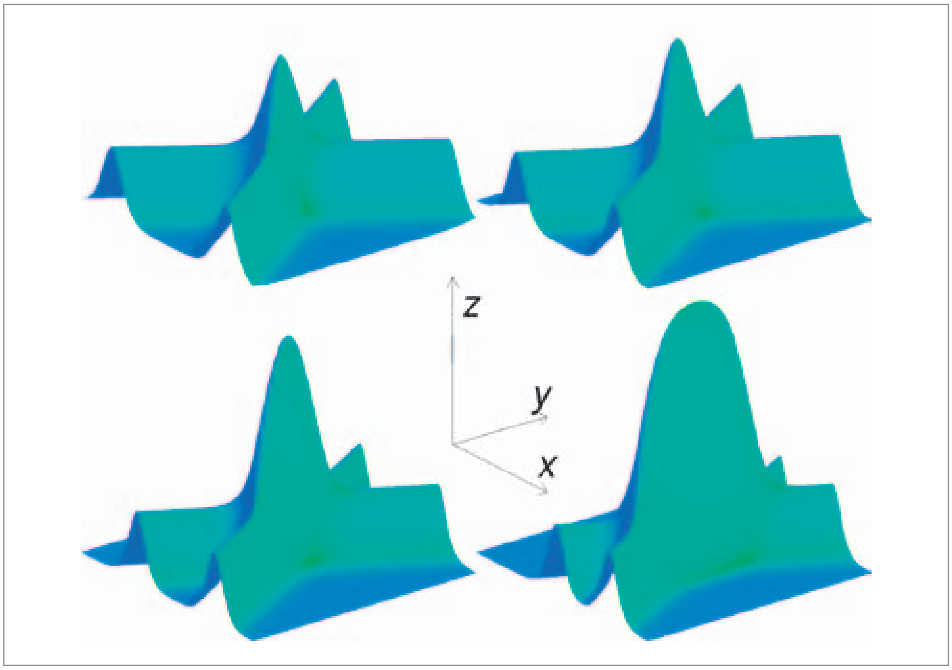


Figure 2.
Surface elevation in the vicinity of the interaction area, corresponding to the interaction of solitons with equal amplitudes intersecting under different angles (Peterson et al. 2003).

relatively soft sediment and are sensitive to large hydrodynamic loads. Quantification of wave-driven coastal processes (Soomere et al. 2008) and developing the relevant models for coastal engineering are thus extremely important (Emelyanov 2005), and also provide important results and experience for the global science (Soomere, Healy 2005).

PREVENTIVE METHODS FOR COASTAL PROTECTION

An additional coastal hazard is the current-driven pollution transport. Like the wave field, it is mostly created remotely by a spatially widely distributed impact of several forcing factors and hits the nearshore in a highly variable manner. The ever increasing ship traffic in vulnerable sea areas such as the Baltic Sea calls for novel methods for mitigating the relevant risks. The major bottleneck is that there is not yet a deterministic method to reproduce the floating object drift or the drift of oil pollution. It is, however, possible to quantify the potential of different offshore domains to serve as a source of danger (e.g., oil pollution) to the coastal environment by means of statistical analysis of a large number of trajectories of propagation of pollution particles. The technology binds together a 3D ocean model, calculations and analysis of a large set of Lagrangian trajectories, and a method for construction of the optimal fairway (Soomere et al. 2011b).

This approach has highlighted the presence of very rich, usually concealed internal structure of semi-persistent transport patterns in certain sea areas (Soomere et al. 2011a). The 'fair way' of dividing the risks equally between the opposite coasts is a local solution that does not normally provide the minimum level of risk for the entire water body (Soomere et al. 2011b). Better approximations to the global minimum of risks are fairways roughly following the minima for probabilities of coastal hit or the maxima for the time it takes for the potential pollution to reach the coast (Soomere et al. 2011b) (Fig. 3).

THE GAPS AND AMBITIONS

The processes of surface wave generation and propagation have been understood quite well for the open ocean and relatively deep shelf seas. There has also been excellent research into storm surges, wave-induced setup, properties of waves attacking beaches with relatively simple geometry, wave overtopping over breakwaters and seawalls, wave-induced near-bottom velocities

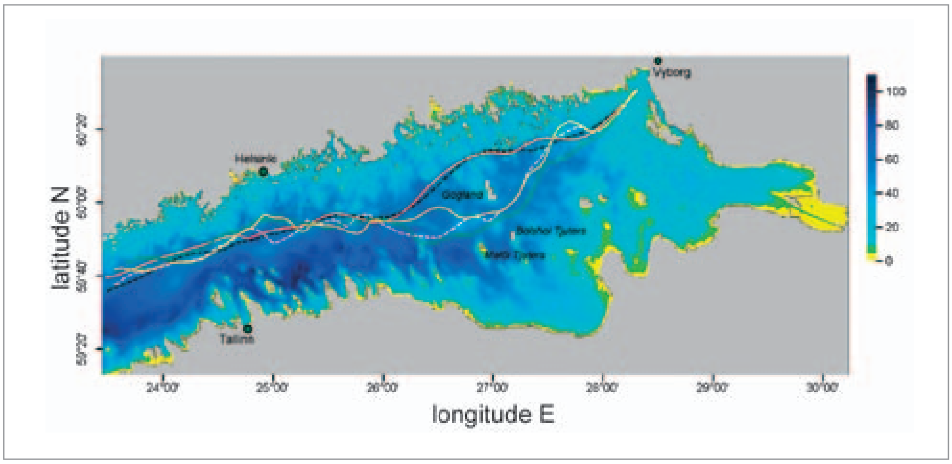


Figure 3.
Optimum fairways from the Baltic Proper to Vyborg according to the spatial distributions of the probability for coastal hits (solid lines) and of the particle age (dashed lines) at resolutions of the underlying ocean model of 2 nautical miles (nm) (red and black respectively), 1 nm (green and cyan) and 0.5 nm (yellow and white). The depth scale to the right of the map is given in metres (Andrejev et al. 2011).

and their impact on sediment, etc. The wave transformation and shoaling of water waves in basins of variable depth is a classical task of fluid dynamics and in physical oceanography. This field is, however, far from being completed and actually is in a rapid development phase.

There are extensive gaps in the coastal science, including vague knowledge of the course and consequences of the conversion of wave motion (in which water particles have moderate speeds) into motion of water masses. The crucial but so far insufficiently understood aspect in wave-driven marine hazards is how the energy and momentum supplied to the wavelike motions of water masses create a highly organised onshore motion of water masses that may penetrate far inland, cause either rapid devastation or slow degradation in the nearshore and at the coasts. The key questions here are:

- what happens with the wave motion in the nearshore, especially when waves become essentially nonlinear and/or interact with each other and with the coast;
- how the wave-induced movements of the waterline position follow changes in the geometry of the nearshore;
- how high velocities of water particles are systematically created during wave transformation over beach profiles of different shape and owing to different kinds of interactions;

- how natural coasts react to changes in wave parameters and accompanying changes in hydrodynamics.

The relevant studies are highly complicated because of the enormous forces occurring during this conversion of ocean waves, the high variability of the geometry and other properties of nearshore areas, problems with the representation of some phenomena (such as tsunami waves) in laboratory scales, virtual impossibility to study the reaction of natural beaches to different wind wave climates, etc.

Another large gap concerns wave climate changes as well as potential impacts of such changes on the nearshore and the coast. Only issues connected with changes to the wave height have been addressed in some detail (Weisse, von Storch 2010) whereas changes to wave propagation direction, periods or group structure may lead to much more drastic reactions of the coasts (Soomere et al. 2011c).

While the classical wave dynamics is commonly understood as a research field where quite exact estimates of the properties and impact are possible, it is an opinion widely spread that physical coastal science (incl. coastal oceanography and the management of coastal processes) is mostly empirical, should rely on simplified concepts and models, and should largely be directed towards applications. The relevant research should reach out farther in the future, applying mathematical methods (theoretical, computational and statistical ones) to wave phenomena to make substantial progress in the analysis and modelling of factors that shape the coasts and to advance towards predictive methods in coastal oceanography and coastal engineering.

A reasonable ambition is that the research into wave dynamics in coastal environment should be directed towards a unified framework for the analysis of the mechanisms responsible for wave-driven phenomena in the nearshore and at the coasts, and their impact. This includes analysis of wave-induced loads on coastal and offshore structures and research into marine-induced coastal hazards; from large-scale disasters (tsunamis and storm surges) down to small-scale phenomena responsible for long-term erosion processes, and breakthroughs in the mathematical analysis and modelling of various wave-driven coastal hazards. This framework, based and further developed in terms of fundamental studies into surface and internal waves, provides a platform for scientifically sound applications for coastal engineering and planning, and mitigation and management of coastal marine hazards.

RESEARCH DIRECTIONS

Realising this vision will partially entail research in the zone not yet covered by classical (both linear and nonlinear) wave theories. Progress in this field is urgently needed: even now scientists and engineers rely here on empirical relations (usually established in site-specific or scale-specific conditions) or simplified (semi-) operational models. Progress is specifically important in the context of the Baltic Sea where key properties of waves differ from those on the open ocean coasts (Soomere 2008).

The investigations for the nearshore and coastal zone will address a cluster of key phenomena:

- shallow-water solitons and wave interactions;
- wave amplification and reshaping in the nearshore;
- wave run-up along the coast and overtopping coastal engineering structures;
- extreme waves and wave-driven floodings;
- impact on coastal processes of changes in the wind wave climate;
- anthropogenic pressure on the coast through wake waves from high-speed vessels and through current-induced transport of adverse impacts.

This cluster, although it does not include all possible important phenomena for the coastal science, is representative of the most energetic processes, drivers and hazards in the nearshore. It is universal in the sense that it involves a minimum set of aspects covering the basic features and/or consequences of the wave energy conversion at different scales, effects amplifying its potential impact on the coast, and test and study 'tools' allowing for systematically identifying and quantifying the role of wave-driven impact and hazards upon the coast in different time scales. All the listed items play a large role in the marine-induced coastal hazards whereas there are large gaps in our relevant knowledge or understanding and substantial progress in closing the major gaps is feasible within one or two coming decades.

CHANGES TO THE WIND WAVE CLIMATE

Climate change is leading to many potential problems for the coastal areas. For Estonian coasts, the largest problems are connected with the simultaneous increase in the water level and storminess. This increases the risks of devastating flooding but also the potential impact of the erosion of the coastline accompanied by gradual degradation of beaches and/or depositional coasts due to the more frequent occurrence of a combination of high water levels and intense waves, and shortening of the ice cover (Orviku et al. 2003). The key driving agent of coastal erosion is the supply of energy to the coastal zone by approaching wind waves. The basis for the development of theoretical concepts and practical applications is a thorough knowledge of the properties of waves of different kind in all phases of their life-cycle, from generation, propagation, interactions and transformation owing to various reasons until their breaking at the seaward border of the surf zone.

It is natural to expect that properties of wind waves change along with the climate change. An important issue is, therefore, the analysis of spatio-temporal variations in the local wave climate (from weekly scales to long-term variations), with emphasis on the northern Baltic Sea and the Estonian coasts, including their consequences to coastal development. These problems have been thoroughly studied in terms of wave height (Broman et al. 2006; Weisse, von Storch 2010). There is, however, an essential gap in such studies: changes in other key properties of wave fields (wave period, propagation direction, group structure, flow velocity) may have an equal, if not larger, impact on the nearshore, coastal processes and coastal management (Soomere et al. 2011c). These changes affect wave refraction and thus may lead to substantial changes in the spatial distribution of areas sheltered from intense wave impact. The research into changes in wave periods and directions (Fig. 4) has a great potential to locate nontrivial, not yet identified or understood, changes in the wind patterns (e.g., in the intensity, size and duration of storms) over north-eastern Europe, and in this way contributes to the understanding of the status of and changes to the climate system, and also provides valuable ground truth for numerical simulations by atmospheric models.

It is not straightforward and at times hardly possible to identify changes in these parameters (and the related impacts) on the open ocean coasts. Studies of wave statistics and wave storms in the Baltic Sea, however, offer a possibility to consider separately the impact of each storm and effects of changes to the wave periods and directions that are usually masked by swell on the ocean coasts (Soomere, Räämet 2011).

Reliable and detailed directional wave climatology is a necessary precondition for progress in the research into the implications of nonlinear shallow water wave interactions in realistic conditions. The quantification of these changes forms a crucial part on the way towards understanding how the wave fields impact the nearshore and how wave-induced local coastal hazards (e.g., wave-induced water level setup) are changing. An equally important question in terms of wave energy potential is how the properties of highest waves

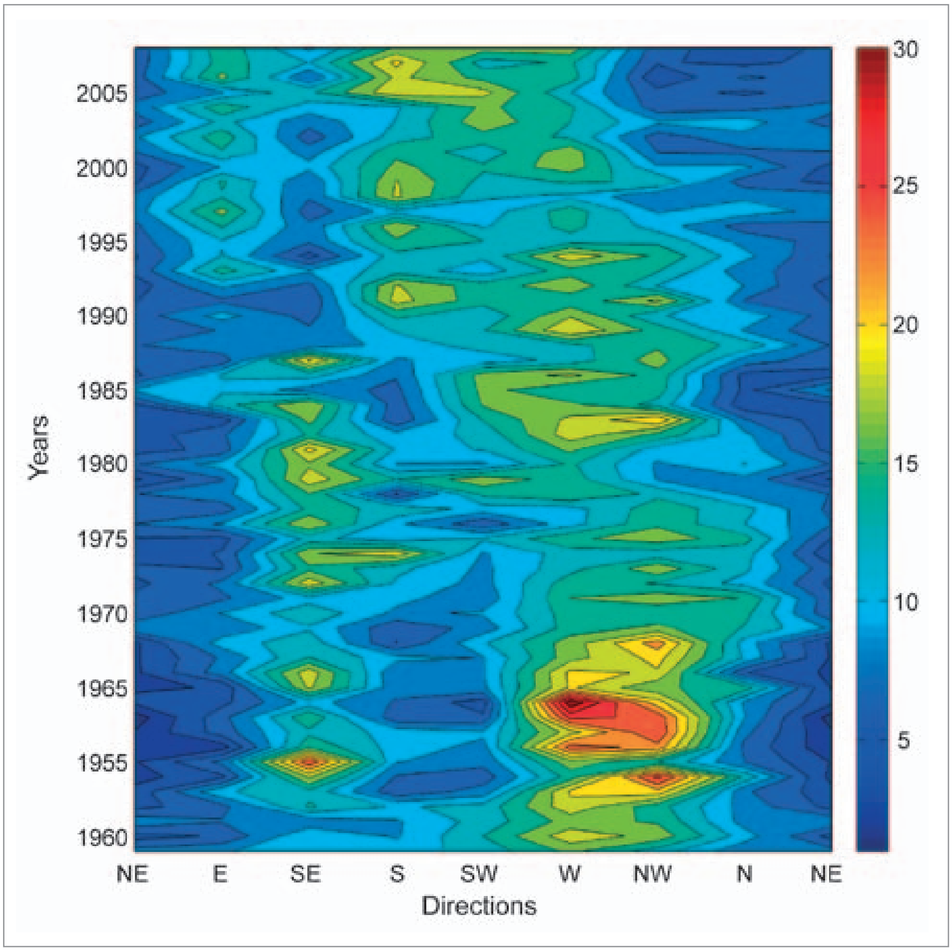


Figure 4.
Observed directional distribution of the wave approach at Narva-Jõesuu for 1954-2008. The colour code shows the frequency of occurrence (%) of waves from a particular direction (Soomere, Räämet 2011).

change in time and space and which combination of wave properties might have the most extreme Baltic Sea storms.

Further research using high-resolution numerical reconstructions of the Baltic Sea waves has an obvious potential for a much more comprehensive understanding of the variations in the nearshore wave properties and will contribute to the analysis of the reaction of the coasts to potential changes in the wind regime. These efforts may greatly benefit from studies concerning the improvement of reconstructions of open sea wind fields. Given the limitations of wave hindcast, the policy to merge numerical simulations with historical visual wave observations is, today the only way towards understanding the reasons behind the systematic mismatch of the overall changes in wave intensity and wind speed (Soomere, Räämet 2011).

A closely related aspect is the climatology both upper-layer and surface winds over Estonia that seems to contain abrupt shifts rather than trends or periodic variations. The changing forcing (e.g. structure and/or trajectory of the storms) may give rise to substantial changes in the wave field and drastically influence shipping safety and coastal and offshore structures. A major bottleneck is that even the best existing atmospheric models are not yet able to reconstruct specific features of the wind field in Baltic Sea sub-basins such as the Gulf of Finland.

IDENTIFICATION OF CONSEQUENCES TO COASTAL DEVELOPMENT

Natural changes in wave properties have already affected the status of several sections of the Baltic Sea coasts. Detailed research into these changes is one of a few feasible ways to provide 'ground truth' for reconstructions of the historical wave climate. The 'hot spots' are located in the eastern part of the Gulf of Finland and in certain sections on the Estonian coasts (Orviku et al. 2003; Soomere et al. 2008). Given the relatively slow evolution of most of the Estonian coasts, the concept of (almost) equilibrium beaches and methods derived for their analysis (Soomere, Healy 2011) apparently are appropriate tools for studies of their past, present and future.

A feasible way to gather progress in studies on the implications of wind wave climate change on the coast, nearshore and semi-sheltered ecosystems consists in using ship-induced waves as a test tool (Didenkulova et al. 2011). The impact of changing wave properties can be studied from a variety of different viewpoints by a proper choice of the study site.

A major bottleneck in the analysis of the reaction of the bottom sediments to the potential changes in hydrodynamic activity is the lack of express methods for the quantification of this reaction. A feasible way is to rely on the relevant optical methods. A promising nucleus has been worked out in (Erm, Soomere 2006). Studies into the described issues will result not only in a better understanding of the course of coastal processes and the potential degradation of beaches but also contribute to the development of timely countermeasures and a much better basis for the general decision-making procedures in coastal affairs.

WAVE AMPLIFICATION, RESHAPING, RUN-UP AND OVERTOPPING

Studies into the role of surface waves in marine-induced hazards and coastal zone management form a classical part of coastal engineering. There has been rapid progress in studies into the transition of surface waves in the nearshore and accompanying amplitude amplification. Most of the studies have been performed in the framework of different versions of the long wave theory. The amplification of waves and extremely strong reflection in the immediate vicinity of the shoreline of some types of natural convex bottom profiles (Didenkulova et al. 2009) apparently are generic for many beaches just offshore from the seaward end of the equilibrium beach profile (Soomere et al. 2008). Their implications (especially for the impact of tsunami-like events) evidently are substantial but the scope of their occurrence is unclear. The possible scenarios for wave propagation in such conditions include the potential for the resulting high waves to propagate through the surf zone until the immediate vicinity of the coastline, contribution of certain 2D effects towards additional amplification of the wave impact and the option for the wave shape to become essentially asymmetric followed by extensive increase in the steepness of the wave front as described above.

The potential increase in the steepness of the wave front is the most alarming issue because specifically this quality characterises the extremes of the run-up process (Didenkulova et al. 2006b). There is thus an acute need for understanding what actually happens with such waves in the nearshore and with which intensity, and how large the probability is of having unexpected penetrations of high-energy events and accompanying large water velocities to the immediate vicinity of the coastline. This includes the research into the influence of the incident wave form on the extreme characteristics of wave run-up and their dependence on different aspects of wave shape (such as front/back or crest-trough asymmetry).

The combination of the phenomenon of an increase in the steepness of travelling nonlinear waves over some profiles with the subsequent increase in the run-up height needs specific attention. Steeper wave fronts generally mean larger near-bottom velocities and more intense reaction of bottom sediments to these, and also larger potential of the wash by such waves to reach the dune toe or unprotected sediments far from the waterline. It is necessary to systematically quantify their potential impact in terms of generic methods such as peak-over-threshold analysis. The results have a number of applications in the management of beaches and mitigation of wave-driven hazards.

The possibility of being hit by a long-wave-induced (e.g. tsunami) or a storm-surge-driven flooding has a great relevance not only for low-lying Estonian cities (Pärnu or Haapsalu) but also for many other cities (Venice, Hamburg, St. Petersburg). The key message from (Didenkulova et al. 2006b, 2009) is that the devastation crucially depends on the wave shape. It is therefore important to make clear the role of the extreme run-up potential specifically in the context of large-scale hazards, accounting for the interaction (or Mach reflection) of shallow-water solitons. Further understanding, detailed analysis, ways of modelling and, if possible, derivation of express methods for estimates of the level of risk are crucial in tsunami studies where run-up intensification of unknown nature has often been observed in post-tsunami field surveys. Progress in this area is crucial for improvement of the predictive capacity of models of natural hazards and worst-case scenarios.

As the run-up process of long waves is largely scale-invariant and waves in the immediate nearshore are effectively long waves for realistic sedimentary beaches, the results obtained from one particular scale can be directly used for processes over a wide range of spatial and temporal scales such as tides, tsunamis, swells, ship waves, and down to local wind waves. It is very likely that progress here can be obtained analytically (Quak, Soomere 2009). The results obtained so far are the most encouraging. In particular, the derivation of express estimates of run-up characteristics for cases when exact information about the form of the incident wave is not available, has direct applications in problems concerning the stability of sedimentary coasts, overtopping of seawalls and breakwaters, in terms of safety for beach users, etc. An equally important feature is the potential for producing science-based recommendations and raising awareness of both scientific and coastal communities with respect to both short-term hazards and long-term wave-driven impacts.

UNDERSTANDING ANTHROPOGENIC PRESSURE: WAKE WAVES FROM HIGH-SPEED VESSELS

Although the basic features of linear ship wakes are well known, wakes from high-speed vessels provide a multitude of important new features, such as monochromatic packets of short waves, solitary and cnoidal wave trains ahead of the ship and associated depression areas (Soomere 2007). Recent studies have revealed other specific properties of vessel wakes, such as their extremely high spatial variability or the group structure of wakes. Latest experiments have produced evidence about other potentially important features such as asymmetry of wave shapes (that may substantially change the run-up properties, see above) or the ability of transport of large volumes of water (Soomere et al. 2011c) (Fig. 5).

There is a definite need for further studies into the properties of wakes from high-speed vessels, with the particular goal to identify the emerging features (e.g., transport of large water masses or depression areas followed by rapid, bore-like events (Soomere 2007)). This may provide a key to understand the phenomenon of the drastic reaction of some coasts ship wakes (Soomere et al. 2011c). Quantification of this influence on the marine ecosystem and its comparison with the impact caused by natural factors is a major challenge for areas hosting intense ship traffic. Moreover, smart use of vessel waves may lead to substantial progress in studies into implications of wind wave climate changes to the coast, nearshore and semi-sheltered regions.

As the impact of a moving ship on water surface is equivalent to that of a moving atmospheric pressure patch, this research is of paramount importance for the understanding of the nature, properties, appearance and potential impact of (meteorological) tsunamis and related flooding scenarios in safe, controlled and repeatable conditions. The dynamical similarity of some vessel waves and landslide-induced tsunamis (Didenkulova et al. 2011) makes ship waves an extremely valuable resource for modelling of dangerous phenomena.

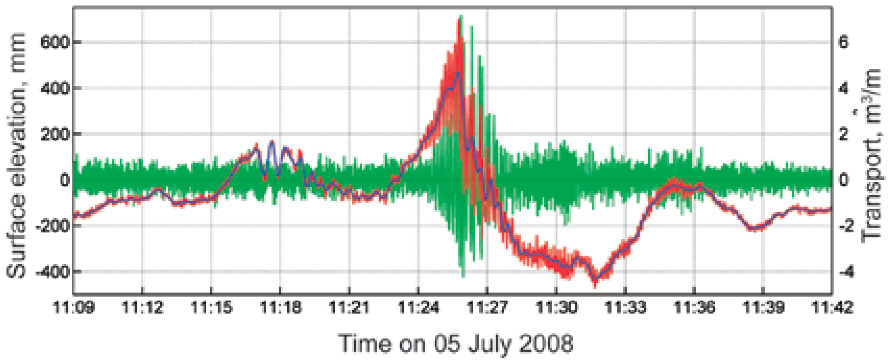


Figure 5.
De-meaned water level variations (green, rapid background oscillations), the instantaneous (red) and smoothed (blue) values of water transport for a wake (per each meter of the coastline) from a large high-speed ferry "Superstar" on 05 July 2008 near the Aegna jetty, Tallinn Bay (Soomere et al. 2011c).

An intrinsic component of the coastal science is the research into fundamental issues such as nonlinear wave interactions or general applied wave mathematics. This is a feasible way towards further understanding of the nature and appearance of particularly high and steep rogue waves (sneaker waves in the nearshore) that may attack beaches and structures from an unprotected side and in unusual locations (Soomere 2010) and serve as a specific source of danger to offshore and coastal structures. A comprehensive picture about such phenomena is missing even in the simplest framework – the (nonlinear) shallow-water theory (Soomere 2009, 2010). It is not known how the structures, created at the intersection (or by Mach reflection) of solitary waves (that lead to extreme slopes and/or changes in the orientation of wave crests (Soomere 2010)), will behave when they hit a harbour or river mouth: whether they disperse or keep their shape. Their potential in causing high-impact and/or run-up events has not been addressed at all.

Rapid progress apparently is feasible through studies of nonlinear interactions of directionally spread and/or crossing shallow-water waves in the vicinity of sheltered fairways and harbour entrances. Ideally, the research into the generation and persistence of steep wave humps in the nearshore (incl. probability distribution functions for such events) should be performed simultaneously with the above-described studies into the local wave climate and studies of the run-up and overtopping events at various incident waves. The results may serve as the basis for identifying the 'worst case' scenarios of wave attacks and also for quantifying the status of and the changes to sediments in areas where such waves may lead to high water velocities.

Large-amplitude or solitonic internal waves pose serious (currently neither well understood nor accounted for) danger for oil and gas platforms, pipelines, and to other bottom objects in stratified water bodies, and may cause substantial transport of near-bottom water masses. Studies of internal waves in the Baltic Sea basin are infrequent and fragmented. There is an obvious need for further refinements of both theoretical research and ocean modelling in this area specifically to account for the complex dynamics associated with the presence of both surface and internal waves. This research might considerably contribute to soliton science (Soomere 2009) and widen our understanding of fundamental problems of mathematical physics such as the integrability of certain classes of equations.

The coastal science should also foster studies into problems of coastal pollution. It is important to identify the regions, where human activities pose an increased risk to the coastal zone, and develop ways to reduce the risk by, e.g., placing these activities in other areas (for example by rerouting the ship traffic). The idea is to determine beforehand the regions, for which the probability of pollution propagation to vulnerable areas is considerably larger or smaller than expected from purely geometric considerations. In essence, this is the inverse problem of quantification of the offshore areas in terms of their potential to serve as a source of coastal pollution. It can be approximately solved using statistical analysis of a large pool of direct solutions (trajectories of water particles) (Soomere et al. 2011b).

The major challenge here is how to extract useful information from a large collection of various simulated data. Therefore, a key contribution from methods of this type is the possibility of extracting and visualising important information that usually remains concealed in classical methods of the analysis of ocean currents. Results from this research field will eventually serve as a scientific basis for establishing in the vicinity of vulnerable coastal sections limitations for ship traffic, which are both economically feasible and tolerable by the local coastal community. In particular, the developed method can be used as a preventive tool supporting decisions about how far the fairway for ships carrying dangerous cargo should be located from any vulnerable areas.

FROM COASTAL SCIENCE TO SOCIETY

The sketched research is intrinsically cross-disciplinary and lies on a border of several disciplines in Earth sciences (physical coastal oceanography, coastal processes, and climate studies), coastal engineering and management of complex systems (coastal management, mitigation of marine natural hazards, etc.). The results evidently have a substantial societal dimension for Estonia – a country with a 3800 km long coastline. The research focuses on a strategic area of R&D in Estonia: protection of the environment.

The above research and gathered understanding is a key component for the comprehensive analysis, modelling, hindcast and forecast of coastal processes for common types of Estonian coasts, incl. mapping of their reaction to changing wave conditions. The results are expected to lead to a number of direct applications in a variety of marine and coastal matters. For example, knowledge of the basic features of the forcing factors is the starting point of reliable wave and water level forecast. Especially in the context of Estonian coasts it is important to develop various site-specific but scientifically sound applications for coastal engineering and coastal zone management (express methods for calculations of wave loads in areas with specific wave climates, technologies of the use of marine hydrodynamics for coastal protection, models of local coastal processes, wave energy aspects, mitigation of marine-induced coastal hazards, etc.). Probably the most acute need of the society is the comprehensive understanding and forecast of the location and course of 'worst-case' scenarios of marine-induced hazards.

Last but not least, the specific conditions of the almost tideless and sheltered (in terms of waves) Baltic Sea allow for separate studies of many wave and coastal phenomena that are masked by tides and remote swells on the open ocean coasts. The results, if put into proper context, may serve as a unique contribution to understanding marine-induced coastal hazards. Many other natural features of the Baltic are also favourable for the described studies, e.g. sedimentary coasts together with a long history of wave studies and a dense observation network.

Acknowledgements

Several research ideas presented above reflect the material in research proposals submitted to European Research Council Advanced Scientist grant and another proposal to Estonian Ministry of Education and Science. Contributions by Ewald Quak and Ira Didenkulova to these documents are gratefully acknowledged. The preparation of the manuscript was partially supported by the targeted financing by the Estonian Ministry of Education and Science (grant SF0140007s14), the Estonian Science Foundation (grant 7413), and the EC FP7 (grant 217246) through BONUS BalticWay project.

REFERENCES

Andrejev, O., et al. 2011. The role of spatial resolution of a three-dimensional hydrodynamic model for marine transport risk assessment. Oceanologia, 53, 1-TI, 309-334.

Arst, H. 2003. Optical Properties and Remote Sensing of Multicomponental Water Bodies. Springer Praxis, Berlin Heidelberg New York.

Broman, B., et al. 2006. Trends and extremes of wave fields in the north-eastern part of the Baltic Proper. Oceanologia, 48, S, 165-184.

Didenkulova, I., et al. 2006a. Freak waves in 2005. Natural Hazards and Earth System Sciences, 6, 1007-1015.

Didenkulova, I. I., et al. 2006b. Runup of nonlinearly deformed waves on a coast. Doklady Earth Sciences, 411, 8, 1241-1243.

Didenkulova, I., Pelinovsky, E., Soomere, T. 2009. Long surface wave dynamics along a convex bottom. Journal of Geophysical Research, 114, C07006.

Didenkulova, I., Pelinovsky, E., Soomere, T. 2011. Can the waves generated by fast ferries be a physical model of tsunami? Pure and Applied Geophysics. DOI: 10.1007/s00024-011-0289-z.

Elken, J., et al. 2001. Hydrodynamical and geological investigations of possible deep harbour sites in north-western Saaremaa Island: Overview and conclusions. Proceedings of the Estonian Academy of Sciences. Engineering, 7, 2, 85-98.

Erm, A., Soomere, T. 2006. The impact of fast ferry traffic on underwater optics and sediment resuspension. Oceanologia**,** 48**,** S, 283-301.

Emelyanov, E. M. 2005. The Barrier Zones in the Ocean. Springer, Berlin Heidelberg New York.

Kharif, Ch., Pelinovsky, E., Slunyaev, A. 2009. Rogue Waves in the Ocean. Springer, Berlin, Heidelberg.

Kont, A., et al. 2008. Implications of sea-level rise for Estonia. Journal of Coastal Research, 24, 2, 423-431.

Kutser, T. 2004. Quantitative detection of chlorophyll in cyanobacterial blooms by satellite remote sensing. Limnology and Oceanography, 49, 6, 2179-2189.

Ojaveer, E., Rannak, L., Laevastu, T. 2000. One and a Half Centuries of Sea and Fisheries Investigations in Estonia. Estonian Academy Publishers, Tallinn.

Orviku, K. 1974. Morskie berega Éstonii [Estonian coasts]. Institut geologii, AN ÉSSR, Tallinn, (in Russian).

Orviku, K., et al. 2003. Increasing activity of coastal processes associated with climate change in Estonia. Journal of Coastal Research, 19**,** 2, 364-375.

Peterson, P., et al. 2003. Interaction soliton as a possible model for extreme waves in shallow water. Nonlinear Processes in Geophysics, 10, 6, 503-510.

Quak, E., Soomere, T. (eds.) 2009. Applied Wave Mathematics: Selected Topics in Solids, Fluids, and Mathematical Methods. Springer, Heidelberg.

Solli, D. R., et al. 2007. Optical rogue waves. Nature, 450, 7172, 1054-1057.

Soomere, T. 2005. Fast ferry traffic as a qualitatively new forcing factor of environmental processes in non-tidal sea areas: a case study in Tallinn Bay, Baltic Sea. Environmental Fluid Mechanics, 5, 293-323.

Soomere, T. 2007. Nonlinear components of ship wake waves. Applied Mechanics Reviews, 60, 120-138.

Soomere, T. 2008. Extremes and decadal variations of the northern Baltic Sea wave conditions. Pelinovsky, E., Kharif, Ch. (eds.) Extreme Ocean Waves. Springer, 139-157.

Soomere, T. 2009. Solitons interactions. Meyers, R. A. (ed.) Encyclopedia of Complexity and Systems Science. Springer, 9, 8479-8504.

Soomere, T. 2010. Rogue waves in shallow water. The European Physical Journal Special Topics, 185, 81-96.

Soomere, T., Engelbrecht, J. 2005. Extreme elevations and slopes of interacting solitons in shallow water. Wave Motion, 41, 2, 179-192.

Soomere, T., Healy, T. 2011. On the dynamics of "almost equilibrium" beaches in semi-sheltered bays along the southern coast of the Gulf of Finland. Harff, J., Björck, S., Hoth, P. (eds.) The Baltic Sea Basin. Springer, Heidelberg, Dordrecht, 257-281.

Soomere, T., Räämet, A. 2011. Spatial patterns of the wave climate in the Baltic Proper and the Gulf of Finland. Oceanologia, 53, 1-TI, 335-371.

Soomere, T., et al. 2008. Modelling of wave climate and sediment transport patterns at a tideless embayed beach, Pirita Beach, Estonia. Journal of Marine Systems, 74, S, S133-S146.

Soomere, T., et al. 2011a. Patterns of current-induced transport in the surface layer of the Gulf of Finland. Boreal Environment Research, 16 (Supplement A), 49-63.

Soomere, T., et al. 2011b. The use of Lagrangian trajectories for identification the environmentally safe fairway. Marine Pollution Bulletin, 62, 7, 1410-1420.

Soomere, T., Parnell, K. E., Didenkulova, I. 2011c. Water transport in wake waves from high-speed vessels. Journal of Marine Systems, 88, 1, 74-81.

Suursaar, Ü., et al. 2006. Cyclone Gudrun in January 2005 and modelling its hydrodynamic consequences in the Estonian coastal waters. Boreal Environment Research, 11, 2, 143-159.

Tamsalu, R., Myrberg, K. 1995. Ecosystem modelling in the Gulf of Finland.1. General features and the hydrodynamic prognostic model finest. Estuarine Coastal and Shelf Science, 41, 3, 249-273.

Weisse, R., von Storch, H. 2010. Marine Climate and Climate Change. Storms, Wind Waves and Storm Surges. Springer, Berlin, Heidelberg, New York.

OPTIMAL PRODUCT DEVELOPMENT AND MANUFACTURING ENGINEERING

R. Küttner, J. Majak, K. Karjust, M. Pohlak, M. Eerme

The planning of R&D in industrial engineering at the Institute of Machinery at Tallinn University of Technology has followed the principles of EU Manufuture Platform. The paper is organised in the following way: in chapter "Research and Industrial priorities for developing manufacturing industry" an overview of the EU Manufuture platform and main trends of modelling products and manufacturing processes is given. In chapter "Development of techniques for integrated modelling of product- process-material interactions", the development of techniques for integrated modelling of product-process-material interactions at TUT according to recent trends, is described. Finally the future trends and prospects of research are briefly discussed.

RESEARCH AND INDUSTRIAL PRIORITIES FOR DEVELOPING MANUFACTURING INDUSTRY

EU Manufuture Platform

Manufacturing industry exerts a strong technology pull on research and innovation. Many high-potential fields of research and industrial priorities have been integrated into EU Manufuture platform.

Manufuture Platform

It was launched in 2004 to provide a vision of strategic research agenda, basic R&D activities and pilot actions, to support the key issues and goals in all fields of manufacturing. Its mission was to propose, develop and implement a research and innovation strategy, capable of speeding up the rate of industrial transformation to high value-added products, processes and business models.

Today a high level group of European executives, from research organisations and industry, invited by European commission, offer their expertise and insights as a basis leading to ManuFuture vision 2020 (Jovane et al. 2009).

As a result of many strategic discussions within the European stakeholders from manufacturing industries and the related research community, it has been determined that the successful development of high value-added technology should consider the following strategic sub-domains (Jovane et al. 2009):

- Sustainable manufacturing
- ICT-enabled intelligent manufacturing
- High performance manufacturing
- Exploiting new materials through manufacturing.

Concentration on these actions has attracted industry, as well as universities and research centres.

Sustainable manufacturing

Manufacturing development Strategies of Manufuture include partial strategies of product development, engineering planning and networking of manufacturing processes. All this together is the core of enterprise's strategic development and comprises the set of models, technologies and tools for efficient engineering of products and processes.

The available collection of methods to systemise the management of manufacturing processes includes specific aspects like planning of processes and supply chains, planning of resources etc.

The supply chain planning is characterised by a decentralised decision making. In the near future, traditional supply chains will be much more reconfigurable, agile, collaborative and responsive inevitably posing new and demanding challenges on production planning tools. The main development issues and targets are collaborative planning, including strategic production planning, and optimisation of production resources in company networks as well as distributed planning models and supporting tools.

ICT-enabled intelligent manufacturing

The contribution of Information and Communication Technologies (ICT) to manufacturing aims to improve the efficiency, adaptability and sustainability of production systems and their integration within agile business models and processes in an increasingly globalised industry. Further integration of any newly developed ICT tool into the industrial environments requires complementary research and innovation efforts. These integration aspects will play a key role in generating and using smart production systems.

ICT is a key enabler for improving manufacturing systems at following levels:

- agile manufacturing and customisation involving process control, planning, simulation and optimisation technologies, and robotics;
- value creation from global networked operations (supply chains);
- better understanding and design of production and manufacturing systems for product life cycle management, involving simulation, modelling and knowledge management.

Modern ICT technologies enable the management of the factories in real time to distribute the information to all the actors in production system.

High performance manufacturing

New complex products and processes will be developed through efficient knowledge reuse for decision-making.

The goal will be to increase economic growth, efficiency and local value added. The main tasks of R&D will be the implementation of new technologies and innovations into operations and keeping enterprises on the highest level of performance.

New basic models of simulation techniques must be developed together with the continuous optimisation of manufacturing.

Exploiting new materials through manufacturing

New advanced materials (e.g. adaptive or functional materials and multi-materials) in a product offer high potential product innovations in all industrial sectors, and will require acquisition of new enabling technologies.

To improve the performance of single products it is necessary to implement more and highly complex functionalities. One way to fulfil these requirements is the usage of multiple materials and the development of new technologies to be able to realise graded material properties. The controlled integration into products of graded properties is representing one step forward into complete new design and process technologies, which enable the production of structures, part geometries and properties hitherto unknown. Another challenging research topic is multi-scale modelling of materials.

Today's challenges faced by manufacturing R&D are so complex that they can only be solved through the help of advanced techniques of mathematical modelling.

Main trends of modelling products and manufacturing processes

The main objective of modelling products and manufacturing processes is sustainable, more resource efficient, and competitive economy. In the following the main trends of modelling products and manufacturing processes are discussed based on European Mathematics and Industry Forward Look 2010 (Main Trends 2010).

There are several aspects to be taken into account. Modelling is typically an iterative procedure, because if results do not explain or fit the observations, one has to modify and adapt the model, and repeat the cycle until the model is appropriate to optimise the products and manufacturing process that it describes.

The use of mathematical techniques opens the way to analysis of the created models by numerical simulation, and the validation of the model with experimental data.

Due to the variability in the production process, confidence issues and uncertainty quantification are critical but also stochastic and statistical methods must be taken into account. In addition, it is important to investigate the robustness and sensitivity of the model.

Many real-world applications give rise to multi-scale, multi-physics problems where the behaviour of the system to be optimised depends on various physical phenomena that occur on a multitude of spatial and/or time scales.

The traditional numerical solution of optimal design is based on a 'sequential approach'. A significant reduction in computational time can be achieved by so-called Multidisciplinary Design Optimisation (MDO) methods, featuring the simultaneous solution of the state equation and realisation of the optimisation strategy. MDO has been rapidly gaining recognition as an engineering discipline for development of complex engineering systems.

Response surface methodology has received much attention in the MDO community in the last dozen years. MDO and response surface methodology are particularly suited to the design process in which analysis of different disciplines may be accomplished by different teams and for use of parallel systems for computing.

Possible techniques for development of models include proper decomposition**.** Reduction of the number of parameters while keeping the accuracy of the approximate solution can be achieved by adaptive model refinement and coarsening.

The last decade has also been marked by advances in the construction of 'Reduced-Order Models' using a variety of methods. Unlike high-fidelity models, these seek the simplest mathematical representation that captures the dominant behaviour of the system to be simulated.

Current research spans different techniques of product development and testing, from manufacturing planning to multi-stage product and manufacturing process development.

DEVELOPMENT OF TECHNIQUES FOR INTEGRATED MODELLING OF PRODUCT-PROCESS-MATERIAL INTERACTIONS

The objectives of the current research draw on needs of Estonian Industry and can be formulated as:

- Development of the generic framework for describing the integrated strategic planning for products and manufacturing process for a supply chain. (SC)
- Development and adaptation of numerical methods and algorithms for solving complex engineering design problems.
- Development and application of novel manufacturing technologies.

Development of Methods

Hierarchical approach and decomposition

To match the planning model more closely with the real situation, the engineering design and manufacturing planning is recommended to be decomposed. A natural response to the complexity is to manage the various decision steps independently, i.e., to allow each entity to use local information and to implement locally optimal polices.

Hierarchical approach raises different engineering concerns, sharing of information and coordination between tasks of different levels. This leads to the multilevel hierarchical optimisation scheme (Haftka 2003; Pohlak et al. 2010).

Hierarchical decomposition facilitates employment of the decentralised optimisation approaches that aid engineers to identify interactions among elements at lower levels and to transfer this information to higher levels, and has in fact become standard design practice.

To develop methods that are both well suited to their specific application and mutually consistent across applications, we use the following steps in developing a planning framework:

1. Divide the overall system appropriately. Different methods for different portions of the process, different product categories, different planning horizons, etc. can be used. The key is to find a set of divisions that make each piece manageable, but still allow integration.
2. Identify links between the divisions.
3. Use feedback to enforce consistency. As the system runs, we should continually use updated information and knowledge to force planning tools to use timely consistent information.

Hierarchical decomposition procedure has been for example used by the authors for design of large composite parts, including FEA, topology- and free-size optimisation in lower level and meta-modelling, global optimisation, etc. in higher levels (Pohlak et al. 2010). As an example the bathtub with optimal thickness distribution of reinforcement layer is depicted in Fig. 1.

The reinforcement problem of the bathtub shell can be formulated as a multi-objective optimisation problem and expressed in mathematical form as

$$\begin{aligned} \min F(x) &= [F_1(x), F_2(x)], \\ F(x) &= C(x_1,x_2, \dots x_n), \\ F_1(x) &= T(x_1,x_2, \dots x_n), \end{aligned} \tag{1}$$

subjected to linear constraints

$$x_i \le x_{i^*}, \dots -x_i \le -x_{i^*}, \dots i = 1, \dots , n \tag{2}$$

and nonlinear constraints

$$u(x_1,x_2, \dots x_n) \le u^*, \quad \sigma^e(x_1,x_2, \dots x_n) \le \sigma_*^e \tag{3}$$

In (1)-(3) $C(x)$ and $T(x)$ are cost of the glass-fibre-epoxy layer and manufacturing time, respectively, x is a vector of design variables, x^*_i and x_{i*} stand for upper and lower bounds of the i-th design variable, respectively. In nonlinear constraint (3) u and σ^ε stand for displacement and effective stress, respectively, u^* and σ^ε are corresponding upper limits. The linear and nonlinear constraints proceed from technological (maximum layer thickness), exploitation (displacement limit) and safety (stress limit) considerations.

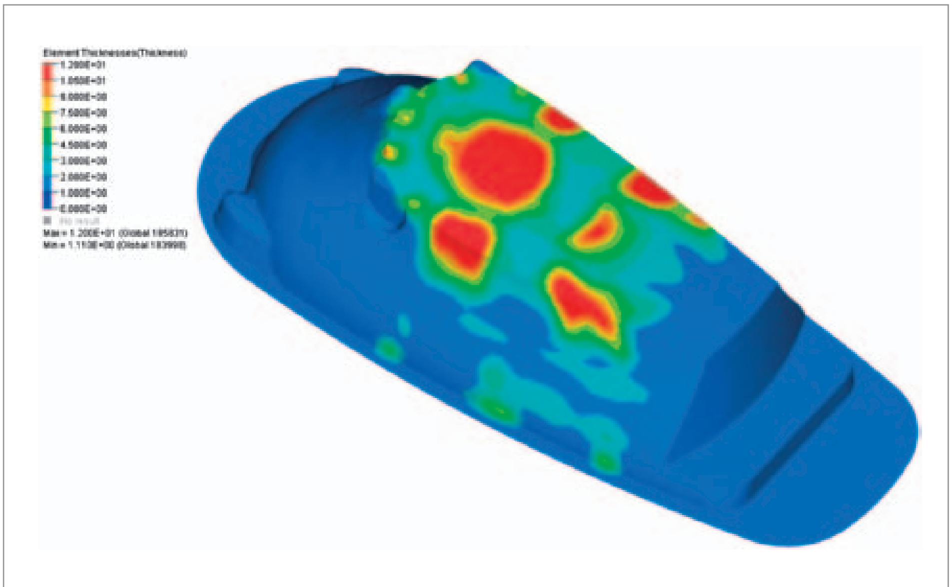


Figure 1.
The bathtub with optimal thickness distribution of reinforcement layer.

Coordinating design sub-systems

It is assumed that each planning task *Pi*, is concerned with the decision problem related to its own design parameters domain and has its own performance functions *Gi*. If there were no coordination among *Pi*, the overall optimum could not be achieved. Consequently, a coordinator has to be introduced. The task of the coordinator is to choose the suitable values of design parameters such that the planning activities on the lower level subsystems would yield a result consistent with the requirements of optimality for the overall task.

To have results consistent with the requirements of optimality for the overall task, and to coordinate and eliminate step-by-step possible discrepancies between different tasks, the coordinator provides (Küttner 2009; Kyttner et al. 2006):

- the prognosis of 'design parameters', typical and recommendable solutions, for main attributes of product family;
- specification of the constraints
- an adaptation of the objective functions for the subtasks.

Adding constraints and focus on constraints satisfaction is one of the major approaches for coordinating design sub-systems. The process of adding and processing constraints is traditional for engineering design practice, for example use of Design for Manufacturing and Design for assembly, etc. rules for product development.

To measure performance, evaluate decisions and coordinate the objective functions of subtasks, optimisation with multi-objectives is proposed as a general framework. Much effort has been devoted to linking the already commonly used CAD, CAE methods and structural optimisation techniques.

Mechanics of constitutive modelling

Constitutive modelling is the mathematical description of how materials respond to various loadings, including theories of elasticity, plasticity, creep, FEM, etc. Some studies of the authors are focused on describing stress-strain behaviour of composite and sheet metal structures. One subtopic is modelling plastic anisotropy. In order to describe the plastic anisotropy a number of new anisotropic yield criteria have been introduced by Barlat, Karafillis and Boyce and others. Plastic anisotropy is modelled by these authors using a linear transformation of the stress tensor.

The results of the project team in this subtopic are related to: development of instability criteria; material parameters identification for advanced yield criteria; formability analysis; modelling incremental forming process; optimal design of materials and structures. In (Majak et al. 2007) a new approach for modelling plastic anisotropy of sheet metals has been introduced. The instability criterion in tensor notation was derived for wide class of homogeneous yield functions as:

$$\frac{1}{\varphi}\frac{\partial\overline{\sigma}}{\partial\sigma_i}A_{ij}^{Instability}\frac{\partial\overline{\sigma}}{\partial\sigma_j}=\frac{1}{\sigma_{SH}}\frac{\partial\sigma_{SH}}{\partial\overline{\varepsilon}_P}. \tag{4}$$

In (4) orthotropic symmetry is assumed, σ and ε stand for the equivalent stress and equivalent plastic strain, respectively, σ_{SH} is a flow stress of the material, σ_i and σ_j are stress components. The instability tensor $A_{ij}^{Instability}$ is equal to Hill and Swift instability tensor on the left and right side of the forming limit diagram, respectively:

$$A_{ij}^{Instability}=\begin{Bmatrix} A_{ij}^{Hill} & for & d\varepsilon_2/d\varepsilon_1<0 \\ A_{ij}^{Swift} & for & d\varepsilon_2/d\varepsilon_1>0 \end{Bmatrix}. \tag{5}$$

Note, that the Hill and Swift instability tensors, defined by authors in (Majak et al. 2009a), are dimensionless quantities depending on the stress ratio only.

Later the non-homogeneity was incorporated by introducing the non-homogeneity factor. As result, the instability criteria proposed has some new features: the dimensionless instability tensors are introduced; the non-homogeneous yield functions are covered.

Future research in this area is related to including micro/nanostructures in constitutive models and yield criteria.

Development and adaptation of numerical methods and algorithms

Design algorithms require an optimisation strategy in a search for the best design. The choice of approach depends upon data availability and the goals of the decision-makers.

It is important to discuss in which sense one optimisation algorithm might be considered better than another. Users generally demand that a practical optimisation technique should fulfil three requirements. First, the method should find the true global minimum. Second, convergence should be fast. Third, the program should have a minimum of design parameters so that it will be easy to use. Actually, there are a huge number of techniques for achieving the desired results.

In some recent studies of the authors, the main attention is paid to solving engineering design problems, focusing on development of:

- Optimisation techniques (structural optimisation, supply chain planning),
- Discretisation methods for solving governing differential equations.

Complex engineering problems often contain complicated objective functions of which it is not possible to compute a derivative, real and integer variables, a number of local extremes, and multiple optimality criteria. In this situation the conventional approaches based on traditional gradient technique fail or perform poorly.

For that reason optimisation approach based on hybrid genetic algorithm (HGA) was developed. Genetic algorithms belong to the larger class of evolutionary algorithms, which generate solutions to optimisation problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. An overview of evolutionary computing techniques in structural engineering can be found in (Kicinger et al. 2005). Most known evolutionary computing techniques are genetic algorithms (GA), evolution strategies and evolutionary programming. These approaches differ in the types of generation-to-generation alterations and in representation of population. The standard genetic algorithms deal only with unconstrained functions, to solve the engineering design problems it is necessary to transform the initial design problem using the penalty function technique, improvement-adaptation of the GA operators, limitations in design space, etc.

The multistage optimisation procedure containing global optimisation and local optimisation (shape, size, topology, etc.) has been developed and validated on a number of practical problems: for example optimal design of composite bathtub (large composite plastics), design of a car frontal protection system, design of wind turbine tower, etc. The basic idea has been to analyse an initial design to generate data that could be used to construct approximations of the objective and constraint functions. The optimisation algorithm is then applied to minimise this approximate model subject to the approximated constraints. After solving this approximate optimisation problem, the models are updated and the process is continued until suitable convergence criteria are met.

Large number of engineering design, production planning and scheduling problems require that decisions be made in the presence of uncertainty. The random variations might arise from a variety of sources including the geometry of the problem, material properties, boundary conditions, initial conditions, and models based on incomplete knowledge etc. Optimisation under uncertainty has experienced and still experiences rapid development in both theory and applications. Once the randomness is included in the formulation of the design problem, it is not immediately clear how to formulate a well-posed optimisation problem. A number of convenient formulations of the problem are possible depending on when decisions must be taken relative to the realisation of the random variable.

In the current study the stochastic optimisation methods are applied in global optimisation algorithms, manufacturing resource planning and supply chain design.

Sensitivity analysis

The role of the sensitivity analysis is difficult to overestimate in engineering design. Optimal solutions, sensitive with respect to design variables, should be replaced with more robust, but less sensitive ones. The aim of the robust optimisation is to minimise the variation of the performance as a result of uncontrollable factors. Significant for this approach is the understanding of the rules for creating robust models.

Multidisciplinary design optimisation

Multidisciplinary design optimisation (MDO) is a concurrent engineering design tool for complex systems design. MDO allows designers to incorporate all relevant disciplines simultaneously. The optimum of the simultaneous problem is superior to the design found by optimising each discipline sequentially, since it can take into account the interactions between the disciplines.

In some recent studies of the authors, a bi-level optimisation architecture of MDO approach has been applied for example to study large composite parts. The disciplines considered in the design of large composite parts are structural analysis, product family planning, process and manufacturing planning. The tools of CAD/CAE, and global optimisation are used. The structural analysis has been employed for determining stress-strain state of the components and substructures in lower level. Next the artificial intelligence tools and methods are used for modelling. In top level the global optimisation has been performed by applying GA, and Pareto optimality concept (Fig. 1).

Discretisation

Discretisation of the differential equations describing constitutive behaviour of the structures plays key role in solution of engineering design problems. One possibility to obtain simple, time and cost effective solutions is development of algorithms and methods for particular class of problems instead of using general methods. Latter approach allows considering the specifics of the problem. In the current study the Haar wavelet based discretisation method (HWDM) has been developed for solving certain class of non-linear differential equations arising from analysis and design problems of solid structures (Majak et al. 2009b; Kers et al. 2010). Some new features of the discretisation method developed can be outlined as:

- using weak formulation in wavelet based discretisation scheme;
- decomposition of the solution to local and global term.

Suggested approach allows combining the advantages of the both, FEM and properties of the Haar wavelets.

However, the HWDM is still in development stage, the problems solved can be classified as rather simple ones. One obvious future direction is development of methods employing 2D and 3D wavelets for solving partial differential equations. Another not less important issue is validation of the HWDM results and comparison with ones available in literature. Possible upcoming research topics considered herein are related to the design optimisation of the wind turbine tower (in Fig. 2 the results of modal analysis are depicted).

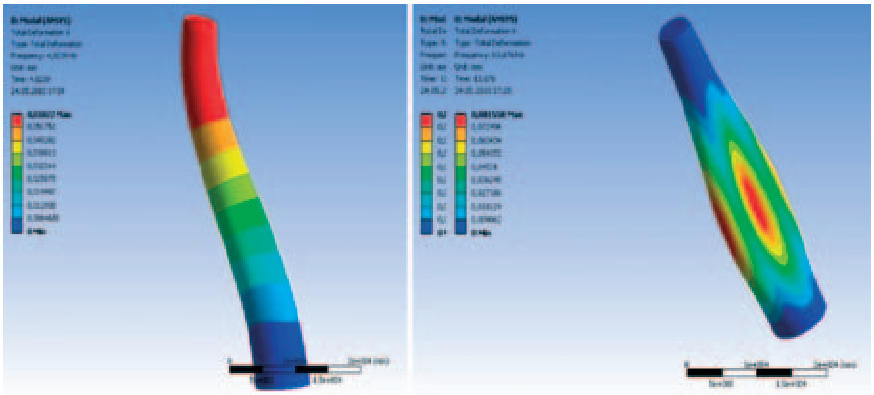


Figure 2.
The results of modal analysis of the wind turbine tower.

The optimality criteria considered are design for maximum frequency and design for minimum vibration. Further research plans involve also study and application of meshless methods (radial function based techniques, etc.).

Reduced-order models

Development of methods for construction reduced-order models is one way to simplify complex engineering design problems. A symbolic-numeric algorithm has been developed for reducing order of the non-linear algebraic system of equations. This algorithm is based on Computer Algebra Systems (CAS-es) methods and is implemented in MAPLE software package. The general idea of the algorithm is to reduce the design space of the optimisation (analysis) problem considered. The cost for this model reduction is that in general the order of algebraic equations in reduced design space is higher (depends on a particular problem). An algorithm has been applied successfully for solving the following engineering design problems (Majak, Pohlak 2010ab; Küttner 2009; Majak et al. 2009a):

- 3D optimal material orientation problems (non-linear algebraic system of equations has been reduced to one higher order equation with respect to one variable and set of linear expressions for determining remaining variables);
- material parameters identification for advanced yield criteria (non-linear algebraic system of equations has been reduced to two higher order equations with respect to two variables and a set of linear expressions for determining remaining variables).

Future research plans include development of the symbolic computation based tools for sensitivity analysis, derivation of governing equations, optimality conditions, etc.

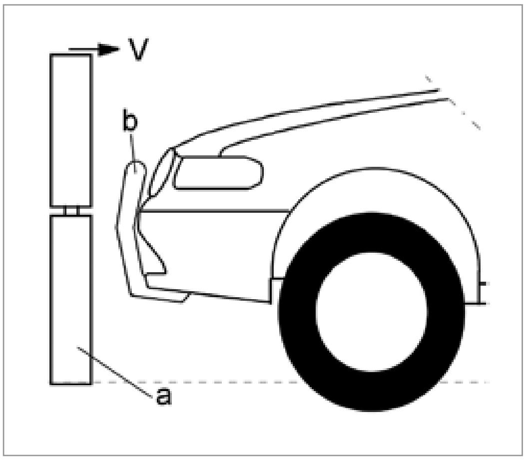
Product Development

Engineers have always pursued the aim of designing products as well as they can. In last decades, a major development is changing the approach of design engineers how to deal with their design problems. The reason is the broad use of computer-aided design, which allowed designers to quickly analyse and modify their designs. The methodology is an engineering design framework based on efficient use of computer aided design and engineering (CAD/CAE) tools. Much effort has been devoted to linking the CAD methods and structural optimisation techniques.

The methodology is illustrated with case study from industrial project of optimising car frontal protection system. A company producing frontal protection systems needed to upgrade its products to conform to European directive 2005/66/EC. The tubular extra accessories that are mounted to the front of vehicle will worsen considerably the situation for pedestrian in case of accident, so only minimum requirements can be met without adding very sophisticated systems (like airbags etc.). Minimum test, required by the directive, is the lower legform impact test. The test is depicted in Fig. 2 . In the test, the impactor (a in Fig. 3) is shot at the speed of 11.1 m/s at the frontal protection system of the vehicle. There are 3 types of sensors mounted inside the impactor: acceleration sensor, bending angle sensor and shear displacement sensor. For all three, there are certain limits defined by the directive. While with bending angle and shear displacement it is easier to fit between the limits, with acceleration limit the situation is more complicated.

To absorb the impact energy, several different design possibilities exist: several types of energy absorbing tubular, laminate and honeycomb structures can be used; materials can vary from metals to plastics and foams. In

Figure 3.
Lower legform impact testing.
a – legform impactor,
b – frontal protection system,
V – velocity of impactor.



the study, based on technological and economical reasons simple energy absorbing components made from sheet metal were investigated.

The solution procedure proposed was decomposed into following subtasks:

- design of simulation data, the set of the design variables must be prepared for a number of crash simulations,
- dynamic FE analysis (explicit FEA) – crash simulation,
- static FE analysis (implicit FEA) – structural stiffness analysis,
- application of the response surface method (RSM),
- search for optimal solution.

The simulation data are obtained by varying the design variables within certain limits. For the study different engineering analysis systems were employed, e. g. LS-Dyna, LS-Opt and Matlab.

The optimised components were tested experimentally for validation and certification.

The correct application of these engineering concepts can be easily extended into new product at its development stage. Rather than directly linking the CAD and CAE systems with optimisation tools, the basic approach involves the incorporation of models for the design parameters and the development of tools for their propagation through the response surface equations.

Exploiting new materials in products

New advanced materials in a product offer high potential product innovations and will require acquisition of new enabling technologies. Current research activities of the workgroup are closely related with modelling and application of advanced materials including orientational design of composites

(Majak, Pohlak 2010b; Küttner 2009), modelling plastic anisotropy (Pohlak et al. 2005; Majak et al. 2007), modelling new composite from recycled GFP (Kers, Majak 2008; Bendsoe, Sigmund 2004), etc.

In (Majak, Pohlak 2010b; Küttner 2009) a new methodology for determining the stationary points of the strain energy density in anisotropic solids has been developed. The methodology proposed is based on new problem formulation, derivation and analysis of optimality conditions and decomposition method. The complexity analysis of the existing optimality conditions has been performed for 3D linear elastic materials with orthotropic symmetry and the efficiency of the proposed approach has been proved (Küttner 2009).

The future research plans in the area of material design are related to free material design, functionally graded materials, nano-materials and technologies, etc. Free material design (FMD) is formulated as the problem of finding the minimal value of a certain objective function, in case when both the distribution of material as the material itself can be freely varied (Guz et al. 2007). The functionally graded material (FGM) can be considered as a special case of free material, where the material properties are fixed. The components built using FGM or FMD are analysed by employing a multi-scale approach. The computational homogenisation technique is used. The optimisation is planned by means of evolutionary and combinatorial algorithms.

In order to enhance the dynamic performance and to raise the overall (stiffness/mass) level of a FGM structures, appropriate optimisation models are planned to develop. Both, continuous and discrete distributions of the volume fractions of the composite material should be considered in the model.

Reinforcing composites with nano-particles improves their integral mechanical properties considerably. Current study is focused on study of nanomechanical models and their application to investigate the mechanical behaviour of nano-materials. One of the subtopics of the current study is modelling and design of the effective mechanical properties of micro-/nano-composites according to the needs of the products and manufacturing processes (Dong, Bhattacharyya 2010; Bhardwaj et al. 2008, Boutaleb et al. 2009; Petersson, Oksman 2006). Both, theoretical (Halpin-Tsai, Hui-Shia, Modified rule of mixture) and numerical (FEM based) micromechanical models for composites are considered.

Optimal planning of manufacturing processes and resources for a SC

Proposed integrated optimisation strategy combines engineering decisions with business decisions. The work in the engineering disciplines focuses on development of different alternatives of product family, as well as satisfying performance targets set in the production management. Each subspace of engineering design, as a local optimiser, operates on its own set of design variables with the goal of matching target values posed by the system level as well as satisfying local constraints. The design variables in this problem are comprised of variables relating for example to the geometry and structural characteristics of the product (e.g. a bathtub), and optimal thickness distribution of reinforcement layer (Fig. 1).

Business optimisation combines technological processes planning and manufacturing planning in an all-at-once optimisation approach. The business decisions result in a set of engineering performance targets, engineers seeking for performance improvements in order to yield higher profits, effective use of resources, etc.

The objective is to investigate how to analyse potential effect of implementation of novel technologies and optimise manufacturing processes for the family of products for an enterprise or network of enterprises.

Similarly to the J.F.Renaud (Gu, Renaud 2000) the two-level Collaborative Optimisation method was used.

An application of the proposed optimisation framework has been tested on the example for a family of products of hydro-spa treatment in Wellspa Inc. in Estonia.

Technological process planning

The process planning problem encompasses generation of feasible manufacturing plans, evaluation of different feasible solutions and selection of the optimal plan(s).

Modelling of the manufacturing process planning tasks is not so much generating fully specified operation sequence, while generating a set of correct and complete precedence graphs of technology operation sequences for manufacturing of the product family. The word complete refers to the generation of precedence graphs from which all possible manufacturing sequences can be derived. The word correct implies that all these sequences are feasible, i.e. they satisfy all manufacturing constraints.

The example of a generalised structure of manufacturing plan for a family of products of hydro-spa treatment is represented as an example in Fig. 4.

In Fig. 5 $Op_{1,1}$ represents reverse draw forming with two heaters; $Op_{1,2}$ represents straight vacuum forming; $Op_{2,1}$ represents automatic trimming with saws; $Op_{2,2}$ represents automatic trimming with 5-axis NC routers; $Op_{2,3}$ represents manual trimming; $Op_{3,1}$ represents manual reinforcement; $Op_{3,2}$ represents automatic reinforcement; $Op_{4,1}$ represents sub-assembling; and $Op_{5,1}$ final assembling.

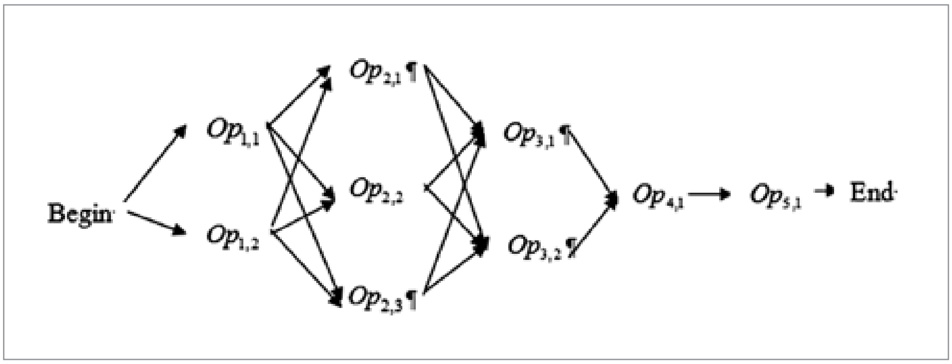


Figure 4.
The generalised structure of the manufacturing plan for product family (Küttner, Karjust 2006).

The necessity of operation is defined by logical conditions of design parameters and features of products. The technology planning model for the manufacturing of the product family is based on the maximisation of the total profit and minimisation of the manufacturing time and is subject to all constraints of operation establishment (operation necessity and operation precedence constraints), workstation time capacities, material availability, etc. For modelling and deriving systematised technological knowledge from complicated or imprecise data, derived from observations, study and experimentation, a non-linear statistical data modelling tool, an artificial neural network (ANN), was used.

Generation and selection of manufacturing (operation) plans for product family is a problem of great practical importance with many significant cost implications.

Manufacturing resource planning

Manufacturing resource planning is oriented to the effective planning of resources of a manufacturing company or a supply chain (SC).

The objective of the manufacturing resource planning tasks for a SC is to plan for given time periods $T_t = 1, ..., tl$ the volumes of products produced $X_{i,t}$, sold $S_{i,t}$ and hold as inventory $I_{i,t}$. The problem is named as multi-period aggregate planning (AP). Input for that task can be taken from the technology planning and product family planning tasks.

From theoretical point of view different variability and uncertainties are considered: such as supplier lead-time, supply quality, stochastic demands, available capacity. Planning system must be able to autonomously respond to the stochastic environment. Robustness can be thought of as an important aspect of plan quality. Model flexibility is the most common technique for providing robustness of planning. More robust plans can be preferred, but this must be weighed against other preferences.

The proposed stochastic approach is based on statistical simulation and on use of a sample of an average approximation scheme (Küttner 2009). The approach proposed is acceptable for SMEs in the make-to-order environment.

Model for integrated optimal planning of a SC

In order to understand how different SCs work, a simple, yet representative SC network $\Theta = (N, A)$ is considered (Fig. 5), where N is the set of nodes and A is the set of arcs.

The set N consists of the set of suppliers S, the set of manufacturing enterprises E and the set of customers C, i.e., $N = S \cup E \cup C$. The manufacturing enterprises E_i; $i = 1, n$ include different manufacturing centres M and assembling facilities F, i.e., $E = M \cup F$. Components are purchased by different suppliers S_j; $j = 1, m$, and it is supposed that there are different customers C_u ; $u = 1, l$.

The model of optimal manufacturing planning maximises overall profit and minimises the manufacturing time considering available resources and demand for multiple products subject to upper and lower bounds on sales and capacity constraints.

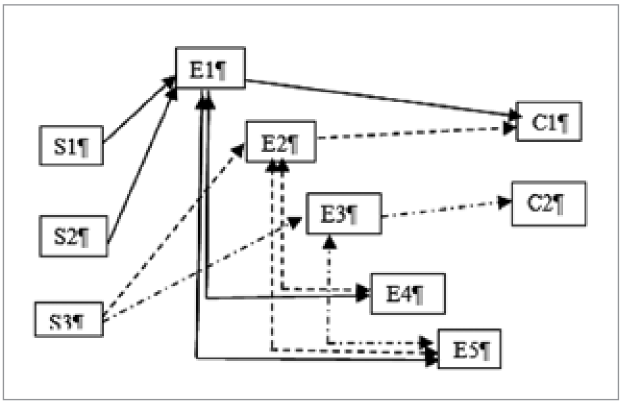


Figure 5.
Schema of a simple supply chain network.

The problem could be solved using the Pareto-optimal curve (Fig. 6) (Küttner 2009).

An approach for optimisation under uncertainty is the multistage stochastic programming, which deals with problems involving a sequence of decisions that react to uncertainties that evolve over time. At each stage, the decisions are made based on currently available information, i.e., past obser-

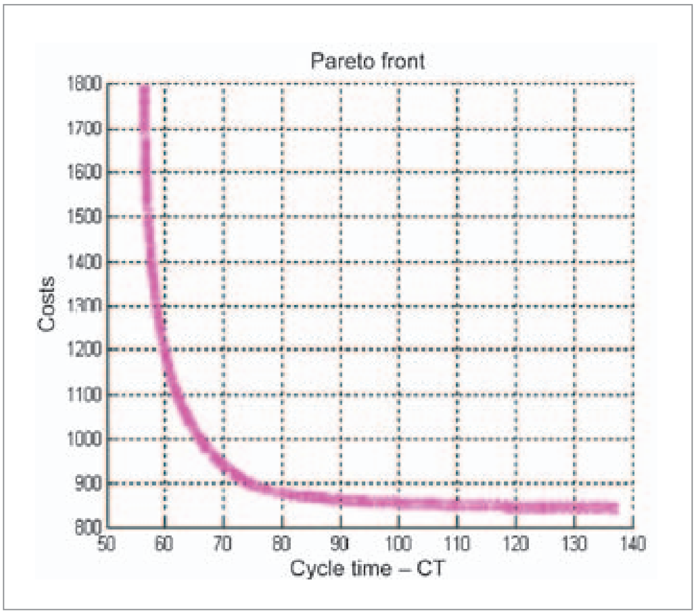


Figure 6. Example of Pareto-optimal curve for capacity costs and CT for enterprise E1.

vations and decisions prior to the realisation of future events. In our model, the two-staged stochastic optimisation approach is used (Küttner 2009).

Analytical techniques and simulation are used to estimate the influence of different sources of variability, to determine the robustness of planning decisions and to tune a supply chain.

FUTURE TRENDS OF RESEARCH

Recent trends in engineering design show the direction spreading from product-level optimality to optimality for the portfolio of products (product family) from one SME to a network of cooperating SME-s.

Current research covers modelling and simulation technologies in the fields of product family development and manufacturing.

The objective of the R&D is to investigate how to optimise the family of products and its manufacturing processes, and to take full advantage of the computer-based optimal engineering decision making process. The parametrical models of modules are planned to use for linking production planning tasks with the CAD and CAE systems.

The methods based on decomposition of global engineering design and planning problems into a number of local problems, solving the local problems in coordinated manner and then combining the local solutions to a global solution are described. An attempt is planned to estimate analytically and empirically that the proposed approach leads to improvements over the methods that consider the problems separately.

Various optimisation methods are planned to analyse to determine the optimal way of obtaining the maximum added value from their products and manufacturing processes for SMEs, including the development of product by cooperative and networking engineering, considering increasing complexity of products.

The research is focused on careful planning of the product – process – material usage with better exploitation of the material's functional potentials.

In the area of theoretical study main aims of the project team are related to the research of constitutive models of composite and sheet metal structures.

In the case of composite structures special attention is paid to development of non-linear constitutive models (Küttner 2009). Future plans are focused on taking into account the effect of micro and nanostructures. One idea is introducing neural network based constitutive models for fibre reinforced composites, nanocomposites, etc. It is planned to continue current research related to stationary conditions of the strain energy density, optimal material orientations problems of linear and non-linear elastic 3D materials (Majak, Pohlak 2010b; Küttner 2009) and topology optimisation.

In the case of sheet metal structures main attention is paid to more precise modelling of plastic anisotropy. The subtopics considered include formability analysis, development of algorithms and methods for material parameters identification for advanced yield criteria (Majak et al. 2009a), and study of micro/nanostructure based yield criteria. Specific subtopic is development of symbolic computing based algorithms for analysis and simplification of non-linear algebraic systems (reduced order model development).

In the area of applied research the future plans include solving engineering design problems arising from Estonian industry, also study of certain specific engineering problems like design of composite laminates, sandwich structures, etc.

Applied research planned involves

- ➢ development and application of novel manufacturing technologies:
    - incremental sheet metal forming,
    - additive manufacturing technologies (selective laser sintering), etc.
- ➢ design, adjusting and application of optimisation algorithms:
    - modelling of manufacturing processes,
    - global optimisation techniques, evolutionary algorithms,

- combined use of meta-modelling and optimisation techniques,
- complexity analysis and comparison of a different optimisation algorithms.

Continuation of the current research – development of hierarchical decomposition based multistage optimisation procedure is anticipated.

The novel trend is to address the production planning problems by including the uncertainty and randomness that exists in manufacturing environment in the models. In order to model and simulate at all scales of manufacturing systems, the integration between the heterogeneous models is necessary.

Future research should focus on developing techniques that are more robust and cost-efficient and the development of decentralised planning methods for the collaborative optimisation of production in non-hierarchical company networks.

REFERENCES

Bendsoe, M. P., Sigmund, O. 2004. Topology Optimization. Theory, Methods and Applications. Springer, Berlin.

Bhardwaj, S., Sharon, M., Ishihara, T. 2008. Taguchi optimization of the carbon anode for Li-ion battery from natural precursors. Current Applied Physics, 8, 71-77.

Boutaleb, S., et al. 2009. Micromechanics-based modelling of stiffness and yield stress for silica/polymer nanocomposites. International Journal of Solids and Structures, 46, 1716-1726.

Dong, Y., Bhattacharyya, D. 2010. Mapping the real micro/nanostructures for the prediction of elastic moduli of polypropylene/clay nanocomposites. Polymer, 51, 816-824.

Gu, X., Renaud, J. E. 2000. University of Notre Dame, India. Decision based collaborative optimization . 8th ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability. PMC2000-217.

Guz I. A., et al. 2007. Developing the mechanical models for nanomaterials. Composites, Part A, 38, 1234-1250.

Haftka, R. T. 2003. Combining global and local approximations. AIAA Journal, 29, 1523-1525.

Jovane, F., Westkämper, E., Williams, D. 2009. The Manufuture Road. Towards Competitive and Sustainable High-Adding-Value Manufacturing. Springer-Verlag.

Kers, J., et al. 2010. Extremes of apparent and tap densities of recovered GFRP filler materials. Composite Structures, 92, 9, 2097-2101.

Kers, J., Majak, J. 2008. Modelling a new composite from a recycled GFRP. Mechanics of Composite Materials, 44, 6, 623-632.

Kicinger, R. Arciszewski, T., De Jong, K. 2005. Evolutionary computation and structural design: a survey of the state of the art. Computers & Structures, 83, 23-24, 1943-1978.

Küttner, R. 2009. Production planning for a supply chain in a low-volume and make-to-order manufacturing environment. Estonian Journal of Engineering, 15, 1, 48-60.

Küttner, R. 2009. Production planning for a supply chain in a low-volume and make-to-order manufacturing environment. Estonian Journal of Engineering, 15, 1, 48-60.

Küttner, R., Karjust, K. 2006. Coordination of complex tasks of engineering product and manufacturing process optimization. Proceedings of the Estonian Academy of Sciences, Engineering, 12, 1, 163-175.

Kyttner, R., Karjust, K., Pohlak, M. 2006. Integrated optimal planning of product family and manufacturing technology of its components. Morgan, M. N., Jenkinson, I. D. (eds.) Advances in Manufacturing Technology – XX: 4th International Conference on manufacturing Research, ICMR2006, Liverpool, UK. Liverpool John Moores University, Liverpool, 55-60.

Main Trends of Modelling of Products and Manufacturing Processes. Mathematics and Industry Forward Look. 2010. European Science Foundation.

Majak, J., et al. 2009a. Weak formulation based haar wavelet method for solving differential equations. Applied Mathematics and Computation, 211, 2, 488-494.

Majak, J., Pohlak, M. 2010a. Decomposition method for solving optimal material orientation problems. Composite Structures, 92, 9, 1839-1845.

Majak, J., Pohlak, M. 2010b. Optimal material orientation of linear and non-linear elastic 3D anisotropic materials. Meccanica, 45, 5, 671-680.

Majak, J., Pohlak, M., Eerme, M. 2009b. Application of the Haar Wavelet based discretization technique to orthotropic plate and shell problems. Mechanics of Composite Materials, 45, 6, 631-642.

Majak, J., Pohlak, M., Küttner, R. 2007. A simple algorithm for formability analysis. Journal of Achievements in Materials and Manufacturing Engineering, 1, 57-60.

Petersson, L., Oksman, K. 2006. Biopolymer based nanocomposites: Comparing layered silicates and microcrystalline cellulose as nanoreinforcement. Composites Science and Technology, 66, 2187-2196.

Pohlak, M., et al. 2010. Multicriteria optimization of large composite parts. Composite Structures, 92, 2146-2152.

Pohlak, M., Küttner, R., Majak, J. 2005. Modelling and optimal design of sheet metal RP&M processes. Rapid Prototyping Journal, 11, 5, 304-311.

BIODIVERSITY PATTERNS AND CONSERVATION – GENERAL FRAMEWORK AND CONTRIBUTIONS FROM ESTONIA

M. Zobel, M. Pärtel, A. Lõhmus

Biodiversity refers to the number, abundance, and composition of the genotypes, populations, species, functional types, communities, and landscape units in a given system. We address the main challenges in studying biodiversity patterns and underlying processes, as well as ways how to conserve biodiversity in human impacted landscapes and ecosystems. Our emphasis is on research conducted in Estonia.

INTRODUCTION

Biodiversity, including the number, abundance, and composition of genotypes, populations, species, functional types, communities, and landscape units, strongly influences the function of ecosystems and through it the provision of ecosystem services and therefore human well-being. Biodiversity affects numerous ecosystem services, both indirectly and directly. Some ecosystem processes confer direct benefits on humanity, but many of them confer benefits primarily via indirect interactions. Although a reduction in the number of species may initially have small effects, even minor losses may reduce the capacity of ecosystems for adjustment to changing environments. In order to be able to address those problems locally and globally and to ensure the functioning of ecosystems under the increasing anthropogenic impact, it is vital to understand biodiversity patterns and processes underlying those patterns.

In a current review, we address the challenges and prospects of biodiversity studies and biological conservation. In particular, we address biodiversity patterns in local, landscape and global scales, as well as main issues related to biodiversity conservation in different scales. Although we address the general theoretical frame of biodiversity studies, our emphases are related to the research conducted in Estonia.

BIODIVERSITY PATTERNS IN LOCAL AND REGIONAL SCALES AND PROCESSES UNDERLYING PATTERNS

State-of-the-art

Diversity and composition of biological communities can be studied at different scales, starting with small scales from which it is possible to describe the occurrence and abundance of individuals and ending with regional and even global scales. The terms 'local' and 'regional' thus refer to the spatial scales at which ecological and biogeographic processes, respectively, predominate. The local community is always a subset of the pool of species existing at larger scales, but the question concerning the relationship between local and regional communities is not trivial. Regional species pool determines the potential set of species in the area, but the way how they are assembled in local communities depends on different ecological processes, notably on environmental sorting, dispersal and biological interactions. In the following we shall address biodiversity patterns in terrestrial ecosystems. Biodiversity patterns and processes underlying patterns have been addressed in aquatic ecosystems as well, but it will not be the subject of the current review.

In Estonia, there is a solid bulk of information with respect to regional and local biodiversity pattern. In particular, vascular plants and vegetation are well studied. There are thorough surveys of vegetation types and their species composition (Paal 1997). Long-term field works of botanists have resulted in compilation of Estonian flora atlas (the public version was made available by Kukk, Kull 2005). There are also a great number of different descriptive case studies of Estonian flora and vegetation. The local situation has thus been favourable to address not only the locally important questions, but also the very basic scientific questions about the relationships of local and regional diversity, and about the processes behind the observed patterns. These theoretical works started in late eighties.

Most frequently, the issues of local and regional communities are considered within the context of species richness. The relationship between local and regional species diversity is seen as a means to express the ability of ecosystems to accommodate additional species locally with increasing regional species pools. If local richness is independent of biotic interactions occurring in the local habitat and increases proportionately with regional richness we speak of 'unsaturated communities'. However, if biotic interactions limit local richness, which then saturates and becomes independent of regional richness, we alternatively speak of 'saturated communities'. The study of relationship between regional and local diversity is thus of great importance for generating the most general understanding about processes structuring assemblages of species.

In order to address the processes assembling plant communities, we may consider that local environmental conditions act as a filter, removing all species that belong to the local flora or fauna but lack the traits required to survive under local conditions. This idea was put forward by several authors more or less simultaneously (e.g. Keddy 1992; Zobel 1992, 1997; Diaz et al. 1998). Such an approach is useful because it allows addressing the relationship between large- and small-scale diversity and the composition of communities within the set of species potentially able to pass through the environmental filter. Much of the current understanding of patterns of species richness has come from the theory of island biogeography, because islands are indeed suitable model objects for studying biodiversity patterns. When island is the study object, researchers are interested in species composition across all habitats there, so in fact the study is made at the level of landscape. But if the goal is to understand the processes that generate patterns of diversity, the including of a mixture of habitats in the data pool confounds the effects of evolutionary processes, dispersal and biotic interactions. For example,

Figure 1.
Laurel forest in La Gomera (the Canaries Islands) has a large species pool attributed to the extensive pre-human area and long age of this habitat type.

it makes no sense to compare the number of species in a dry grassland plant community with that in a region that includes wet forest and seashore as well as dry grassland – confounding effect of data from various habitat types may mask true relationships. Because of that, the introduction of the idea of habitat filtering, described above, was the major step towards disentangling the processes underlying patterns.

The theoretical frame addressing diversity patterns in different spatial and temporal scales simultaneously was called the concept of species pool (Zobel 1992). In particular, addressing relationships between regional and local diversity might indicate whether or not communities are saturated only when the comparison is made within the same species pool, i.e. within the set of species that is capable of coexisting in the target community, bound to a certain set of environmental conditions that provide suitable ecological niches for the occurring species. Consequently, the delineation of the species pool should be based not only on area but also on ecological factors.

Forward look

The traditional study of biodiversity patterns in ecology seeks finding relationships between environmental gradients and community patterns, and explains the observed distribution patterns through abiotic tolerances of species, as well as through local biotic interactions. Recent decades have raised the issue of how significant is the effect of dispersal limitation in shaping local communities. There is accumulating empirical evidence addressing the role of dispersal limitation (Myers, Harms 2009) and the experimental studies made in Estonian plant communities (Zobel et al. 2000) have contributed to changes in current understanding. But the potential role of biogeographic history in shaping current within-region biodiversity patterns has been mostly overlooked.

It is well established that species diversity shows contrasting values among biogeographic regions. Explanations of diversity variation at a biogeographic scale have relied on large-scale processes (Mittelbach et al. 2007). For instance, the most thoroughly studied pattern in biogeography – the latitudinal diversity gradient – has been attributed to the 'time and area hypothesis', which maintains that tropical climates, being older and historically larger, have experienced more opportunity for diversification. Although the potential role of large-scale processes in generating local diversity patterns has been emphasised for considerable time as well, this question has garnered more theoretical interest only recently (Zobel, Pärtel 2008). In particular, the role of historical factors in generating local patterns of species richness, e.g. change of diversity along edaphic and climatic gradients within region, has

Figure 2.
Semi-natural alvar grasslands in Estonia have decreased dramatically and many populations there are persisting close to their extinction threshold. The extinction debt estimated for this site is around 40% of its current species number.

been largely ignored. We see here a huge perspective to integrate approaches in different scales and take a 'biogeographic view' when addressing local biodiversity patterns.

In particular, the 'species pool hypothesis' in ecology, launched basically through the works of Hodgson (1987) and Taylor et al. (1990) and developed by the Estonian team (Zobel 1992), is an analogous concept to the 'time and space hypothesis' in biogeography, aiming to address local diversity patterns through large-scale historical factors. This hypothesis suggests that more species are expected to occur in conditions that have been more abundant (in space or time) throughout evolutionary history. The predictions made by species pool hypothesis are habitat type specific and do not consider diversity at the landscape level, where different habitat types co-occur.

The species pool hypothesis has received indirect support from large-scale studies as referred to above. However, such a support remains indirect and needs further validation through different empirical data. One possible approach to disentangle the role of biogeographic history in generating local biodiversity patterns is to study empirical diversity-environment relationships in parallel in different regions with different history. Such an approach was pioneered by Pärtel (2002) who showed that empirical ecological relationships may vary geographically. For example, a positive relationship between plant diversity and soil pH has been documented in the temperate and boreal zones, but an opposite relationship is usually found in the tropics. Later more information became available about the geographic variation of empirical ecological relationships provided through meta-analysis (Pärtel et al. 2007).

However, such an approach – studying ecological relationships in biogeographic context – will provide ecologists a powerful tool for understanding processes underlying the currently observed biodiversity patterns. This understanding will also improve our ability to understand forces behind biodiversity decline and to manage biodiversity in a sustainable way. In particular, contrasting empirical relationships between diversity and environmental parameters among geographic regions may reflect the different evolutionary history of the respective species pools. If empirical relationships between diversity and environmental factors vary among regions, one may deduce that regional evolutionary history has resulted in different numbers of species adapted to particular environmental conditions in different regions. Since species exhibit niche conservationism – a tendency for modern species to retain ancestral traits that govern niche boundaries – relationships between diversity and environmental conditions can be viewed as products of history and evolution. For example, the different local diversity-productivity relationships in the temperate zone and tropics can be attributed to the divergent evolutionary history of species pools present in different areas (Zobel, Pärtel 2008). Productive habitats in moist lowland tropical regions have been more prevalent throughout evolutionary history, while highly productive sites in the temperate zone are less common or are fairly young. Thus, one can expect that diversification in the tropics has been greater in abundant productive habitats, and in temperate regions in abundant unproductive habitats, and the locally different diversity in productive and unproductive habitats can be attributed to different evolutionary history of respective species pools. Alternatively, if the empirical relationship is constant across regions or phylogenetic lineages, one may suspect ecological mechanisms – either interaction- or dispersal-related – to account for the pattern.

The next logical step in the study of the role of large-scale long-term processes in shaping local biodiversity pattern would be addressing locally represented species pools with different origin. Being the first to establish a direct link between evolutionary history and diversity, Harrison and Grace (2007) showed that biogeographic affiliation of plant species is related to their distribution along productivity gradients and local diversity patterns might reflect the different evolutionary history of species pools in particular habitat types. The main bottleneck of analogous studies is the lack of relevant information about the evolutionary origin of particular species and sets of species. Stepwise accumulation of phylogenetic and biogeographic information will hopefully fill these gaps in the future and make it possible for ecologists to put the current patterns into the historical context.

In specific cases the evolutionary history of local flora and fauna might be better known. For instance, one would require areas containing locally radiated endemic species, because the diversity pattern of local endemics makes it possible to evaluate the role of evolutionary history in generating local diversity. Also, the presence of clearly distinguishable habitat types strongly facilitates an understanding of diversity patterns. Oceanic islands, where the flora and fauna is the result of species accumulation through time by chance dispersal from mainland sources or neighbouring islands, and in situ speciation and extinction, constitute a suitable locale for analysis of diversity patterns, in contrast to mainland biogeographic areas, which might have had historically different degrees of connectivity. In particular, the floristic history of the Canary Islands is well documented and detailed information on the distribution of habitat types has been recently made available. In addition, these islands also allow study of the diversity patterns of ancient immigrants. Because of that, we addressed the legitimacy of the species pool hypothesis (analogous to time and area hypothesis in biogeography) in explaining current local diversity patterns, using the Canary Islands as a model system. We defined species pools for five main habitat types of seven islands and hypothesised that historical diversification (speciation minus extinction of single island endemics) is positively related both to habitat area and age (Zobel et al. 2011). We found indeed that the largest portion of the local variation in plant species diversity was attributed to the historic (pre-human) habitat area, although island age was also important. Our study supports the species pool hypothesis, demonstrating that natural local patterns of species diversity in different habitats mirror the abundance of those particular habitats in evolutionary history. These results allow us to conclude that much of the observed local variation in plant diversity can be attributed to the differing sizes of species pools evolved under particular habitat conditions, and that historic parameters are far more important determinants of local diversity than suggested by ecological theory

Because of the scarcity of the evidence of geographic variation of empirical environment-diversity relationships, as well as of biogepgraphic origin of different species pools responsible for the observed local biodiversity patterns, we emphasise the urgent need for broad-scale case studies which have been specifically designed to address these questions. First of all, we need appropriate study systems, both oceanic archipelagos, as well as regions in mainland where the biogeographic history of flora and fauna are traceable. In parallel to that we need background phylogenetic and biogeographic information about the local species. In other words, we have to go beyond the boundaries of our 'home ecosystems' and use the 'global biogeographic laboratory'. But even if the need for such studies is acknowledged, the implementation of ecological studies on broader geographical scale is financially and logistically a challenging task. It needs collaboration of scientists of different countries and resolving both gaps in knowledge of local diversity in certain areas, as well as lack of local resources and logistic support. Given scientists from different countries will find a feasible way to collaborate, the near future of biodiversity studies will provide attractive novel information of empirical ecological relationships worldwide.

Besides enlarging the geographic scope of biodiversity studies, there is also a necessity to elaborate the taxonomic scope and resolution. Current theory mostly addresses either vascular plants (the main concern of the Estonian team) or vertebrates. At the same time, even the biodiversity patterns of those taxonomic groups have rarely been analysed in accordance to each other. Biodiversity patterns of less studied taxonomic groups and patterns of cryptic organisms in particular, are mostly now known. This knowledge is, however, of great importance not only because it would enable to distinguish the role of history for organisms having fundamentally different life history traits, but would also allow an insight into the distributions and evolution of cross-taxon relationships. There has been increasing interest towards cryptic biodiversity patterns in recent years. For instance, Estonian teams have addressed the global diversity patterns of ectomycorrhizal (Tedersoo, Nara 2010) and arbuscular mycorrhizal (Öpik et al. 2010) fungi which live in symbiotic associations with plants. We expect that increasing taxonomic scope and resolution will characterise the biodiversity studies in the coming decades.

LANDSCAPES

State-of-the-art

In island biogeography, the relationship between local and regional richness has been considered in the context of island-mainland flora and fauna. Similarly, landscape ecology addresses the dependence of flora and fauna of small landscape elements either on neighbouring elements or the whole habitat matrix in the context of meta-communities. In their simplest forms, ‘flora’ and ‘fauna’ refer to a list of species or to the plant and animal taxa occurring in a defined area. Comparisons are usually made among spatial units (island, patch) and emerging relationships are strongly dependent on the diversity of habitat conditions within and among those spatial units.

According to the theory of island biogeography the equilibrium species richness in an island is defined mainly by its isolation and size. Isolation is defining the chance of immigration: islands close to the mainland are likely to get more immigrants than isolated islands. Area is defining extinction – large islands can host large populations which are less prone to extinction. Similar processes are also occurring in landscapes. Here the ‘mainland’ is often missing and isolation is rather measured based on the distance to other similar patches. By tradition, the landscape ecology is using the inverse of isolation – an index called connectivity. This can be just an inverse to distance to the nearest neighbouring patch or a more sophisticated index which is nonlinearly taking into account distance to all other patches in the region, and size of the neighbouring patches. There are many works which have documented that biodiversity is related to patch area or connectivity (Rosenzweig 1995). The positive relationship between species richness and sample area is probably the most commonly found pattern in biodiversity research. Therefore it was rather unexpected in landscape ecology when sometimes a number of species was found to be independent on the patch area (Pärtel, Zobel 1999; Krauss et al. 2004). During the last years this controversy has been clarified – modern landscape have been so dynamic that both patch areas and connectivities have changed so quickly that biodiversity is not in equilibrium yet. In order to understand biodiversity patterns in current landscapes, we need to consider their historical legacy, including past anthropogenic effects.

Patterns of modern landscapes are mostly shaped by man through different human land use: agriculture, forestry, mining and spread of urban areas. Consequently, centuries long human land use history can underlay current biodiversity patterns. A widespread wisdom is linking human influence and biodiversity negatively and there are good evidences for that. Interestingly, however, sometimes human influence can also be positively correlated with biodiversity. There are many large-scale examples where high biodiversity co-occur with high human population density (Luck 2007). Both humans and biodiversity might be influenced by a third factor (e.g. productivity) but there can also be direct causal effects. In Europe it is recognised that extensive agriculture often enhanced biodiversity by creating a diverse landscape with small fields, pastures, woodlands and other landscape elements (Billeter et al. 2008). The manifestation of mutualistic relationships between humans and biodiversity are semi-natural grasslands – representatives of the most diverse ecosystems in Europe. Semi-natural grasslands were created and maintained by human activities during millennia. There are also examples how forest biodiversity has been enhanced by moderate human-caused disturbances (Reier et al. 2005). Modern agriculture and forestry where large monocultures and intensive weed control is used, however, has usually negative effects on biodiversity since its causes extinction of a large part of biodiversity and remaining areas are often very isolated. In addition, previous agricultural activities with freely moving livestock across landscape (including meadows, forests and fields after crop was harvested) helped seed dispersal among different patches. Current agriculture and forestry has a very weak support to such seed dispersal. Thus, human activities have shaped landscape patterns throughout history and determined two main processes underlying biodiversity: immigration and extinction.

Immigration is a slow process including lot of stochasticity. However, if a landscape patch is older, it has a larger chance to 'collect' more species and genes. Oceanic islands have age in millions of years and older islands have both more naturally immigrated and locally evolved species (Zobel et al. 2011). Age of landscape patches can be measured in hundreds or thousands of years but mostly we do not know the exact age of a grassland or forest. In human-modified landscapes, however, human history can be used to detect history of landscapes. Historical maps reach back in some hundreds of years but this might still be too short time. Semi-natural grasslands occurred only in these areas where human lived, thus, it is possible to estimate age and historical extent of grasslands in some regions by using archaeological evidences of human population density. In Estonia we found that current plant diversity in semi-natural calcareous grasslands was positively correlated with the late Iron Age (c. 800–1000 years ago) density of human settlements, indicated by Late Iron Age fortresses and villages (Pärtel et al. 2007). In contrast, species richness was the highest in the intermediate current human population densities, indicative of moderate land-use intensity. The human-aided dispersal is also hypothesised to explain relatively high diversity of very isolated calcareous grasslands close to St. Petersburg, NW Russia. Closest extensive calcareous grassland regions originate from Estonia some hundreds of kilometres away. Nevertheless, the trade of domestic animals during the past centuries has probably promoted diaspore dispersal from Estonia to Russia. Tens of thousands of sheep and cattle were shepherded from Estonia to the market in St. Petersburg and these grasslands were exactly on this route. Domestic animals potentially acted as dispersal vectors. Thus, human influence can indeed enhance biodiversity if it is not too intensive, which evidently cause extinction of many species.

Extinctions occur if population has become too small (e.g. due to loss of habitat area), or habitat conditions have changed (e.g. too intensive agriculture or forestry practices). Extinctions, however, often occur with a time delay and populations living close to their extinction threshold might survive for long time periods before they go extinct (Helm et al. 2006; Vellend et al. 2006). This time delay in extinction is called the 'relaxation time' and the phenomenon that declining populations will eventually go extinct in fragmented or degraded habitats has been described as an 'extinction debt'. It would be alarming for conservation planning if extinction debts were a common phenomenon in present-day fragmented and perturbed landscapes, because many species or local populations might be on a deterministic path to extinction without any further habitat loss occurring. In landscapes with extinction debt we expect that current biodiversity is rather linked to past landscape parameters (which have been stable for a long time), not to current parameters after recent changes. We examined the response of plant species richness to habitat loss and fragmentation of Estonian calcareous grasslands (Helm et al. 2006). We had the privilege to use a large-scale vegetation maps from 1930s to determine past landscape patterns. The current number of habitat specialist species was indeed not explained by their current areas and connectivities but it was explained by their areas and connectivities 70 years ago, indicating that extinction debt is present in these grasslands. Later we participated in a joint group to explore extinction debt in semi-natural grasslands at the European scale ranging from Scandinavia to Mediterranean region (Krauss et al. 2010). Present-day species richness of plants was better explained by past than current landscape patterns, indicating an extinction debt. These results indicate that management strategies maintaining the status quo of fragmented habitats are insufficient, as time delayed extinctions will lead to further biodiversity loss in the future.

In summary, biodiversity patterns at the landscape level are very complex and human land-use legacy starting from prehistoric times is an important aspect, which could be an important contributor to the current variation in diversity.

Forward look

Currently we are just at the stage where we can detect the presence of extinction debt but there is still much work needed to elaborate methods to quantify extinction debt, predict the relaxation time and indicate which species are most threatened. Then it is possible to consider the extinction debt not only as a major problem for biodiversity conservation but as a possibility. As long as a species that is predicted to become extinct still persists, there is time for conservation measures such as habitat restoration and landscape management.

Our work from Estonian grasslands is still among very few which have quantified the magnitude of extinction debt (Helm et al. 2006). We used species–area relationships for a subset of grassland sites that had lost only a small amount of habitat to predict species richness for sites that lost a larger proportion of their original area. The extinction debt estimated for individual grasslands was around 40% of their current species number. Mostly, however, such subsets are not available. Since basic data about extinction curves are still rare, long-term monitoring schemes have a great value (Kuussaari et al. 2009).

If we know which species groups are most prone to go extinct, we can better plan conservation actions. In our pan-European study we sampled both long-living plant species and short-living butterflies (Krauss et al. 2010). If extinction debt was found for plant species at the time scale of c. 40 years, butterfly richness was related only to modern landscape patterns. Thus, butterflies at the European scale had already paid their 'debt'. Thus, at higher trophic levels where life-span is shorter, conservation measures must be applied more quickly. In a more detailed study in Estonia, however, both historic and current landscape patterns were linked to butterfly richness (Sang et al. 2010). The independent effect of past habitat area presumably indicates the presence of extinction debt in the study region. Butterfly species requiring large habitat areas were more likely to show signs of extinction debt. Species richness of other grassland butterflies (those not confined to the focal habitat type) was not associated with past or current habitat area, the pattern indicating their lower sensitivity to changes in focal habitats. The response to habitat loss in specialist butterflies appears faster than in specialist plants, studied in the same landscape earlier (Helm et al. 2006). Our ongoing works are looking for functional and phylogenetic plant groups which are most prone to extinction. Using long-term data where actual extinctions are documented is a promising way to combat extinctions in similar ecosystems where the extinction debt is still partly unpaid.

Can we indicate the threat of extinction for a specific population? According to theory, genetic diversity might react more quickly to environmental changes than species diversity and thus be a warning sign of future extinctions. We studied the behaviour of genetic diversity of a relatively common habitat specialist grass in fragmented calcareous grasslands in which the extinction debt has been previously documented (Helm et al. 2009). In contrast to species diversity, genetic diversity was indeed best described by current connectivity of grasslands. Additionally, genetic diversity was negatively related to current human population density, indicating an adverse effect of contemporary human settlements on the studied species. The faster response of genetic diversity to changed environmental conditions compared to species richness was further supported by the absence of an expected correlation between species richness and genetic diversity. We concluded that genetic diversity in fragmented communities is reacting quickly to changes in landscape structure and anthropogenic pressure. Our results confirm that species can be prone to genetic deterioration due to habitat fragmentation and negative anthropogenic impact even if the decline in species richness has been delayed by extinction debt. Thus, a decrease of population genetic diversity in fragmented communities can be taken as the first indication of future species losses.

Similar concepts as used for extinction debt theory can be applied to biodiversity restoration as well. If a deteriorated site has been restored, it takes time until suitable species inhabit it and genetic diversity is 'building up'. The amount of expected species is called colonisation credit, which has relaxation time similarly to extinction debt (Jackson, Sax 2010). At the landscape scale, however, we must consider whether these species are still present in the region (part of the dark diversity, Pärtel et al. 2011) or already regionally extinct. In the former times human activities often enhanced species dispersal but nowadays artificial introduction of suitable species might be a reasonable conservation measure. This, however, calls for a careful consideration of genetic background of introduced individuals. The fate of ecological restoration can also depend on the order of (re)immigrated species and community assembly rules. In conclusion, the more we know underlying processes behind biodiversity patterns at local, landscape and regional scale, the more successful we can be in conserving and restoring biodiversity.

CONSERVATION BIOLOGY

State-of-the-art

During the approximately 4 billion years of the history of life on Earth, there has been a gradual diversification of biota with only a few geologically short periods of its sharp loss. The causes of those losses documented for the last 0.5 billion years (so-called mass extinctions) have been mostly related to geological and climate processes on the Earth or external influences such as asteroid impacts. The current global concern over the loss of biodiversity (e.g. Millennium Ecosystem Assessment 2005; Butchart et al. 2010) is related to the fact that, apparently for the first time in Earth's history, one biotic species – humans – is causing a similar effect. That effect is related to the dramatically increased human population via resource consumption (including land conversion for production and overexploitation of natural populations), its side-effects (human-assisted spread of invasive species, pollution, etc.) and indirect consequences (such as the suspected anthropogenic climate change and extinction cascades in impoverished ecosystems). The frequency of the resulting species extinctions alone is estimated to exceed the average natural rates by 2-3 orders of magnitude.

Conservation biology is the scientific discipline, which documents and explains the ongoing human-caused biodiversity loss, predicts its consequences and analyses the possible solutions. Inevitably, this includes not only understanding general processes at large scales but also, or even foremost, focusing on specific situations at smaller scales subject to practical conservation activities. Because different conservation targets span over different scales, there is additionally a need to integrate local-level solutions to various larger-scale strategies (Groves et al. 2002). As such, conservation biology is a data rich discipline, organising the diversity of life within the diversity of natural and social contexts. In terms of the research, the Estonian contribution has been >250 conservation-related international peer-reviewed papers in the last 10 years, while from the practical side, Estonia has offered some excellent examples of implementing and analysing conservation activities. This has been possible because of its relatively low-intensity land use and, at the same time, a relatively long history of nature conservation starting from the classic works on fish conservation by Karl Ernst von Baer.

There is a broad consensus that, for practical reasons, biodiversity conservation should focus on two levels of organisation – ecosystems and species (e.g. Groves et al. 2002; Teder et al. 2007). Those levels allow complementary approaches to three land-use paradigms (biodiversity preservation in reserves; sustainable management of land and marine areas; the emergence of novel anthropogenic ecosystems) combined with species-level priorities derived from assessments of extinction risk. Traditionally, reserve network planning has been the main systematic conservation approach (Groves et al. 2002). Here, the main Estonian contribution has been in actually implementing some major gap-analyses of reserve deficiencies based on broad vegetation types and disturbance dynamics (forests: Lõhmus et al. 2004) or distributions of well-known taxa (birds: Lõhmus et al. 2001). Additionally, Estonia has been at the forefront of developing the idea of ecological networks in landscape planning.

Although reserves are crucial for preserving a part of the most sensitive taxa and ecosystems, accumulating knowledge has highlighted major limitations of relying on reserves only (Franklin, Lindenmayer 2009). These limitations refer to the understanding that reserves can cover only a small part of the surface area of the Earth: (i) Conservation targets are not necessarily concentrated to the same sites and/or reserve systems established using easily surveyable taxa or land-cover types have appeared to only poorly incorporate other taxa or values. Despite of its 18% land being protected for nature conservation, such incongruence problems are also prominent in Estonia (Vellak et al. 2010) and may critically depend on the distribution of particular resources on the landscape (Lõhmus et al. 2007). (ii) Historically developed patterns of land ownership and use by humans restrict the optimal selection of sites and even ecosystem types for protection, particularly in areas having long history of human use (such as Europe; Tuvi et al. 2011). (iii) Not all species can have viable populations in, or even inhabit, strict reserves as they appear in the planning process, notably because of the small size of the reserves and the suppression of large-scale disturbances for safety reasons (Swanson et al. 2011). This has led to compromises, such as the principle that naturally frequently disturbed habitats require less strict preservation, which is being put into practice in the Estonian forests (Lõhmus et al. 2004).

The limitations of reserves have highlighted the biodiversity value of the extensive seminatural areas managed for human production, notably grasslands and managed forests. Theoretically, the quality of managed areas surrounding strict reserves may compensate for up to twofold differences in the reserve area (Fahrig 2001) and, clearly, those areas could provide distinct early-successional habitats. During the last centuries, European seminatural grasslands have provided essential habitat for many species whose natural habitats – such as natural grasslands, maintained by abiotic conditions or wild herbivores, or wetlands – have been largely lost. However, those habitats are now threatened and being lost themselves in the processes of agricultural intensification and land abandonment, causing local and regional extinction in a wide variety of taxa (e.g. Helm et al. 2006; Rannap et al. 2007; Sang et al. 2010). Similarly, timber harvesting has profoundly impoverished forests structurally and biotically worldwide, even where the forest cover has been largely retained (Lindenmayer 2009). The last decades have provided extensive knowledge on specific key structures, conditions, and their combinations on the landscape, the maintenance of which could greatly enhance the biodiversity in such managed forests (e.g. Rosenvald, Lõhmus 2008; Remm, Lõhmus 2011). A particularly useful practical approach has been increasing the similarity between harvesting-related and natural disturbances, assuming that species and communities have evolved under the conditions of natural disturbance regimes.

Forward look

Gathering information about the global, regional and local distributions of species and relating distributions to anthropogenic drivers, as well as to networks of conservation areas, is a vital approach in conservation biology. The understanding of the impact of biodiversity drivers and thus improvement of conservation planning would greatly be enhanced when the 'invisible' biodiversity will be taken into account as well. It might be biodiversity of cryptic organisms which has until now been mostly overlooked and where we gather gradually more and more information due to the development of molecular methods. Second, we might also address the 'invisible part' of diversity of those organisms which are well studied, like vascular plants and vertebrates. It is possible to apply the concept of species pool and compile the list of potential inhabitants for particular habitat types. It will enable to address the so-called 'dark diversity' – a set of species potentially able to grow and reproduce in given conditions, but missing from actual communities for some reasons (Pärtel et al. 2011). Relating local and dark diversities enables biodiversity comparisons between regions, ecosystems and taxonomic groups, and the evaluation of the roles of local and regional processes in ecological communities. Dark diversity can also be used to counteract biodiversity loss and to estimate the restoration potential of ecosystems

Another general trend in studies addressing the global biodiversity loss is that biologists are increasingly involved in trans-disciplinary work. Major trans-disciplinary research fields include, for example, ecological economics, projecting the biodiversity effects of climate change, spatial planning of land use and developing biodiversity information systems, and understanding and enhancing public and policy acceptance of biodiversity issues. Such approaches are crucial, particularly for global issues and in regions having high human population density or developed intensive land use where biodiversity-related conflicts can be severe. The solution of such conflicts requires well-informed management and international collaboration.

Understanding the social dimension of environmental issues (including biodiversity conservation) is also leading to new approaches to the research planning. Collaborative platforms of scientists, conservation practitioners and policy makers have emerged to sort out the broad research questions that are most important to address globally or in particular societies and jurisdictions. In addition to showing the way to applied scientists, such open science-policy platforms serve the aim of educating different stakeholders about the scientific method and the findings (i.e., using existing knowledge more effectively), and of assisting funders in directing funds.

On this dramatically broadened global stage of conservation biology, Estonia hopefully continues to be a model area of reasonable size to analyse the opportunities of multiple-value land use and sustainable management. For example, new solutions are under development for ecologically sound, socially acceptable and economically viable integrative management strategies to maintain seminatural grasslands. The cultural context also matters because the biodiversity values known and acknowledged by the general public in the past are increasingly forgotten. One particularly valuable type of contribution are the success stories of practical conservation that effectively complement systematic reviews because all combinations of conditions do not exist or are otherwise unavailable in the condensed decision-making process. Taking advantage of its long history of relatively low-intensity land use, such broadly relevant Estonian cases already range over various types of ecosystems and taxon groups (e.g. Lõhmus, Lõhmus 2008; Rannap et al. 2009) and their further documentation is important.

Great scientific challenges are posed by the ecological functionality and persistence of ecosystems intensively managed for production. Here, biodiversity is included as a part of key processes or as integrative indicator of ecosystem health. Particularly when such areas become dominant in a region, there is an urgent need for quantitative biodiversity targets for applied purposes at various spatial scales. Because those ecosystems and areas can be also seen as large ecological experiments, their future development is also likely to shed light on theoretical issues of biodiversity research.

REFERENCES

Billeter, R., et al. 2008. Indicators for biodiversity in agricultural landscapes: a pan-European study. Journal of Applied Ecology, 45,141-150.

Butchart, S. H. M., et al. 2010. Global biodiversity: Indicators of recent declines. Science, 32, 1164-1168.

Diaz, S., Cabido, M., Casanoves, F. 1998. Plant functional traits and environmental filters at a regional scale. Journal of Vegetation Science, 9, 113-122.

Fahrig, L. 2001. How much habitat is enough? Biological Conservation, 100, 65-74.

Franklin, J. F., Lindenmayer, D. B. 2009. Importance of matrix habitats in maintaining biological diversity. Proceedings of the National Academy of Sciences of the United States of America, 106, 349-350.

Groves, C. R., et al. 2007. Biogeographic affinity helps explain productivity-richness relationships at regional and local scales. American Naturalist, 170, S5-S15.

Helm, A., Hanski, I., Pärtel, M. 2006. Slow response of plant species richness to habitat loss and fragmentation. Ecology Letters, 9, 72-77.

Helm, A., et al. 1987. Why do so few plant species exploit productive habitats? An investigation into cytology, plant strategies and abundance within a local flora. Functional Ecology, 1, 243-250.

Jackson, S. T., Sax, D. F. 2010. Balancing biodiversity in a changing environment: extinction debt, immigration credit and species turnover. Trends in Ecology and Evolution, 25,153-160.

Keddy, P. A. 1992. Assembly and response rules: two goals for predictive community ecology. Journal of Vegetation Science, 3, 157-165.

Krauss, J., et al. 2004. Effects of habitat area, isolation, and landscape diversity on plant species richness of calcareous grasslands. Biodiversity and Conservation, 13,1427-1439.

Krauss, J., et al. 2010. Habitat fragmentation causes immediate and time-delayed biodiversity loss at different trophic levels. Ecology Letters, 13, 597-605.

Kukk, T., Kull, T. 2005. Atlas of Estonian Flora. Institute of Agricultural and Environmental Sciences of the Estonian University of Life Sciences, Tartu.

Kuussaari, M., et al. 2009. Extinction debt: a challenge for biodiversity conservation. Trends in Ecology and Evolution, 24, 564-571.

Lindenmayer, D. B. 2009. Forest wildlife management and conservation. Annals of the New York Academy of Sciences, 1162, 284-310.

Lõhmus, A., et al. 2001. Bird species of conservation concern in the Estonian protected areas and important bird areas. Hirundo, Supplement, 4, 37-167

Lõhmus, A., et al. 2004. Loss of old-growth, and the minimum need for strictly protected forests in Estonia. Ecological Bulletin, 51, 401-411.

Lõhmus, A., Lõhmus, P. 2008. First-generation forests are not necessarily worse than managed long-term forests for lichens and bryophytes. Restoration Ecology, 16, 231-239.

Lõhmus, A., Lõhmus, P., Vellak, K. 2007. Substratum diversity explains landscape-scale covariation in the species richness of bryophytes and lichens. Biological Conservation, 135, 405-414.

Luck, G. W. 2007. A review of the relationships between human population density and biodiversity. Biological Reviews, 82, 607-645.

Millennium Ecosystem Assessment. 2005. Ecosystems and human well-being: Biodiversity synthesis. World Resources Institute, Washington, DC.

Mittelbach, G. G., et al. 2007. Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. Ecology Letters, 10, 315-331.

Myers, J. A., Harms, K. E. 2009. Seed arrival, ecological filters, and plant species richness: a meta-analysis. Ecology Letters, 12, 1250-1260.

Öpik, M., et al. 2010. The online database MaarjAM reveals global and ecosystem distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). New Phytologist, 188, 223-241.

Paal, J. 1997. Eesti taimkatte kasvukohatüüpide klassifikatsioon. KKM Info- ja Tehnokeskus, Tallinn.

Pärtel, M. 2002. Local plant diversity patterns and evolutionary history at the regional scale. Ecology, 83, 2361-2366.

Pärtel, M., et al. 2007. Grassland diversity related to the Late Iron Age human population density. Journal of Ecology, 95, 574-582.

Pärtel, M., Laanisto, L., Zobel, M. 2007. Contrasting plant productivity-diversity relationships in temperate and tropical regions: the role of evolutionary history. Ecology, 88, 1091-1097.

Pärtel, M., Szava-Kovats, R., Zobel, M. 2011. Dark diversity: shedding light on absent species. Trends in Ecology and Evolution, 26, 124-128.

Pärtel, M., Zobel, M. 1999. Small-scale plant species richness in calcareous grasslands determined by the species pool, community age and shoot density. Ecography, 22, 153-159.

Rannap, R., Lõhmus, A., Briggs, L. 2009. Restoring ponds for amphibians: a success story. Hydrobiologia, 634, 87-95.

Rannap, R., Lõhmus, A., Jakobson, K. 2007. Consequences of coastal meadow degradation: the case of the natterjack toad (Bufo calamita) in Estonia. Wetlands, 27, 390-398.

Reier, Ü., et al. 2005. Threatened herbaceous species dependent on moderate forest disturbances: a neglected target for ecosystem-based silviculture. Scandinavian Journal of Forest Research, 20, S145-S152.

Remm, J., Lõhmus, A. 2011. Tree cavities in forests – the broad distribution pattern of a keystone structure for biodiversity. Forest Ecology and Management, 262, 579-585.

Rosenvald, R., Lõhmus, A. 2008. For what, when, and where is green-tree retention better than clear-cutting? A review of the biodiversity aspects. Forest Ecology and Management, 255, 1-15.

Rosenzweig, M. L. 1995. Species Diversity in Space and Time. Cambridge University Press, Cambridge.

Sang, A., et al. 2010. Indirect evidence for an extinction debt of grassland butterflies half century after habitat loss. Biological Conservation, 143, 1405-1413.

Swanson, M. E., et al. 2011. The forgotten stage of forest succession: early-successional ecosystems on forest sites. Frontiers in Ecology and the Environment, 9, 117-125.

Taylor, D. R., Aarssen, L. W., Loehle, C. 1990. On the relationship between r/K selection and environmental carrying capacity: a new habitat templet for plant life history strategies. Oikos, 58, 239-250.

Teder, T., et al. 2007. Monitoring of biological diversity: a common-ground approach. Conservation Biology, 21, 313-317.

Tedersoo, L., Nara, K. 2010. General latitudinal gradient of biodiversity is reversed in ectomycorrhizal fungi. New Phytologist, 185, 351-354.

Tuvi, E.-L., et al. 2011. Establishment of protected areas in different ecoregions, ecosystems, and diversity hotspots under successive political systems. Biological Conservation, 144, 1726-1732.

Vellak, K., et al. 2010. Vascular plant and bryophytes species representation in the protected areas network on the national scale. Biodiversity and Conservation, 19, 1353-1364.

Vellend, M., et al. 2006. Extinction debt of forest plants persists for more than a century following habitat fragmentation. Ecology, 87, 542-548.

Zobel, M. 1992. Plant species coexistence – the role of historical, evolutionary and ecological factors. Oikos, 65, 314-320.

Zobel, M. 1997. The relative role of species pools in determining plant species richness: an alternative explanation of species coexistence? Trends in Ecology & Evolution, 12, 266-269.

Zobel, M., et al. 2000. Is small-scale species richness limited by seed availability or microsite availability? Ecology, 81, 3274-3282.

Zobel, M., et al. 2011. The formation of species pools: historical habitat abundance affects current local diversity. Global Ecology and Biogeography, 20, 251-259.

Zobel, M., Pärtel, M. 2008. What determines the relationship between plant diversity and habitat productivity? Global Ecology and Biogeography, 17, 679-684.

BIODIVERSITY INFORMATICS

U. Kõljalg, K. Abarenkov

INTRODUCTION

In this paper we define biodiversity informatics as a research field that stores, organises and analyses taxonomical, ecological and genetic data. Among other things, one of the more important fields in biodiversity informatics is the development of new scientific methods for analysis, and the development of corresponding software. Since the data is stored in databases, biodiversity informatics is actively engaged in the creation and employment of international standards required for the development of such databases. In general, biodiversity informatics guarantees the construction and functioning of e-infrastructure for biological diversity studies.

During the past twenty years, there have been two major changes within the fields that study biodiversity. First, in the early 1990s, polymerase chain reaction, or PCR, was widely adopted. In molecular biology, PCR is used for the amplification of a DNA fragment. During the reaction, one can acquire millions of copies of a single DNA fragment. This, in turn, allowed the application of new and effective methods for sequencing the amplified DNA. Soon these methods migrated from molecular biology to other scientific fields. In biodiversity studies, DNA sequencing became to be used for analysing the phylogeny of taxa. Later on, they were used for identifying species from soil, air and water samples and from the tissues of other organisms. Application of DNA sequences for identifying species was quickly adopted by microbiology, since up to that time the identification of bacteria in soil and water samples was based on relatively imprecise methods. They were followed by researchers who used DNA to identify the species of fungi from plant roots and other similar samples. DNA sequencing became increasingly simpler and cheaper, leading to an explosive growth of corresponding data.

This led to the other major change that has influenced biodiversity research: the fast development of bioinformatics, the primary purpose of which was the development and application of software required for analysing the new types of data. A separate issue was the storing of nucleotide sequences in databases that would allow free access to published data. For this purpose were created public databases, the so-called gene banks that today form the consortium International Nucleotide Sequence Database Collaboration (INSDC). The consortium comprises nucleotide sequence databases of

DDBJ (Japan), NCBI (USA) and ENA (Europe) (http://www.insdc.org/). These three databases exchange data, allowing access to all DNA sequences from any one of these databases. DNA sequences obtained through biodiversity research are usually stored in INSDC databases and in general they are associated with the name of a species or a higher taxon. For this reason, these databases have turned into important primary sources for biodiversity research, as has too the software used for DNA sequence analyses developed in bioinformatics.

BIODIVERSITY INFORMATICS VERSUS BIOINFORMATICS

Over time, bioinformatics increasingly specialised in serving the needs of molecular biology, especially the analysis of DNA sequences. In biodiversity research, however, it is necessary to associate DNA sequences with other data, such as the name of a taxon, classification, coordinates of the site of discovery, description of the habitat, etc. Such data was at least initially outside of the sphere of interest of bioinformatics. For this reason the 1990s saw the beginning of the development of new computer science solutions and databases required for analysing this kind of data. The concept 'biodiversity informatics' was most likely first used by the Canadian Biodiversity Informatics Consortium already in 1992 (http://www.bgbm.org/BioDivInf/TheTerm.htm). But the proper development of biodiversity informatics began after the publication of the OECD "Report of the Working Group on Biological Informatics" in 1999. 2004 saw the inaugural issue of the scientific journal "Biodiversity Informatics". The OECD report focused, among other things, on the basic problems of developing biodiversity informatics and proposed a plan for creating the Global Biodiversity Information Facility (GBIF), which by now has been implemented. GBIF develops a web portal for searching biodiversity information simultaneously from all databases that have joined its system (http://www.gbif.org/). As of now, 56 countries have joined the GBIF network, including Estonia. Initially, GBIF focused on specimen information, such as name of species and location, stored in scientific collections. Later on, data on observations of species was also added, and now GBIF is already interested in DNA sequence based observations as well.

Making the data stored in hundreds of different databases accessible simultaneously is connected to many severe problems. For example, a data field with a particular name may contain different data in different databases. The data field 'species name' may sometimes include the name of the genus in addition to the species epithet, etc. That is why data can be displayed in the GBIF portal only after the data fields of the so-called donor database have been coupled with those of GBIF. The best solution for data exchange is the development of standardised data fields which most database developers comply with.

STANDARDS IN BIODIVERSITY INFORMATICS

Standards for biodiversity data were developed before the creation of computer databases and thus before the appearance of biodiversity infor-ma-tics as a distinct discipline. Biodiversity Information Standards, TDWG (http://www.tdwg.org/) is probably the most important of the organisations engaged with this field. The two primary standards that TDWG develops are DarwinCore (http://www.tdwg.org/activities/darwincore) and ABCD (http://www.tdwg.org/activities/abcd/). The main focus of both standards was initially the terminology associated with scientific collections, but recently they have been supplemented with the terminology of other fields, including molecular biology. In 2005, the Genomic Standards Consortium (GSC) was founded with the goal of streamlining the creation of standards for genomics and for genome descriptions. GSC publishes the journal Standards in Genomic Sciences and includes several active workgroups. Notably, the INSDC is an active participant in the operations of this consortium, and thus new standards find their way quickly into the three major public gene banks. With respect to the development of biodiversity informatics, the increasing interest from the GSC towards biodiversity data is integral. For this purpose, the Biodiversity Working Group was created by the GSC in 2010. The object of this workgroup is to bring together molecular biologists, taxonomists, ecologists and biodiversity informaticians to work on the coherence of ontologies developed in different disciplines. The first unified version of the GSC and DarwinCore standards will be probably completed by fall 2011. The integrated use of gene, taxon and environmental data in scientific research, including the creation of databases, has been of interest to the authors of this document for nearly a decade.

BIODIVERSITY INFORMATICS IN ESTONIA

The origins of biodiversity informatics in Estonia must be sought for in the 1970s and 1980s. The present document surveys the development of this field since the publication of the OECD report in 1999, during which this concept was widely adopted. Nor does this document claim to be exhaustive; there are bound to be studies and developments by Estonian scientists in this field that the authors have overlooked.

Figure 1.
UNITE – web output of the database for the molecular identification of fungi.

In 2001, the relational database software MySQL was installed at the former Institute of Zoology and Botany of the Estonian Academy of Sciences (currently part of the Institute of Agricultural and Environmental Sciences (IAES) of the Estonian University of Life Sciences). The very next year this was used to create a database model, probably the first in Estonia and one of the first worldwide, that integrated DNA, taxa and environmental data. That same year, the network of Baltic and Nordic scientists began to use this model to develop a new database, UNITE (Fig. 1), available online since 2003 (http://unite.ut.ee). This database is used for molecular identification of fungi and has been created with the collaboration of taxonomists, ecologists and informaticians (Kõljalg et al. 2005; Abarenkov et al. 2010a). That same year one part of the scientific research on biodiversity informatics was shifted to the University of Tartu, but close collaboration with IAES continued. The scientists from both universities are the most important developers and users of the current system. During the past five or six years, the same group of scientists has participated in developing a brand new biodiversity database platform, the PlutoF cloud (Abarenkov et al. 2008, 2010b).

THE PLUTOF CLOUD

The PlutoF cloud (Fig. 2) allows the user to create and manage different biodiversity databases, from taxonomy to metagenomics (Abarenkov et al. 2010b, 2008). It is not literally a database, but rather a so-called home for databases that includes servers and standardised and unique software. In other words, PlutoF provides a thin client-based cloud database service.

The PlutoF platform consists of the relational database software MySQL running on Red Hat Linux Apache web server, and of a web-based workbench. Currently, the MySQL data model consists of more than 150 tables and it is being continually updated. The developers of the PlutoF cloud have

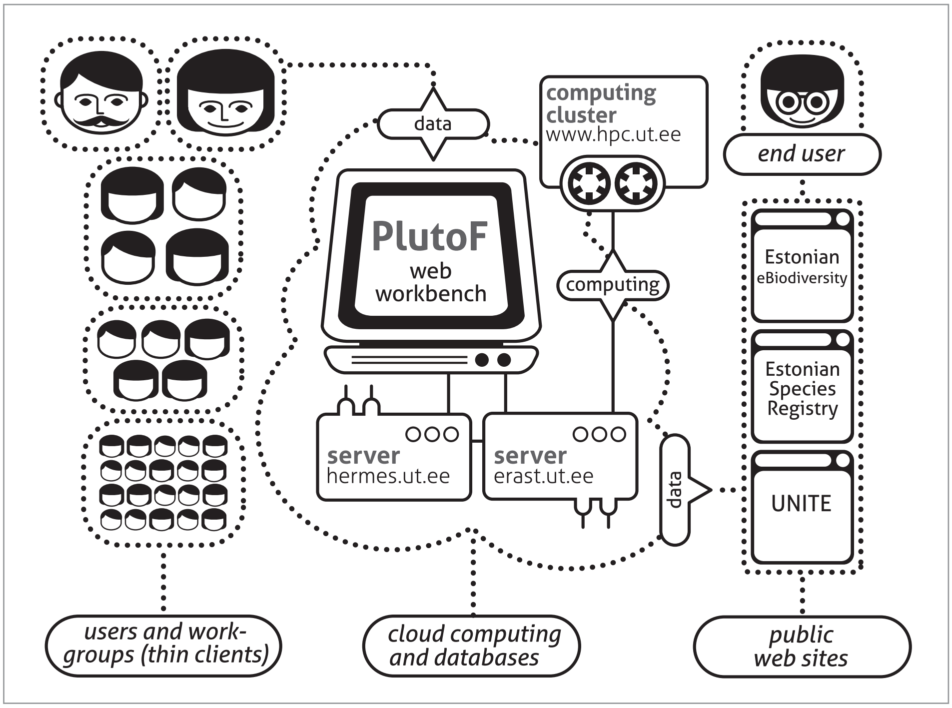


Figure 2.
A schematic drawing of the functioning of the PlutoF cloud. PlutoF provides the thin client (users, including workgroups) with cloud database services through the web browser. Every user can own several databases or be part of different workgroups with their own databases and projects. Every database/project can have its own separate web output or display data together with other databases. In cooperation with the High Performance Computing Center of the University of Tartu, PlutoF provides cloud computing services for substantial, DNA sequence based analyses. Figure by Marie Kõljalg.

always proceeded from an idea of one single database model that would allow the storage of biodiversity-related ecological, genetic, taxonomical and other information. At the same time, the developers set the goal of providing users with the option of creating an endless number of different databases, the data of which would be manageable and analysable simultaneously and in real time.

For the end user, the most prominent part of the PlutoF cloud is the web-based workbench for managing biodiversity databases (Fig. 3). This has been programmed with the following web technologies: HTML, CSS, JavaScript, AJAX, PHP, and Perl. The workbench can be used for managing both the personal and the workgroup or project databases of which the user is a member. At login, the username determines the databases and tools that the user has access to, what sort of data they are allowed to view and/or change. Thus for example only chosen experts can make changes to biological classification, only the employees of a few chosen institutions have rights to manage Estonian animal collections, the database on molecular identification of fungi can only be supplemented by the members of the international workgroup UNITE, etc. All researchers, institutions, conservationists, teachers, students

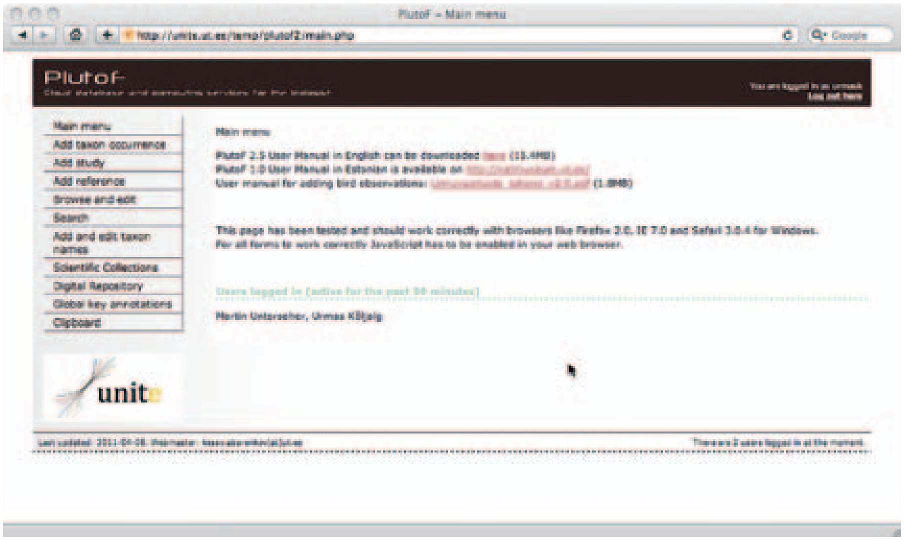


Figure 3.
PlutoF – the web-based workbench for managing biodiversity databases and cloud computing.

and so on who are interested in developing their databases on this platform can apply for usage rights for the web-based workbench.

The users can send sorted data to the workbench clipboard, from which one can commence different analyses, including displaying the data on Google Maps, carry out metagenomic analyses, etc. If the analyses exceed a certain volume, they are sent to the computer cluster of the High Performance Computing Centre of the University of Tartu. The PlutoF cloud is currently extensively used by scientists engaged with molecular databases on fungi; analyses of environmental genomics are currently only possible for fungi. Hopefully other databases on major taxa are developed for this platform that would make maximum use of the possibilities provided by PlutoF cloud.

Currently, the zoological, botanical and mycological collections of the University of Tartu, the zoological and mycological collections of the University of Life Sciences, the collection of the Tallinn Botanic Garden, owners of a few private collections, the international consortium UNITE developing molecular determination of fungi, and others have chosen the PlutoF cloud as the home of their databases. As of April 2011, the PlutoF workbench has had 308 users from around 20 countries, covering more than 30 different databases.

THE BENEFITS OF THE PLUTOF CLOUD FOR RESEARCHERS

In addition to managing databases, several public web applications have been created based on the PlutoF infrastructure for displaying data from different databases. For example, eBiodiversity (http://elurikkus.ut.ee) allows users to search for both the Estonian and English language information on species found in Estonia. As of July 2011, it includes data on 23 477 species: scientific and common names, synonyms, position of the species on the tree of life, references to distribution and ecology in Estonia, photos, maps of distribution, data on the presence of the species in scientific collection, data from the Estonian Red list and/or protection of the species on the governmental level. For example, the information displayed in Figure 4 on the distribution of fungal species *Thelephora terrestris* is drawn from scientific collections, literature, observations and DNA sequences.

This data is drawn from the following databases developed on the PlutoF platform: Registry of Estonian Species, databases of scientific collections of the University of Tartu, University of Estonian Life Sciences and Tallinn Botanic Garden, information from personal databases of researchers, DNA sequence databases UNITE and INSDC, and others. Thus it is a

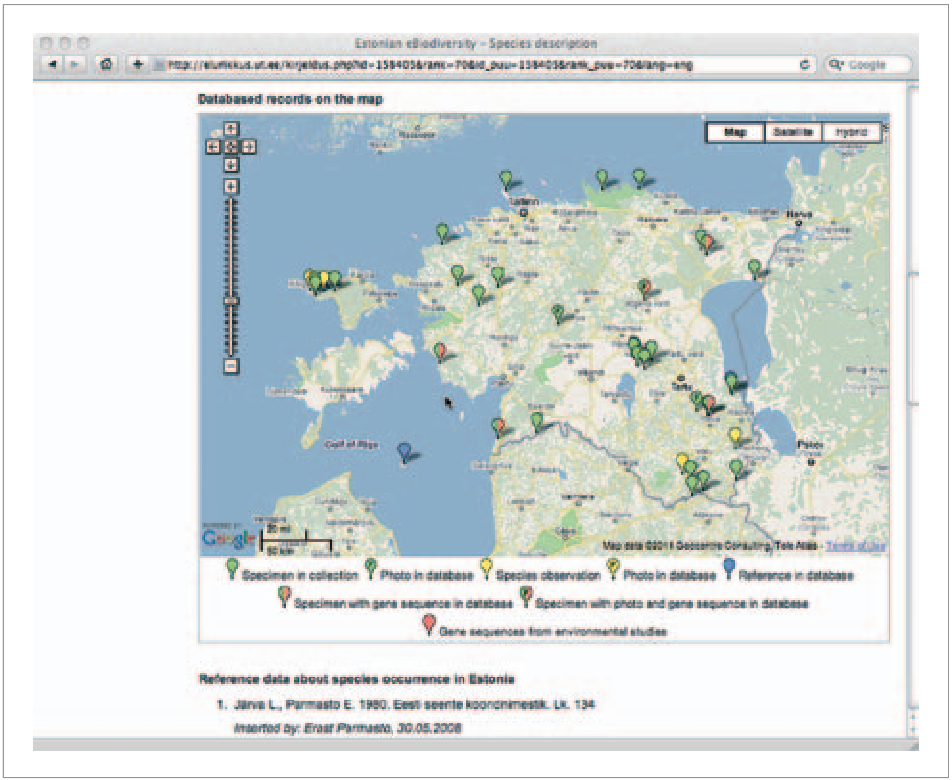


Figure 4.
Distribution map of the *Thelephora terrestris* in Estonia in the eBiodiversity output. Information about this species of fungus is drawn from scientific collections, literature, observations and DNA sequences.

web project that is based on a large number of independent databases and projects, the number of which is bound to grow over time – including the number of species included. The web-based workbench of the PlutoF cloud also allows one to display the global distribution map of *Thelephora terrestris*, based for example on rDNA ITS sequences only (Fig. 5).

By using PlutoF cloud, individual researchers or workgroups are relieved from the costs of developing and maintaining their own required e-infrastructure. Data security and backup have also been resolved centrally. In addition, it is easy to use PlutoF for making databases mutually available even before the data contained in them is published. This deepens cooperation between different workgroups and hopefully will provide an opportunity for accelerating wide-ranging analyses. For example, workgroups that are collecting similar biological samples on different continents can quickly proceed to conducting common analyses.

Figure 5.
Global distribution map of the *Thelephora terrestris* on the PlutoF web-based workbench. The map is based on rDNA ITS sequences that originate from UNITE, INSDC, and private databases.

The PlutoF analysis module is constantly updated with new tools. Currently the following tools, required for DNA sequence analysis (Fig. 6), have been developed the furthest: 1) massBLASTer blasts multiple DNA sequences against UNITE or annotated INSDC datasets using the Mega BLAST algorithm; 2) barcode tagger lets you tag your DNA sequence entries according to original sample, to enable subsequent joint analysis; 3) ITS extractor lets you extract the variable ITS1 and ITS2 sub-regions of fungal ITS sequences, getting rid of any portions (above 19 bp.) of neighbouring ribosomal genes; 4) 454 pipeline allows you to process and oversee the clustering of the 454 sequences and the subsequent, BLAST-based identification procedure; 5) BLASTCLUST automatically and systematically clusters protein or DNA sequences based on pairwise matches found using the BLAST algorithm in case of proteins or Mega BLAST algorithm for DNA; 6) Chimera checker for the fungal ITS region (Hartmann et al. 2010; Nilsson et al. 2010ab, 2011ab).

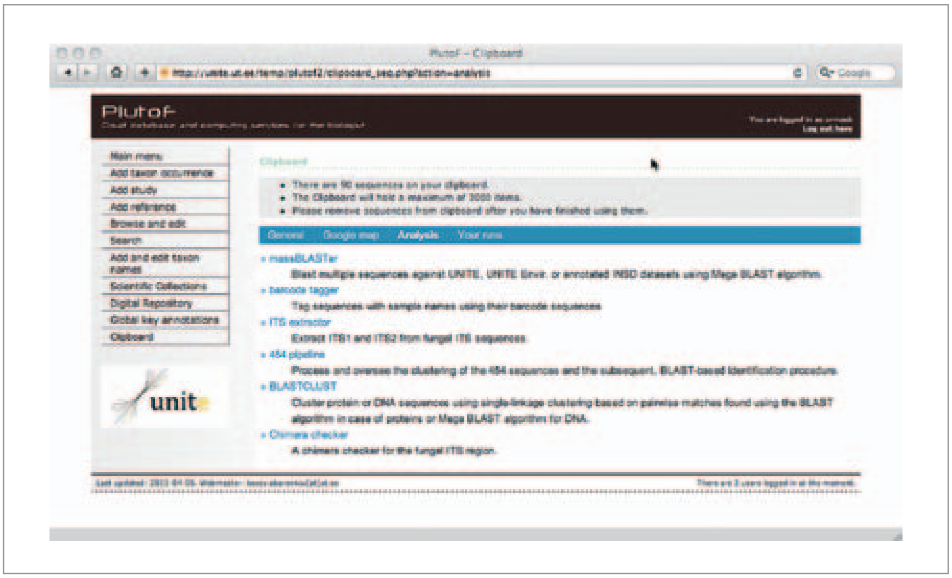


Figure 6.
The analysis module of the web-based workbench of the PlutoF cloud.

In addition to the PlutoF cloud, the software for finding chimeric DNA sequences (Nilsson et al. 2010a) has also been adopted by NCBI databases. In the future, new software for phylogenetics and statistics will be added to the workbench to increase the expediency of sending data to analysis.

PRIVACY AND ACCESSIBILITY OF DATA

An important aspect of the databases within the PlutoF cloud is the issue of privacy and accessibility. Currently, the owner of the database can lock the data for an unlimited period of time. This becomes problematic when the data are published in the form of a scientific paper or elsewhere, but the researcher forgets to make the data public. Or another issue: what to do with locked data the owner of which moves on to other activities (i.e. students) or leaves for good? Should we apply a time limit after which the data automatically becomes unlocked? How long should this period be? The openness of data is currently a hotly debated issue in science. Several overviews and papers have discussed the issue of scientific data and its release (cf. OECD 2007; Field et al. 2009). Major financers of biodiversity research, such as NERC (http://www.nerc.ac.uk/) and NSF (http://www.nsf.gov/) have separate policies for this issue. The general tendency seems to be that the financers and publishers will begin to make increased demands on how and where the data of biodiversity research projects are stored. Without it, the fund will refuse to finance the project and the publishers decline to publicise the study. Many scientific journals already make use of a system in which it is demanded that DNA sequences be deposited in a public database. PlutoF cloud will follow suggestions for publicising data that have acquired a general consensus.

In principle, the PlutoF cloud allows for an unlimited use of locked data. For example, everyone can use the locked DNA sequences of the UNITE database for identifying unknown fungal sequences. The results of the analysis will display the statistics on locked DNA sequences, but one must turn directly to the owners of the database for acquiring locked sequence. Such a practice has hopefully brought together scientists and data that would have been impossible otherwise.

Another issue related to openness and privacy is the question of recognition. How to acknowledge, that is, to cite, owners who have made their data available to others or who have updating pre-existing data? In the near future, a system will be implemented into PlutoF that allows for adding citations to data automatically.

ESTONIAN RESEARCH INFRASTRUCTURE ROADMAP NATARC

NATARC is an abbreviation of the Estonian research infrastructure roadmap "Scientific archives and data network". Within the framework of this project, the infrastructure of the PlutoF cloud will be developed in several important directions during the next four years. The goal is to create an integrated PlutoF-based biodiversity e-infrastructure that would allow researchers from Estonia and abroad to analyse in unison scientific data on organisms, descriptions of ecosystems and DNA sequences. The platform being developed should support breakthrough discoveries in different fields of the natural sciences, and guarantee the adequate documentation of the state and changes of Estonian biodiversity and conditions of the environment. The stored and ordered data will be made available through software applications to scientists in Estonia and abroad, to subordinate institutions of the Estonian Ministry of the Environment, etc.

The long-term storage of data in the PlutoF cloud is guaranteed by the development of a new e-infrastructure in the University of Tartu and the storage of copies of the database in the servers of the Estonian University of Life Sciences.

One of the major challenges for biodiversity informatics and for science in general is the accumulation of a wealth of data that is being generated in an increasing rate. On the one hand, new data are produced by scientists and their equipment. Examples of the latter are DNA sequencers, microscopes (providing pictures and related meta-data), analysers of soil, water and air samples, different monitoring systems, etc. A different, but by no means smaller part of the data is generated by calculations of data coming in from the previously mentioned and other equipment. The task of biodiversity informatics is to store this wealth of data in an organised manner and to make it available for future analysis. Without this, it would be highly difficult to move from this mass of facts and the empirical phase of science on to scientific hypotheses and concepts. On the other hand, science must be accessible and the hypotheses testable. With current massive data sets this is possible by using structured databases (Mesirov 2010). Without this, the taxpayer would pay for scientific labour the data of which would be difficult, if not impossible, to use in 5 to 10 years' time.

In addition to access to data stored in databases, their quality is also relevant. Increasingly, biodiversity-related DNA sequences and related metadata on the locality, habitat, climate, etc., is stored in the databases of the INSDC. The data being stored are organised in an increasingly ordered manner, making it relatively easy to access. More serious is the problem that part of the data is of low quality. For example, a recent study (Longo et al. 2011) found that the public genome sequence databases NCBI, Ensembl, JGI and UCSC include a large quantity of human genome sequences filed under bacteria, plants and fishes. The use of such data by other scientists, lacking quality control, would understandably lead to incorrect results. Several papers have dealt with the issue of fungal DNA sequences in the INSDC databases (Bidartondo et al. 2008; Nilsson et al. 2007, 2011b). The rDNA ITS sequences are used in environmental genomics analyses for a DNA-based identification of fungi from samples of soil, air, tissue of other organisms, etc. Since approximately 20% of fungal ITS gene sequences in the INSDC databases are defective, the indiscriminate use of this data leads to erroneous results. Yet the worst part of this is that the erroneously identified DNA sequences are also uploaded into INSDC and thus increase the amount of defective data, since as a rule they are stored in the database under the same faulty name. For this reason, the PlutoF cloud includes a separate database for annotating the fungal DNA sequences stored in the INSCD. This allows the user to add a different taxon name to a sequence that the user thinks may be defective, to note whether the sequence is chimeric or of low quality, to supplement or to add the locality, habitat, and other data. The annotated data, including the taxon name, is incorporated into the PlutoF-based analyses, allowing the defective sequences to be avoided.

Undoubtedly, the postulation and testing of scientific hypotheses based on large sets of data requires the help of biodiversity informatics. It is very common that cutting edge analyses in such fields as biogeography, coevolution, environmental genomics, systematics, phylogenetics, nature conservation, etc., are associated with large data sets. For this reason it is expedient to decide about the tools of biodiversity informatics that are to be involved already during the planning phase of the study, and to keep in mind that data and results must be usable several years from now on.

REFERENCES

Abarenkov, K., et al. 2008. PlutoF 1.0. Veebi-töölaud andmete sisestamiseks, toimetamiseks ja vaatamiseks (vaatlused, kollektsioonid, liigid jt. taksonid, DNA-andmed jm). Kasutusjuhend looduseuurijale. Tartu Ülikooli Kirjastus, Tartu.

Abarenkov, K., et al. 2010a. The UNITE database for molecular identification of fungi – recent updates and future perspectives. New Phytologist, 186, 2, 281-285.

Abarenkov, K., et al. 2010b. PlutoF – a web based workbench for ecological and taxonomic research, with an online implementation for fungal ITS sequences. Evolutionary Bioinformatics, 6, 189-196.

Bidartondo, M. I., et al. 2008. Preserving accuracy in GenBank. Science, 319, 5870, 1616.

Field, D., et al. 2009. Omics data sharing. Science, 326, 5950, 234-236.

Hartmann, M., et al. 2010. V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. Journal of Microbiological Methods, 83, 250-253.

Kõljalg, U., et al. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytologist, 166, 1063-1068.

Longo, M. S., O'Neill, M. J., O'Neill, R. J. 2011. Abundant human DNA contamination identified in Non-Primate genome databases. PLoS ONE, 6, 2, e16410.

Mesirov, J. P. 2010. Accessible reproducible research. Science, 327, 5964, 415-416.

Nilsson, R. H., et al. 2007. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PloS ONE, 1, 1, e59.

Nilsson, R. H., et al. 2010a. An open source chimera checker for the fungal ITS region. Molecular Ecology Resources, 10, 1076-1081.

Nilsson, R. H., et al. 2010b. An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. Fungal Ecology, 3, 284-287.

Nilsson, R. H., et al. 2011a. A note on the incidence of reverse complementary fungal ITS sequences in the public sequence databases and a software tool for their detection and reorientation. Mycoscience, 52, 4, 278-282.

Nilsson, R. H., et al. 2011b. Molecular identification of fungi: rationale, philosophical concerns, and the UNITE database. Open Applied Informatics Journal, 5, Suppl 1-M9, 81-86.

OECD. 1999. Report of the Working Group on Biological Informatics. (http://www.oecd.org/dataoecd/24/32/2105199.pdf)

OECD. 2007. Principles and guidelines for access to research data from public funding. (http://www.oecd.org/dataoecd/9/61/38500813.pdf)

THE ESTONIAN BIOBANK –
THE GATEWAY FOR THE STRATIFIED MEDICINE

A. Metspalu
and
L. Leitsalu, L. Milani, T. Esko, A. Allik, K. Lilienthal, M.-L. Tammesoo, T. Haller, A.-T. Kaasik, M. Nelis, M. Hass, K. Fischer, K. Krjutskov, E. Leego, H. Alavere, M. Perola

The Estonian Biobank is the population-based biobank of the Estonian Genome Center of the University of Tartu (EGCUT), a research and development institution within the University of Tartu. Concurrent to research and informational enrichment, the EGCUT is working on building a system for implementing the genomic information generated through research projects into the Estonian health care system thereby moving towards the medical model referred to as 'personalised medicine'.

INTRODUCTION

In the decade following the launch of the Human Genome Project, major developments occurred in the field of human genetics and genomics. The publishing of the draft sequence of the human genome in 2003 can be viewed as the starting point of the 'genomics era' (Guttmacher, Collins 2003). While traditionally genetics has been associated with rare hereditary (Mendelian) diseases involving single genes, the new technologies have provided an opportunity to study hereditary material on a genome-wide scale and a chance to concentrate on common complex diseases. This evolution has required a move from studying families to studying entire populations.

The need and an opportunity to study genetics on a population-wide scale was recognised by Risch and Merikangas (1996) and as a next step at the end of the 1990s the first population based biobanks were established. Among these was the prospective longitudinal population-based biobank – the Estonian Biobank of the Estonian Genome Center of the University of Tartu (EGCUT).

The idea of an Estonian biobank dates back to 1999 (Metspalu 2004). The following year, the Estonian Government decided to fund the creation of the Estonian Human Genes Research Act (HGRA). In 2001, the Estonian

Genome Project Foundation (EGPF) was initiated, with the primary objective of establishing a large-scale population-based biobank for the advancement of genetic research, collection of population based genetic and health data, and implementation of the results from genetic studies for the promotion of public health.

During the first 10 years, the EGCUT focused on recruitment of participants (gene donors) and data collection. By 2010, the anticipated 50,000 participants were recruited, and the main focus shifted towards conducting research based on the biobank samples, health information and genotyping and sequencing data. Simultaneously, the database is regularly updated and expanded by retrieving information received from other registries and databases and by re-contacting the participants. In addition, increasing efforts are being made to bring the results of scientific research into clinical practice.

Prior to the completion of the Human Genome Project the emphasis of genomics research was placed on understanding the structure of genomes (Green, Guyer 2011). After the publishing of the draft sequence of a human genome, the emphasis moved towards understanding the structure and organisation of genomes and the biological basis of disease. These were also the objectives of the EGCUT. The decade ahead is anticipated to be the strongest decade of the genomics research ever carried out in Estonia. The first aim is to conduct genomics research that is based on the information resources of the biobank and can be described as 'from bed to bench-side' research. However, there is also the second aim: to translate the results of the genomic research into practice and bring the research results back – 'from bench to bedside'. Similarly to the vision of the U.S. National Human Genome Research Institute (NHGRI 2009) advancing the science of medicine and improving the effectiveness of healthcare is also part of the main objectives of the EGCUT.

The activities conducted and planned by the EGCUT are in concordance with the Estonian Research and Development (R&D) and Innovation Strategy "Knowledge-based Estonia 2007–2013" (E. M. o. E. a. Research 2008). Two of the three key technologies (biotechnology and information and communication technologies) are either directly related or considerably overlap with the aims and objectives of the EGCUT.

The activities of the EGCUT also closely follow the three main objectives of research and development activities specified in the Estonian R&D and Innovation Strategy:

- the competitive quality and increased intensity of research and development;
- innovative entrepreneurship creating new value in the global economy;
- an innovation friendly society aimed at long-term development.

WHAT HAS BEEN DONE SO FAR?

Infrastructure

The infrastructure of the EGCUT consists of the following three main elements:

- The Estonian Biobank with comprehensive health records, DNA, plasma and white blood cell samples from over 51,000 participants.
- Technology for whole genome analyses, either genotyping, next generation sequencing, transcriptome, epigenetic and metabolome level and associated IT and statistical analysis support.
- Laboratory space (over 1000 m^2) in the new research building specifically designed and equipped for the needs of the EGCUT.

The EGCUT was among the centres chosen for the Estonian Research Infrastructures Roadmap 2011–2015 (M. o. E. a. Research 2010). In order to ensure collaboration opportunities on the national as well as the international level, the roadmap supports improving the quality of the technology and infrastructure in the Estonian Centre of Excellence in Genomics. The developments include a third generation sequencer with the necessary servers and software for large scale whole genome sequencing and data analysis, and ensuring the qualitative development of the EGCUT database to make it more user-friendly for scientists from other fields.

The EGCUT currently has 40 employees, including laboratory staff, IT development and support staff, technology specialists, and well trained researchers in the fields of human genomics, functional genomics, and biostatistics, several of whom have been recruited from abroad. Additionally, the EGCUT supports inter-disciplinary cross-faculty research teams from the medical field to molecular biology and genomics/genetics to bioinformatics and computer science. It is serving as a research base for the students, at all three levels.

The plans for the coming years include increasing the number of scientists with academic degrees and expanding the areas of specialisation. In addition to the Estonian scientists, it is a goal to attract specialists from abroad who have areas of expertise lacking among local scientists. These areas include genomics, genetic epidemiology, bioinformatics, biostatistics, human genetics and genetic counselling. Additional plans include identifying local young medical scientists to be trained at genomic medicine centres outside the country who can help to strengthen the links between biobanking and distinct medical specialties.

The aforementioned infrastructure and excellent training of scientists have made it possible for the EGCUT to join large international research consortia either within the EU FP6 and FP7 (ENGAGE, EUCLOCK, LifeSPAN) or globally (GIANT). Success of these collaborations is demonstrated by several resulting articles presenting Genome-Wide-Association Studies (GWAS) with data from tens or hundreds of thousands of individuals, published in or submitted to high profile journals.

Building the biobank

As of December 2010, the Estonian Biobank reached a cohort of 51,515 participants from over 18 years of age, with a collection of health, lifestyle and genealogy information and DNA, plasma and white blood cells from each. The population-based cohort reflects the age, sex and geographical distribution in the Estonian population (Fig. 1). The ECGUT cohort encompasses 4.7% of the adult population of Estonia (Estonia 2010). Other population databanks, such as the Californian Kaiser Permanente Study (Permanente 2011),

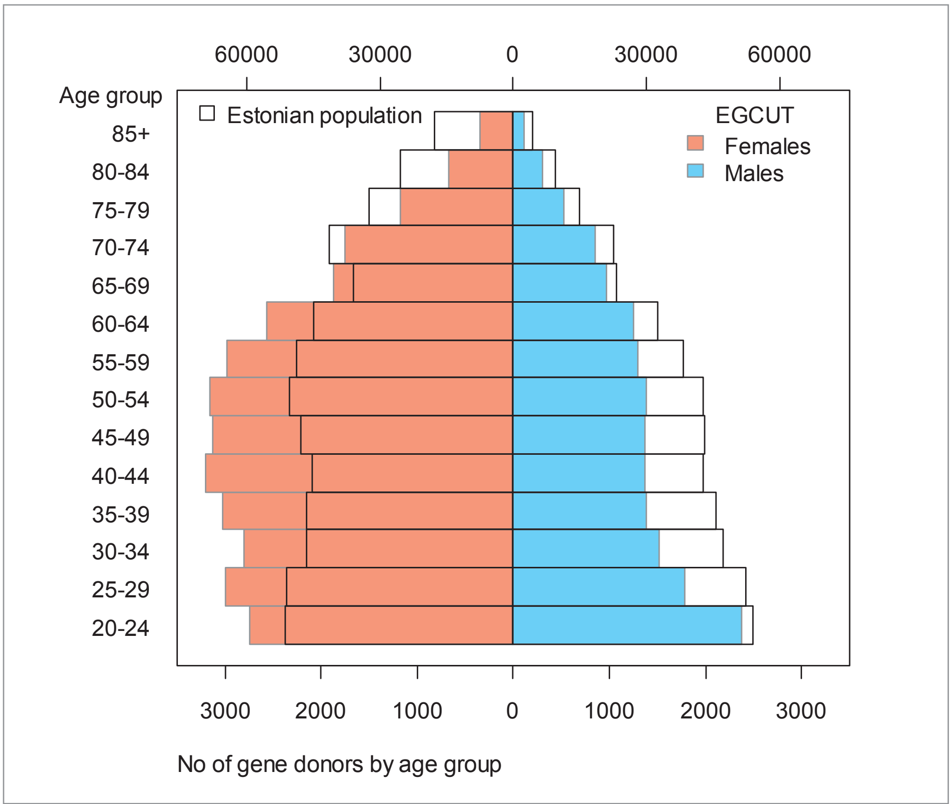


Figure 1.
The age and gender structure of the gene donors as of 01.01.2011 in comparison with the adult (20+) population of Estonia.

the UK Biobank (U. Biobank 2011), and the Swedish Biobank (S. Biobank 2003) exist that have a larger cohort. However, such a high proportion of the population is only topped by the Icelandic Genome Project (deCODE 2010). This places EGCUT internationally on a competitive scale in terms of the absolute number of participants in the cohort, and in terms of the relative number of participants from the population studied.

The recruitment of participants took place through a unique network of data collectors consisting of family physicians and other medical personnel in private practices, hospitals or in the EGCUT's patient recruitment offices. Geographically, all fifteen Estonian counties were included. When comparing the number of women and men in the EGCUT database to the Estonian population, all age groups have adequate representation.

There are several advantages of applying family physicians as recruiters for the EGCUT in terms of health data collection. The HGRA allows preexisting medical information available to be included in the phenotypic data of the participants (HGRA 2000). To differentiate between the reliability of diagnoses, criteria are used to classify them as proven (confirmed by a family physician or a medical specialist and the data collector has got the results of medical investigations performed), probable (some of the clinical features or investigation results present), or possible (diagnosis revealed by the participant, no documental proof) based on the source of information and the existing evidence. This information can be useful when choosing individuals to be included in a cohort for a research project. This system of data collection provides the EGCUT with a bulk of health data. Altogether there are currently 372,892 diagnoses among the 50,155 participants, which translate to 7.6 diagnoses on average per participant.

In addition to the quality and quantity of the data available for research, the transferability of the research conducted on the EGCUT cohort has been demonstrated in one of the first studies of the EGCUT where the genetic ancestry of Estonians was analysed in the European and world-wide context (Nelis et al. 2009). The statistical analysis revealed that there was a correlation between the genetic structure of the European populations and their geographic locations. Estonians were genetically closest to other Baltic populations (Latvians and Lithuanians), Polish people and Russians from the North-West region followed by Finns (Fig. 2). Regardless of the small population size, the Estonians do not represent a genetic isolate and are rather a heterogeneous population representative of the general European population.

This information was crucial for opening the EGCUT data to international research, as the results have implications on the transferability of the research conducted on the EGCUT cohort as a sample population. The re-

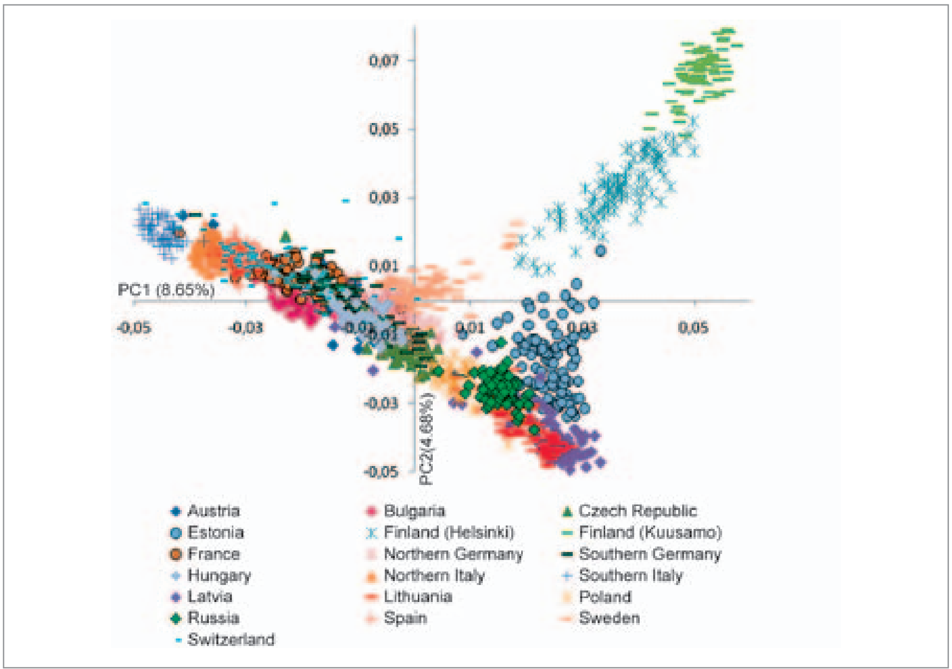


Figure 2.
Genetic map of Europe – Estonian sample representativeness relative to the general European population. The resulting genetic map is a mathematical presentation of principle component (PC) analysis comparing 270,000 genetic markers from the human genome between different European populations. The cloud that represents single individuals forms a triangular structure with Finns, the Balts (Estonians, Latvians and Lithuanians), Polish, and Western Russians, and Italians as its vertexes, and with Central and Western European ancestry samples in its centre.

sults support participation in large international GWAS consortia where the data collected are combined with the other large-scale association studies to discover the genetic variants that have a modulating effect on various common complex diseases and complex human traits.

Impact on Estonia

The developments in the Estonian health-care sector, including both public and private enterprises, are dependent on a strong biobanking infrastructure. The health-care system is moving towards individualised care, introducing an objective need for a national biobank.

As of 2007, the EGCUT has been a research and development institution within the University of Tartu. In order to promote and advance research, the EGCUT is regularly contributing to multidisciplinary research.

The activities of EGCUT support the research and studies carried out in the field of medicine and natural sciences as well as application of the unique equipment and resources of the biobank by making its resources easily accessible, not only to the research institutions in Estonia, but also abroad.

The international links established by the EGCUT activities constitute a role model for research strategies in Estonia. With its strong international collaborations (Karolinska Institutet, University of Helsinki, the Wellcome Trust Sanger Institute, the Broad Institute of MIT and Harvard, Helmholtz Zentrum München, University of Leiden, Erasmus Medical Center in Rotterdam, University of Oxford, etc.) the EGCUT provides an excellent opportunity to increase the international impact of the Estonian science and attract skilled researchers from the other universities.

Collaborations are necessary to be able to conduct the large-scale studies investigating the genetics of common complex disease. Collaborations are also a key for achieving the goals of the EGCUT of implementing the knowledge gained from the genetic research for the improvement of public health in Estonia.

The Estonian Biobank is readily providing samples for research purposes to other institutions, even in cases where the EGCUT is not directly involved in the research conducted (Fig. 3). Such data and sample sharing is not

Figure 3.
Sample preservation at
the Estonian Biobank.

only advancing research in general, but is also directly useful for the EGCUT. The EGCUT has a policy for data sharing, which states that for EGCUT to share its resources the applicant for data release is to return the research results obtained once the project is completed (EGCUT 2011; HGRA 2000). These research results are then added to the EGCUT database, thereby expanding the database. Up to the end of 2010 more than 52,000 samples were released from the Estonian Biobank for research including 8 institutions from Europe and North America.

The developments of the EGCUT also introduce new business opportunities for spin-off companies from the University of Tartu. One example is the Competence Centre on Reproductive Medicine and Biology (Repro-TAK), for which the Estonian Biobank was a strong and essential partner from the beginning, which generated over 64,000 EUR of revenue for the EGCUT during the first year.

National and international collaborations

The EGCUT is part of the national Centre of Excellence in Genomics, together with three other departments (Biotechnology, Bioinformatics, and Evolution) of the University of Tartu and the Estonian Biocentre, and as of July 2011 the EGCUT will coordinate the Centre for Translational Genomics, funded by the University of Tartu. The centre will perform the comprehensive analysis in order to follow the flow and changes of the genetic information from the variations of the DNA over the transcription and epigenetic changes to the metabolomic outcome for one thousand participants from the EGCUT cohort. This cohort will be genotyped and sequenced at the genome level, phenotyped deeply and analysed at the transcriptome, epigenetic and metabolome levels. To create this database and demonstrate its usefulness, the strongest groups from the three Centres of Excellence of the University of Tartu will make a joint effort. This is a unique opportunity and could place the University of Tartu among the leading European research centres on human genomics.

In addition to the internal projects, the EGCUT is engaged in numerous collaborations funded by various European Commission FP7 grants and Estonian funding agencies. For instance, the EGCUT was awarded a large grant from the EU FP7-REGPOT instrument. The aim of this FP project is to make the resources of the EGCUT more available to the European Research Area by upgrading the research infrastructure, promoting further integration into pan-European R&D networking, and improving research capacity by making the EGCUT more attractive as a research partner. Additionally, there have been smaller individual grants from the Estonian Science Foundation, Estonian Research Council and large FP7 collaboration projects. The EGCUT is also relatively well funded from infrastructure projects and is equipped with state-of-the-art genotyping and sequencing instruments, computing power, and computer storage space.

The EGCUT is one of the founding members and current participants of the Public Population Project in Genomics consortium (P3G 2011), a partner in the European Strategy Forum on Research Infrastructures (ESFRI) Roadmap project as well as in the Biobanking and Biomolecular Resources Research Infrastructure, where the EGCUT is the co-leader (with the University of Helsinki) of the Work Package 2 (Population-based biobanks) (BBMRI 2011). The active role played in the BBMRI process ensures the status of an 'expert centre' and a hub within the emerging European biobanking infrastructure. This could strengthen the role of Tartu as one of the 'expert centres' in Europe. Thus, EGCUT is well integrated into European activities, and this collaborative work will further support its international standing. Recently, the EGCUT was incorporated into the Nordic Biobanking Consortium, which opened the information, standard operating procedures, collaborations and funding for the EGCUT on the equal basis of the Nordic countries.

Now that the recruitment phase is completed, the number of projects is continually growing. There are currently over 100 international projects where samples from the Estonian Biobank are used and in which the EGCUT genotyping team and statistical/bioinformatics analysis group are currently involved. The majority of the collaborators are from other universities with their own biobank collections, or technology providers such as BGI-Shenzhen from China, The Wellcome Trust Sanger Institute or the Broad Institute. So far, more than 30 peer-reviewed articles have been published, several in top-ranking journals and additional manuscripts are in different stages of publishing.

The EGCUT is collaborating with many other genome centres in Europe and the USA, investigating tens to hundreds of thousands of patients in each study in order to reveal the genetic component of complex phenotypes. The EGCUT is a partner in many phenotype-driven and genetic epidemiology focused consortia such as GIANT (anthropometrical phenotypes), ReproGen (female health), LifeSPAN (healthy aging), NEOGWAS (personality), ENGAGE (genetic epidemiology) and EUClock (circadian clocks).

Genotype and phenotype data from the EGCUT have actively been used for research since 2008 when the information collection and participant recruitment was still ongoing. The first cohort included 1090 samples. It was a small model of the Estonian population where the participants were selected based on their place of birth and gender. Since then the sample-set with genome wide genotype data (GWAS) has steadily increased to 20,000. In addition the first 100 full genomes are currently under sequencing and this number will increase to 1000 which would enable to predict most of the Estonian's genomic variants above using the genotyping and sequencing data of the EGCUT.

Through the research studies conducted, 5,300 participants have been genotyped using the Illumina CNV370 or OmniExpress beadchips. Two-thirds of these have been selected as random population-based controls. These 5,300 subjects have been included in several multi-centre genome-wide association studies. Genotyping of the next 7,000 samples is in progress. About 5,700 samples are characterised with a custom-designed genotyping arrays (CardioMetaboChip and ImmunoChip) to discover the causative genetic variance and to assess the impact of GWAS findings for health outcome. For 1100 of the subjects with genome-wide genotype data, RNA has been extracted from venous blood for gene expression analysis and measurement of 40 clinical chemistry/biochemistry markers from serum and plasma performed at the University of Tartu hospital laboratory. The analysis of metabolites from the same individuals with both NMR and M/S methods is in progress.

The focus of the research at the EGCUT is currently on anthropometric traits, physiological conditions, health behaviour psychological traits, and pharmacogenomics. The set of anthropometrical traits include height, weight, body mass index, and waist and hip circumference. These phenotypes have been intensively analysed within the framework of the GIANT consortium, which now includes more than 90 institutions worldwide and has contributed to the analysis of more than 300,000 individuals of whom 10,000 are from the EGCUT. This effort has identified more than 250 genomic loci and shed insight on several biological pathways.

Public relations

Since 1999, considerable interest has been expressed in the EGCUT by the Estonian and foreign media. Over the last decade, the international media published over 120 articles on the EGCUT. Estonian media coverage was even greater – in 2010, the EGCUT was featured in various Estonian media channels 162 times. There has been a lively and constructive debate going on in Estonia on the pros and cons of biobanking activities. However, unlike a few other biobanks, the legal and societal issues of biobanking are well covered in Estonia, and the attitude of the Estonian media towards the EGCUT is generally positive.

The EGCUT popularises the field of genetics, promoting the field of science overall by informing the public of new technologies and medical achievements via national and international training programs, conferences, information days, and the Internet. The goal is to raise the standards of science in Estonia in the fields of genomics, human genetics, genetic epidemiology, biobanking, biostatistics and related specialties. Public opinion regarding the EGCUT activities has been periodically monitored by TNS Emor, one of the leading polling agencies in the country.

Annual surveys have been carried out since the establishment of the EGCUT to investigate the levels of awareness and the opinion of the Estonian population regarding the EGCUT (Fig. 4).

According to the last poll conducted in 2011, 85% of Estonians and 65% of the population as a whole stated that they were aware of the activities of the EGCUT, and from those 59% strongly supported the project. The largest increase was observed in the number of people who were 'strongly supportive of the EGCUT'. This change is accompanied by a decline in the number of people who had a 'wait and see' attitude. The number of people who were against the idea of the EGCUT has always kept below 4%.

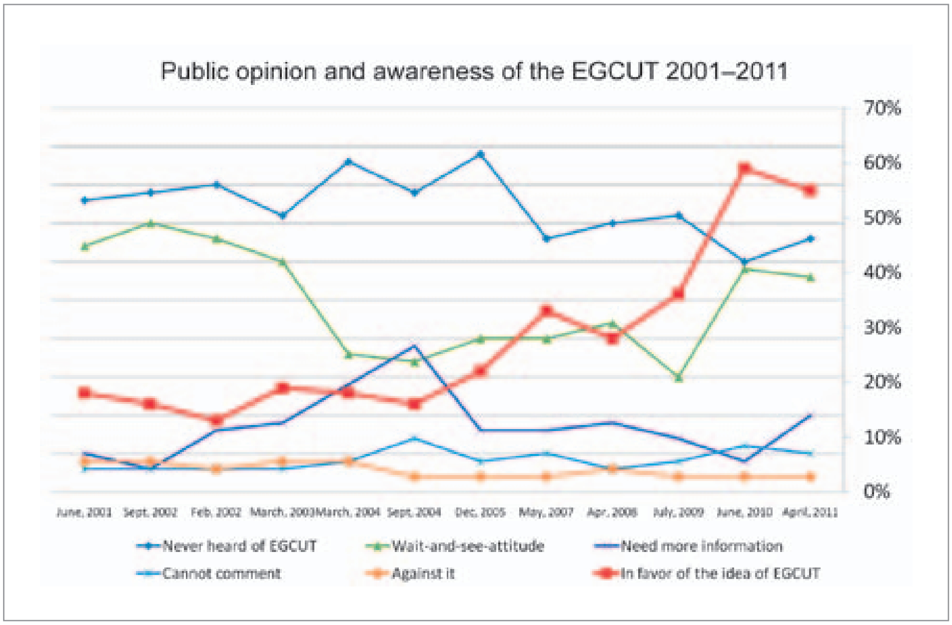


Figure 4.
Public awareness of the EGCUT and opinion on the EGCUT among the respondents who were aware of the EGCUT. Opinions are in 5 groups (in favour of the idea of the EGCUT; wait-and-see attitude; need more information; cannot comment; and against it). 1,000 individuals with ages 15–74 were surveyed by the polling agency TNS Emor.

One of the areas that need attention in public relations involves public perception of the biobanking effort. It is less focused on the basic research questions and rather on risk prediction and improved healthcare (Ormond et al. 2009). This perception leads to high expectations from the public, with increasing potential for frustration. Translating the complex information available into practice is a time consuming process and does not just depends on the activities of the EGCUT but rather on the united efforts of researchers world-wide. To avoid growing frustration from participants about the lack for personal gains from the project a proactive strategy is necessary. The EGCUT has conducted two surveys – one investigating the perspectives of family physicians regarding the implementation of genomic information into health care (Leitsalu 2010), and the second investigating the public's perspectives regarding personalised medicine.

The overall success of the personalised medicine concept constitutes an 'external' threat to the project, because the implementation of personal medicine will depend on the development of the e-health structures and is dependent on collaboration with major hospitals and the health policy of the country. However, with major international players also banking on this concept, this risk is seen as a small one.

FUTURE PERSPECTIVES OF THE FIELD

The phase of constructing a biobank of sufficient size is now considered complete, and the emphasis has shifted from recruitment of participants to other areas. These include participation in projects conducting research to discover the underlying biology of common complex diseases; informational enrichment: receiving the genomic information from genotyping whole genomes, exomes, and transcriptomes and re-sequencing genomes; as well as enriching the phenotypic data via re-contacting of gene donors and through linking to other registries.

Concurrent to research and informational enrichment, the EGCUT is working on building a system for implementing the information generated into the Estonian health care system, thereby moving towards the medical model referred to as 'personalised medicine'.

The concept of 'personalised medicine' refers to a medical model where patients are stratified into several groups and the optimal drug or treatment will be specific for that group, not to an individual itself as is commonly misconceived. Similarly to the department store where the clothing is grouped into several sizes – S, M, L, XL and XXL and 'one size fits all' is not working, it has not been working well in the pharmacy either. Genomics will help here to find 'a right patient for the existing drug', or *vice versa* 'the right drug for the patient in need'. In many cases it is working already today, but on a large scale more information is needed on the gene-disease-environment-lifestyle interactions, and regulatory agencies have to work out how to stimulate the reimbursement mechanisms in order to motivate drug companies to move towards stratification of patients.

Success in personalised medicine vitally depends on the integration and interaction of different levels of information flows (genetic variance, enzyme activity and metabolite levels in the context of lifestyle, health behaviour, individuals, and familial anamnesis) and on the ability to draw high quality conclusions from this massive, complex set of information. This is a complex process and requires multidisciplinary involvement from the medical, legal, and research fields.

The genomic era has been driven by technological advances such as the development of high-density array technology. These advances have led to the discovery of more than 4,000 gene loci that are associated with complex diseases and traits. The technological advances such as Next Generation Sequencing (NGS) technology, nuclear magnetic resonance spectroscopy (NMR), and mass spectrometry can be applied on the genome, transcriptome, epigenome, proteome or metabolome, while the comprehensive, longitudinal data on the patients' health status in biobanks have brought us closer to the realisation of personalised medicine.

The goal of the EGCUT is to develop a comprehensive database for basic research and to implement a genomic approach for medical applications. The planned database could be useful for a wide range of applications, such as a comparative tool for other populations, as a starting point for hypothesis-driven research for *in silico* hypothesis testing, or for browsing of genomic variations and associations. The EGCUT and the Estonian Biobank (its legal and ethical landscape, study design, and funding model) are frequently used as an example for several other countries in the P3G and BBMRI projects. Hence the planned database has the potential to serve as a model database for personal, preventive, predictive, and participative medical applications for other countries.

There is a shift from reactive medicine to proactive medicine – it is the next level of decision prompting to move from descriptive personalised information to predictive and preventive personalised medicine where patients have to participate actively. This personal, preventive, predictive, and participative medicine or '4 P Medicine' was introduced by Leroy Hood, a visionary of genomic medicine (Bousquet et al. 2011).

There are several factors that place Estonia in a unique position, both in terms of the tools available for genomic research and in terms of the preparedness for implementing the use of genomics in medicine. In addition to the relatively large cohort size, these unique factors include the law regulating the EGCUT, the technical infrastructure in Estonia, and the network of participants and recruiters involved with the EGCUT.

The legal framework of the EGCUT

A unique piece of legislation, the Estonian HGRA (HGRA 2000), is regulating the EGCUT and its biobank, the Estonian Biobank. The act was established in 1999, a year before the EGCUT was founded. The HGRA is the main law for regulating the activity of the EGCUT, with the goal of regulating the maintenance of the EGCUT and its data collection and processing.

The HGRA protects participants from misuse of their genetic data and from discrimination based on their health and genetic information, and vests in the government the responsibility of maintaining the Estonian Biobank.

This legislation also foresees the possibility to expand the database through re-contacting the participants for further investigations and through collecting health information from the other databases. The first re-contacting of participants was carried out successfully between 2008 and 2010. The response rate of gene donors was 57.2% (Alavere 2011). The second wave of re-contacting for follow-up purposes will start in 2011. The HGRA also permits the enrichment of phenotypic data by up-dating the health data from medical records available in the Estonian Health Information System and the other national health databases and registries. This has successfully been done with the Estonian Population Register, the Estonian Causes of Death Registry, and the Estonian Cancer Registry. This type of design is predicted to be of major importance for future investigations. The prospective nature of the biobank allows longitudinal and interventional studies, insuring the continued usefulness of the cohort for research purposes and its long term sustainability.

It is planned that results from the research findings generated by the EGCUT will be translated into clinical practice. According to the HGRA, the gene donors have the right to obtain feedback on their genetic information, and the feedback should be accompanied by medical counselling, which most likely would occur through primary care practitioners. When predictive genetic testing results are provided through health care providers the results can be translated and communicated to the patient, together with an appropriate follow-up plan. Since providing the accompanying counselling along with genetic information is required by law, the state is to support the necessary preparation towards implementing this in practice.

Nationwide IT infrastructure

There are several developments in Estonia that have provided the necessary technical prerequisites for genomic information to be implemented into health-care. A nationwide technical infrastructure, the X-road (cross-road), is already established and maintained by the state. The X-road provides a platform for secure data processing and electronic information exchange between all Estonian public sector databases, including the Estonian National Health Information System (ENHIS). Other examples include ID-card authentication and digital signature. These developments provide the basis for implementing electronic health records that eventually will all be available for healthcare providers and patients *via* X-road.

Although the gradual development of the ENHIS is planned to continue until 2013, it has been obligatory for health care service providers to forward medical data into the ENHIS since 2008 (The Health Services Organization 2007). In the future the EGCUT database will be integrated with ENHIS, creating a bi-directional information flow where the EGCUT can regularly update the health information of its cohort using information available in ENHIS, while the information generated from research conducted by the EGCUT will be integrated with the ENHIS database and can then be implemented into clinical care. In this way genetic risk predictions can be taken into account together with all other medically relevant information and patient history, and not in isolation.

Linking the genome centre to the medical field

For the success of future plans close collaboration between the EGCUT and the medical disciplines is vital. This is to avoid a gap between the translational efforts of the fields of genetics and genomics and the clinical research. Clinical research is a necessary step as genomics research is moving from bench to bedside.

Collaboration between the EGCUT and the medical field will deepen as the EGCUT coordinates the Centre for Translational Genomics and has the Medical Faculty of the University of Tartu as a partner. Future plans include developing training courses for medical students, interns and physicians to support the adaptation to genomic databases. The database can be used in the training of medical specialists – from primary care to specialists in, for example, genetics, endocrinology, cardiology, or oncology. This combination of the database of EGCUT with research and education generates an opportunity for the smooth implementation of the genomic tools into clinical practice.

Collaboration with the primary care practitioners that was established when the recruitment started in 2002 needs to continue and develop further for the future plans of integrating genomic information into health care through primary care. Altogether around 640 family physicians and nurses nationwide have volunteered to be involved in recruitment of participants for the EGCUT. This collaboration has been a significant factor helping to achieve the high participation rate of around 5% of the adult population of Estonia. The recruitment phase can be seen as just a starting point in the working relationship between the EGCUT and primary care practitioners. The collaboration throughout the past decade has not only provided high quality phenotypic data through the inclusion of the data available in pre-existing medical records. It also forms a good basis for further collaboration between the EGCUT providing the genomic data and the family physicians.

Family physicians are the likely candidates to facilitate the use of genomic information in clinical setting, providing the necessary interpretation to accompany genomic information once it has sufficient clinical validity and utility (Feero 2008). It has been demonstrated, that the primary care practitioners are currently not prepared to interpret the complex genome scan results (Burke et al. 2002; Kiray Vural et al. 2009; Maradiegue et al. 2005). A survey was conducted on the family physicians collaborating with the EGCUT where the knowledge base and perspectives regarding genomic information use in clinical care of family physicians were investigated (Leitsalu 2010). Overall, family physicians demonstrated eagerness to apply genomic information in practice and willingness to improve their knowledge base in genetics and genomics. Results also indicated that substantial improvements in genetics education were necessary to achieve the EGCUT's elaborate plans for incorporating genomic information into the health-care system.

To be prepared for these innovative plans outlined above, family physicians need further education in the field. The EGCUT will, together with the Faculty of Medicine of the Tartu University, create an education program in order to prepare family practitioners and other appropriately qualified doctors for offering first line medical counselling in the future. Importantly, the Faculty of Medicine of the University of Tartu is the only provider of medical education in Estonia, ensuring that all physicians will receive the same education. Genomics education can be introduced in two stages: first as a special training program planned to be incorporated into the continued education of practicing health care workers, and later as a subject to be included in the medical school training.

CONCLUSIONS

During the first 10 years, the EGCUT focused on participant recruitment and data collection. By 2010, the anticipated 50,000 individuals were recruited. Although the database is continuously updated and expanded, and the follow-up studies are conducted, the main focus has shifted towards conducting research projects based on the biobank samples, health information and genotyping and sequencing data. At the same time increasing efforts are being made to translate the results of scientific research into clinical practice. The EGCUT is collaborating with the medical community developing a system for implementing the information generated into the Estonian health care system, thereby moving towards the medical model referred to as 'personalised medicine'. Thus the information is collected at the Estonian Biobank, it is processed by the EGCUT together with collaborative efforts world-wide, translated to medical and clinical fields through close partnership with the medical community. This development will lead to the advancement of personal medicine, where patients' individual genetic variants and predictive biomarkers will be assessed together with classical risk factors like age, sex, environment, and life style.

Acknowledgements

This work was supported by FP7 grant OpenGENE, Estonian Research Infrastructures Roadmap project, the Estonian Ministry of Social Affairs and the Estonian Ministry of Education and Research, Targeted Financing from the Estonian Government (SF0180142s08 and SF0180026s09), and by the European Union through the European Regional Development Fund in the frame of Centre of Excellence in Genomics.

REFERENCES

Alavere, H. 2011. Further development of the database of the Estonian Genome Center, University of Tartu: is it possible through re-contacting? Unpublished Master's Thesis. University of Tartu.

BBMRI 2011 = Biobanking and Biomolecular Resources Research Infrastructure. http://www.bbmri.eu/

Biobank, S. 2003 = The Swedish National Biobank Program. http://www.biobanks.se/

Biobank, U. 2011 = UK Biobank – improving the health of future generations. http://www.ukbiobank.ac.uk/

Bousquet, J., et al. 2011. Systems medicine and integrated care to combat chronic noncommunicable diseases. Genome Medicine, 3, 7, 43.

Burke, W., et al. 2002. Genetics in primary care: a USA faculty development initiative. Community Genetics, 5, 2, 138-146.

deCODE 2010 = deCODE genetics – the population approach. http://www.decode.com/research/

EGCUT 2011 = Estonian Genome Center of the University of Tartu, Application form for data release. http://www.geenivaramu.ee/index.php?id=481

Estonia, S. 2010 = Population pyramid. http://www.stat.ee/public/rahvastikupyramiid/

Feero, W. G. 2008. Genetics of common disease: a primary care priority aligned with a teachable moment? Genetics in Medicine, 10, 2, 81-82.

Green, E. D., Guyer, M. S. 2011. Charting a course for genomic medicine from base pairs to bedside. Nature, 470, 7333, 204-213.

Guttmacher, A. E., Collins, F. S. 2003. Welcome to the genomic era. The New England Journal of Medicine, 349, 10, 996-998.

The Health Services Organization Act and Associated Acts Amendment Act, §59[1] section 1 C.F.R. 2007.

HGRA 2000 = Human Genes Research Act. http://www.geenivaramu.ee/index.php?langeng&sub18&eetika1

Kiray Vural, B., et al. 2009. Nursing students' self-reported knowledge of genetics and genetic education. Public Health Genomics, 12, 4, 225-232.

Leitsalu, L. 2010. Counseling for Complex Diseases: Incorporating Predictive Genetic Testing into the National Health Care System of Estonia. Sarah Lawrence College, Bronxville.

Maradiegue, A., et al. 2005. Knowledge, perceptions, and attitudes of advanced practice nursing students regarding medical genetics. Journal of the American Academy of Nurse Practitioners, 17, 11, 472-479.

Metspalu, A. 2004. The Estonian Genome Project. Drug Development Research, 62, 97-101.

Nelis, M., et al. 2009. Genetic structure of Europeans: a view from the North-East. PLoS One, 4, 5, e5472.

NHGRI 2009 = National Human Genome Research Institute. http://www.genome.gov/10001772

Ormond, K. E., et al. 2009. Assessing the understanding of biobank participants. American Journal of Medical Genetics A, 149A, 2, 188-198.

P3G 2011 = Public Population Project in Genomics. Http://p3g.org/

Permanente, K. 2011. Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH).
http://www.dor.kaiser.org/external/DORExternal/rpgeh/index.aspx

Research, E. M. o. E. a. 2008 = Education and Research in Estonia [Electronic Version]. www.hm.ee/index.php?popup=download&id=8022

Research, M. o. E. a. 2010 = Estonian Research and Infrastructure Roadmap o. Document Number

Risch, N., Merikangas, K. 1996. The future of genetic studies of complex human diseases. Science, 273, 5281, 1516-1517.

MEDICAL RESEARCH IN ESTONIA: CURRENT STATUS AND EXPECTED DEVELOPMENTS IN FUTURE

E. Vasar

HISTORICAL BACKGROUND OF MEDICAL RESEARCH

There is no doubt that after the re-establishing independence in 1991 Estonia has made major progress as an independent state. This progress has been achieved thanks to radical reforms carried out in the society, including the reform in science and higher education. Today the fruits of this reform can be easily seen and appreciated by the members of society. One may even suggest that the real development in the medical sciences started after this major reform. Currently, it is obvious that the basis of medical research has strengthened substantially compared to the respective situation under the Soviet government. In many cases it is even difficult to find any comparisons, because several successful research directions did not exist in the Soviet Estonia. There are two major centres involved in medical research in Estonia: one is in the university town of Tartu and the other is in the capital city of Tallinn.

However, the role of Tartu in medical research is more apparent, because the only medical faculty in Estonia is at the University of Tartu. Moreover, several other leading institutions in biomedical research are belonging to the university as well. Nevertheless, because Tallinn has got almost four-fold larger population than Tartu, its role in the development of practical medicine cannot be under-estimated. On the other hand, it is becoming more and more obvious that the level of everyday medical practice depends on whether it is connected to actual scientific research or not. In order to analyse the current situation in medical research let us start from the old university town of Tartu. The Faculty of Medicine was historically among the first four faculties at the University of Tartu. Medical research was especially highly developed in the 19^{th} century under the umbrella of the German-language speaking university of the Russian Empire (Fig. 1-3). The Faculty of Medicine played a prominent role in the development of medical sciences during the previous independence of the Republic of Estonia (1918–1940).

In the Soviet period the development of medical sciences was much stagnated and it was very much like keeping alive the traditions coming from the prosperous history.

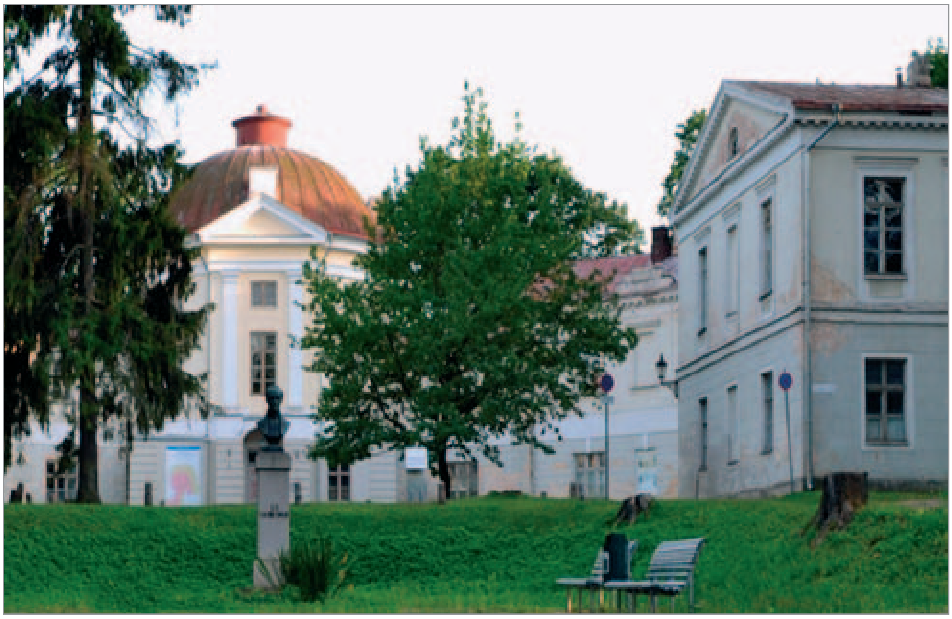


Figure 1.
Old Anatomical Theatre.

Figure 2.
New Anatomical Theatre (1888).

Figure 3.
Clinicum Universitatis Dorpatensis (1808).

However, despite the difficult times several research groups in medical sciences enabled to establish international contacts and subsequently reached the scientific level acceptable by the international standards, neurology (E. Raudam, A.-E. Kaasik, R. Zupping, M. Mägi), psychopharmacology (J. Saarma. L. Allikmets, A. Zharkovsky), gastroenterology (V. Salupere, K. Villako, H.-I. Maaroos) and immunology (R. Uibo).

This period was fruitful for the development of surgery. Research team established by A. Linkberg was widely recognised in the Soviet Union (H. Tikko, E. Tünder, K. Põder, H. Tihane) and they were known abroad as well (especially the research group lead by T. Sulling). There is no doubt that the roots of development of medical sciences in independent Estonia are in many ways based on the work of the above-mentioned research teams.

CURRENT STATUS OF MEDICAL RESEARCH IN ESTONIA

As mentioned above, the situation has changed completely after regaining independence and establishing a fully democratic state (Kaasik, Uibo 2005). In 1999 the preclinical departments, previously disseminated all over Tartu, moved into a new building – Biomedicum (Fig. 4). Probably this has

Figure 4.
The Biomedical Building (1999).

been one of the best investments of Estonia into higher education and science, because this change has been a major driving force for the development of medical sciences in Estonia. This was not just a new building, but it also contained new contemporary scientific equipment, helping to move the medical research from the pleasant memories of 19th century into the reality of 21st century (Kaasik, Uibo 2005). In 2001 the Centre of Excellence of Molecular and Clinical Medicine was established in the Faculty of Medicine. In 2003 it was nominated as the centre of excellence for the Central and Eastern European countries (Uibo 2007). In 2008 new centres of excellence were chosen in Estonia according to inter-national peer-review evaluations. Altogether seven new centres were chosen from 29 proposals presented. The only centre, fully dedicated to the medical research, was the Centre of Excellence for Translational Medicine involving the strongest research groups in the Faculty of Medicine including expertise both in preclinical and clinical research.

The centre of excellence involves the following research groups: virology (I. Lutsar), immunology (R. Uibo), clinical metabolomics (M. Zilmer, U. Soomets), molecular pathology (P. Peterson), neuropharmacology (A. Zharkovsky, A. Kaasik) and physiological genomics (S. Kõks, E. Vasar). The mission of the Centre of Translational Medicine is to raise the international competitiveness of Estonian medical research and to concentrate the skills and funds to cutting-edge areas.

The aim of the centre is to take a step towards multidisciplinary translational approach and to link two major research fields in the Faculty of Medicine, neuroscience and immunology. Altogether, the major goal of this centre is the development of translational medicine aiming from the one side at the quickest application of achievements of basic research into clinical research and after that into the everyday medical practice (from bench to bedside approach). From the other side, clinical practice raises important translational research questions for basic research in order to accelerate the discovery of solutions for urgent medical problems. Among the scientists included into the centre of excellence P. Peterson is widely internationally acknowledged by his research on the functional role of AIRE in the immunological regulation. I. Lutsar is internationally known by her translational studies of HIV infection in Estonia. Her research group participates in different international research networks dealing with HIV infection and antibiotic treatment in neonatal children. R. Uibo is acknowledged as an experienced scientist in the field of immunology. His current research includes immunology of type 1 diabetes in Estonia. In this field he has strong collaboration with the research group from the Department of Pediatrics (Faculty of Medicine, V. Tillmann). This is also matter of international collaboration in the European Union. Moreover, R. Uibo is collaborating with a number of research groups both in Estonia as well as abroad. Among others, collaboration with the University of Leiden should be mentioned. This work is supported by the research grant from the European Foundation for the Study of Diabetes – New Horizons Collaborative Research Initiative. M. Zilmer and U. Soomets are widely known by their studies aiming to reveal a role of different metabolomic factors in the pathogenesis of various diseases, including also neuroimmunological disorders. They have strong research links with different clinical research groups working in the fields of pediatrics (V. Tillmann), anaesthesiology (J. Starkopf), cardiology (J. Eha), dermatology (H. Silm), neurology (T. Asser) and psychiatry (V. Vasar). Under the scope of research groups of A. Zharkovsky and A. Kaasik is the study of probable neuroimmunological mechanisms of neurodegenerative disorders (Alzheimer's and Parkinson's disorder). There is a strong collaboration with the Department of Psychiatry (Faculty of Medicine, E. Maron) and widespread international collaboration within the European Union (A. Veksler, E. Lammertyn and E. Bock). The goal of studies of S. Kõks and E. Vasar is to establish a role of two proteins, wolframin and limbic system associated membrane protein, in the neuroimmunological regulation. There is a collaboration with the clinical research groups from neurology

(T. Asser and P. Taba) and psychiatry (V. Vasar). The research group is included into the COST activity: European systems genetics network for the study of complex genetic human diseases using mouse genetic reference populations (SYSGENET).

From 1991 the new commencement procedure for doctoral degrees was introduced at the University of Tartu, applying rules similar to the Scandinavian countries. First new style PhD degree was commenced by Heidi-Ingrid Maaroos in 1991. For improving the quality of doctoral studies a new track in neuroscience was established in 2001. Today this doctoral school belongs to the Network of European Neuroscience Schools. The number of doctoral degrees commenced in the Medical Faculty during the last 20 years is getting close to 200. It means that on the average 10 PhD dissertations have been defended per year. However, this is definitely not enough to fulfil all the needs for the academic and practical medicine in Estonia. Nevertheless, the quality of PhD dissertations has increased significantly since 1991. The amount of research money allocated for medical research has been substantially raised in Estonia, but it is still below threshold for sustainable development of medical sciences (Uibo 2007). Therefore, significant funding comes from abroad and in recent years its share has been almost 20% of the total research budget in the Faculty of Medicine.

Medical research at the University of Tartu is not restricted to the Faculty of Medicine. There are strong research groups involved in medical research (especially in biomedical studies) in the Faculty of Science and Technology (Institute of Molecular and Cell Biology – T. Maimets, J. Remme, J. Sedman, Institute of Technology – M. Ustav, Ü. Langel, T. Tenson, A. Merits) and the Estonian Genome Centre (A. Metspalu). An independent affiliation Estonian Biocentre (R. Villems) that has strong collaborative links with the Estonian Genome Centre and the Institute of Molecular and Cell Biology is also involved in research in biomedicine. There are two centres of excellence linked to these institutions. The Centre of Excellence in Genomics (M. Remm) is based on the research groups of the Estonian Biocentre, whereas the Centre of Excellence in Chemical Biology (T. Tenson) was formed by the research groups in the Faculty of Science and Technology and Tallinn University of Technology. A good part of research performed in both centres of excellence is dedicated to biomedicine. Besides that the research groups performing research both in biomedicine and public health are in the Faculty of Social Sciences (Institute of Psychology – J. Harro) and in the Faculty of Sports Sciences (M. Pääsuke, T. Jürimäe and J. Jürimäe). These research groups together with the Institute of Public Health (Faculty of Medicine – R.-A. Kiivet and A. Uusküla) and the National Institute for Health Development (T.Veidebaum, M.Rahu) have formed the Estonian Centre of Behavioural and Health Sciences. Their major goal of this centre (J. Harro) is to promote public health oriented research in Estonia.

There is a reasonable collaboration between the scientists working in the field of medicine at the University of Tartu and in the field of veterinary medicine at the Estonian University of Life Sciences. Especially strong connections have been established in the research areas of morphology (A. Arend, T. Haviko, M. Aunapuu and V. Andrianov) and in physiological genomics (S.Kõks, Ü.Jaakma).

Tartu University Hospital (U. Siigur) also plays a major role as the basis of medical research. Ongoing renovation of clinical buildings and obtaining of new equipment definitely improves the quality of medical services, but also affects positively the outcomes of clinical research and application of results into everyday clinical practice (Fig. 5).

One has to underline recent developments taking place in Tallinn University of Technology in relation to medical research. Several research groups are engaged into biomedical research (P. Palumaa, T. Timmusk and P. Kogerman) with outstanding outcomes. Five years ago Technomedicum (K. Meigas, R. Sepper and M. Viigimaa) was established as an independent structure in Tallinn University of Technology. The mission of Technomedicum is to gen-

Figure 5.
Tartu University Hospital.

erate a link between medical and technical knowledge for the promotion of human health and well-being. Technomedicum is an innovative and interdisciplinary scientific institution linking together Tallinn University of Technology, hospitals as well as other organisations and institutions related to healthcare. It provides innovative solutions and know-how in medicine, technology and science, and serves as a generator of synergy within multidisciplinary collaboration of doctors and engineers at a high scientific level. Moreover, Techomedicum started to play a role as a centre for PhD studies. Several PhD degree commencements have been taken place in Technomedicum.

The National Institute for Health Development (M. Jesse, T. Veidebaum, M. Rahu) is mainly oriented on research in the field of public health. The task of the research centre is the performance of research and development works in the form of public procurements, in the framework of international programmes, as target-funded topics and as the Estonian Science Foundation grant projects. Research is performed in areas related to epidemiology and biostatistics, oncology, virology, environmental health and various chronic non-infectious diseases. Research centre includes the Estonian Medical Birth Register and Pregnancy Termination Database, the Estonian Drug Monitoring Centre, the Medical Terminology Commission and the Tallinn Medical Research Ethics Committee. Research centre regularly performs a nationwide study on adult health behaviour. The health status of Chernobyl veterans from Estonia has been regularly explored and registered by the research centre.

The North Estonia Medical Centre (S. Nazarenko) is the second largest hospital in Estonia. However, the capacities of this hospital are not fully used for medical research and educational purposes. The East-Tallinn Central Hospital and West-Tallinn Central Hospital are two other major medical centres in Tallinn. The above mentioned statement is also true about these two hospitals, because their resources can be used in a better way for medical research in Estonia.

CHALLENGES AND OPPORTUNITIES FOR MEDICAL RESEARCH IN ESTONIA

Probably the most important challenge for medical research is education in science. Today the medical curriculum at the University of Tartu is strictly the profession oriented. Many students graduating from the university are not educated in research. They are illiterate in terms of science. The actual development of medical sciences needs medical doctors to be able to analyse the meaningfulness of recent scientific achievements for their quickest application into everyday medical practice. The number of PhD degrees in medicine commenced every year in Estonia is not enough to supply the needs of the country. This is just for satisfying the needs of academic institutions, but definitely not for extending scientific activities and for supplying the field of practical medicine with scientifically oriented medical doctors. What could be the possibilities for overcoming this problem? Today there is a tendency that after graduating from the university, medical doctors start their residency course in parallel with their PhD studies. Indeed, this task is not the easiest one and only few young doctors can handle this challenge. Therefore, this approach cannot be expanded endlessly.

One the other hand, the conditions should be developed for experienced medical doctors wanting to get the PhD degree as externs. However, for achieving a major breakthrough we should introduce the parallel scientific track with the basic and practical undergraduate studies in medicine. It is evident that the existing course of medical studies is rather oriented to the medium level of students and these studies are not challenging enough for people having superior intellectual capacities. Obviously, it is possible that the best students may choose the possibility to take the scientific track in parallel with the studies oriented for obtaining the theoretical and practical skills needed to become a medical doctor. This means that the curriculum of medical studies has to be revised, but definitely this target is achievable. Carefully planned scientific track will be a significant reduction of time for obtaining the MD and PhD degrees. This is obviously the most powerful resource helping to increase the number of medical doctors necessary for increasing the quality of medical research and services in Estonia. Last but not least, PhD studies in the medical field should be opened for the graduated students from the other fields of natural and social sciences as this would definitely help to enlarge the basis of medical research in Estonia.

For several years the possible breakthrough areas have been under the discussion among the scientists in Estonia. A number of Estonian biomedical and medical scientists, working both in Estonia and abroad, believe that drug discovery can be among the strongest opportunities for the country. There are longstanding traditions in this field; the critical mass of scientists is also present to start creating the necessary networking. The existing situation in the Europe is very favourable for starting this kind of activities. There is growing evidence that the big pharmaceutical companies in Europe are moving out from the activities necessary for early stages of drug development and, therefore, outsourcing everything what is below Phase III of clinical trials. Phase III of clinical trials are randomised controlled multicentre trials on large patient groups (300-3,000 or more depending upon the disease or medical condition studied) and are aimed at being the definitive assessment of how effective the drug is, in comparison with current 'gold standard' medication. Because of their size and comparatively long duration, Phase III trials are the most expensive, time-consuming and difficult trials to design and run, especially in therapies for chronic medical conditions. Indeed, only the big pharmaceutical companies have enough monetary resources to conduct these studies. Nevertheless, this leaves a space below Phase III of clinical trials open for academic institutions and small biotechnology companies. It does not necessarily mean that the whole chain of activities in drug discovery has to be established in Estonia. There is rather a clear chance to establish a world-class expertise in specific areas, like developing of disease models in laboratory animals, performing of toxicological and pharmacological studies, performing target validation and so forth. A major breakthrough in a certain area will help to generate the necessary international networks in order to cover all the needed activities below Phase III of clinical trials. Indeed, there is a space under the national biotechnology programme to elaborate the activities necessary for the development of drug discovery platform in Estonia.

Let us give an example about the situation in drug discovery and development. It is no secret that the development of new drugs is getting more and more difficult, and of course, more and more expensive. Therefore, as mentioned above, large pharmaceutical companies try to reduce risks arising from the process of development. One possibility is to skip certain areas and leave them to others. Especially difficult is the situation with the psychiatric drugs. During the last ten years no major breakthroughs have happened. Indeed, we have the psychiatric medications with fewer side effects, but they are not stronger compared to the older ones. Also, in some cases even new side effects have appeared. Instead of extra-pyramidal disturbances the newer antipsychotic drugs cause metabolic syndrome. What is the reason for this failure? Probably our knowledge about the pathogenesis of psychiatric disorders is still rather limited and therefore stronger efforts have to be made in this direction. Frequently we do not realise that a psychiatric disorder is not only a malfunction of the central nervous system, but other measures in homeostatic regulation are corrupted as well. Take, for example, the relation between schizophrenia and type 2 diabetes. Patients with the chronic disease frequently suffer from type 2 diabetes and this disease can be the side effect of the medication with antipsychotic drugs like olanzapine. Therefore, a wider scope is necessary for understanding psychiatric disorders, involving the knowledge about the programming of human development, the wider understanding about the pathological processes in the biological systems (like when disease is *in situ* state and can be still prevented or literally 'dissected out', when disease spreads over the body), the question about endophenotypes and so forth.

There is certain progress in the genetics of psychiatric disorders, but our knowledge is still far from adequate to put together a full picture. In using animal models of psychiatric disorders we do not take into account the full spectrum of the biological background of laboratory animals and to what extent the animal data can be used for understanding human pathology. Indeed, this is obviously enough to give up from the perspective of common sense. Nevertheless, psychiatric disorders and the development of drugs for their treatment is an area where both bottom-up and top-down approaches could be used where necessary as without applying all possible means we have very little chance to get close to the origins of psychiatric disorders.

Estonia is a small and relatively poor country. Therefore, the money available for research is rather limited. Especially it concerns medical research, because contemporary medical research is extensively high-technology dependent and, therefore, the necessary equipment is very expensive. One the other hand, the GLP (good laboratory practice) and GMP (good medical practice) rules must be followed in medical research and the strict following of these rules is not reducing the price of studies either. Indeed, here is an important question for the medical community in Estonia: should we invest in applied science and import the best available knowledge from larger and more developed countries or should we search for ways to support frontline basic and clinical research. The first approach is apparently supported by recently announced National Program for Promotion of Public Health (TerVE 2011). This programme has been generated under the supervision of the Ministry of Education and Research and the Ministry of Social Affairs. The main goals of this programme are as follows:

- Analysis of factors shaping lifestyle and its changes, including risk-taking behaviour, and assessment of their impact.
- Evaluation of environmental risk factors on health.
- Promotion of opportunities for collection and use of information and communication technology-based health data.
- Health technology assessment.
- Elaboration of long-term developmental plan for public health-related research and development, and informing of public about ongoing research and development.

As every other approach, it has both strengths and weaknesses. It is apparently clear that the acute medical services in Estonia are at a rather good level, but the situation is different if the patients need rehabilitation and additional care after acute hospitalisation. This side is far from being ideal in

Estonia and, therefore, corresponding steps should be taken. This is clearly a case where the adaptation of best international practices is necessary. Indeed, considerable steps have been taken by the larger hospitals in Estonia in order to improve the situation. Also, Haapsalu Neurological Rehabilitation Centre has taken initiative in this particular area. However, this is still not enough to cover all the existing needs of Estonia.

In comparison with the highly developed Nordic countries Estonia also lags behind with the health and disease registers. The experience of Scandinavian countries has shown that these registers are extremely effective for making the necessary political adjustments in the medical practice, but their importance for medical research cannot be underestimated also in cases where the involvement of large population cohorts is needed. This well organised information is also useful for the generation of biobanks helping to raise the necessary research questions both for basic and clinical research. Indeed, useful support can be gained from these registers for basic and translational research. The development of health and disease registers is an important challenge in the development of medical services and research in Estonia.

Despite obvious strengths the proposed approach contains important weaknesses, because there is an obvious need for the development of the basic as well as the frontier clinical research. This aspect is completely out of scope of the mentioned program. In the long run this may have a devastating effect for the overall development of medicine in Estonia. There is no doubt that the contemporary medical education has to be based on advanced medical research, because without that we would place ourselves in the same situation where the medical education was in the Soviet Empire. Teaching of students was vastly dominating in the University of Tartu, whereas the research was more like hobby for the university scholars. Nowadays, it is clear that the level of provided medical services in the hospitals is directly dependent on the development of research activities there. According to "The White paper of EMRC: Present Status and Future Strategy for Medical Research in Europe, 2007" the level of medical services in hospitals not performing medical research is almost seven years behind from the existing contemporary level of medicine (Publications 2007). Reduction of curiosity driven medical research also means that we will face the extensive braindrain. We will lose a significant number of talented scientists leaving the country establishing themselves in the leading scientific centres all over the world. There will be very little chance that they will return to Estonia, because the mental and physical infrastructure will not be favourable for the high-class science here. Instead of evidence-based medicine we will extend enormously the questionnaire-based science. Indeed, the questionnaires are the powerful vehicles in the hands of social scientists. However, no questionnaire is replacing evidence collection from the basic, translational and clinical research. The current development is nothing else but a long step from the evidence-based medicine toward the questionnaire-based sciences. This program is designed for limited group of scientists in Estonia. The majority of the above-mentioned institutions involved in medical research in Estonia are completely out of scope of the program. The established program exactly fits the purposes and tasks of the Estonian Centre of Behavioural and Health Sciences.

In preparation for the forthcoming 8th Framework Program for research, an impact assessment exercise has been initiated by the Health Directorate to examine the performance of the Health programme in previous framework programmes (FP6 2002-2006 & FP7 2007-2013). Although the Health programme has been successful in many of its targets, improvements can still be suggested for the remaining period of FP7 and the next FP8 program. Among those needing improvement is that the new member states from the Central and Eastern Europe (EU-12 countries) are lagging behind the 'older' countries by means of scientific excellence. Therefore, scientific excellence should be promoted in EU-12 countries.

Unfortunately the fostering of applied medical sciences does not increase the overall competitiveness of Estonian science. An increase in scientific excellence and competitiveness are the ultimate goals for Estonia. The involvement of Estonian medical researchers in Framework Programmes is significant, but their role is rather secondary and they do not participate in active building up of research networks. What are the possibilities for promoting of scientific excellence in Estonia? The above-mentioned TerVE programme may have only indirect impact if the tasks taken for development of health and disease registers will be successful. Otherwise, this programme has got only limited value for promoting scientific excellence in Estonia. Fortunately, there are other instruments in Estonia helping to foster medical research. Estonian research infrastructures roadmap 2010 identified the national centre for translational and clinical research. The main goal for establishing this centre is to develop a supporting infrastructure to ensure the sustainability of the medical research in Estonia – new equipment and technologies for the planned laboratory complex with a modern experimental animal centre, to fulfil the increased medical research demand.

The established infrastructure incorporates functional units with three different objectives: 1) fundamental research unit aims to understand disease mechanisms and to identify drug targets. This unit has laboratories for working with dangerous infectious diseases (e.g. HIV, hepatitis viruses, etc.), for regenerative medicine, the tissue bank and laboratories with modern analytical technologies (metabolomics, genomics, glycomics); 2) disease models unit aims to validate drug targets and to develop new intervention and diagnostic strategies by using animal models for diseases. Unit involves animal houses for small animals, technologies for modelling diseases in animals, laboratories for studying physiological and behavioural features of animals and modelling infectious diseases; 3) functional imaging unit works on the testing of new drugs and diagnostic methods. This unit is involved with application of modern imaging technologies (magnetic resonance imaging, radiological imaging, fluorescence, luminescence, etc.) on experimental animals.

As mentioned above, Estonia is a small country with limited resources. Therefore, building up of double systems in the medical education and research is out of any sense. Wasting limited money coming from the taxpayers pocket is not the best business plan. A sustainable development, at most cannot be guaranteed using this attitude. Somehow the situation in Estonian medicine resembles the arms race between superpowers rather than intelligent collaboration taking into account the strengths of both sides. Nevertheless, steps should be taken to establish proper collaboration between the largest hospitals in Estonia in order to increase not only the quality of medical services, but also the competitiveness of medical research in Estonia. The size of Estonia does not allow in many cases to gain the necessary power for performing advanced translational and clinical research. It is also clear that not only the medical services have to be reinforced, but also the rehabilitation of chronic diseases has to be facilitated. This means also that academic medicine has to follow the trends and the necessary academic networks should be built. Instead of competition there has to be a substantial increase in collaborative efforts between the leading medical centres in Tartu and Tallinn.

Estonia has to keep the eyes open for all outstanding developments in the medical area. According to the recent publication in New England Journal of Medicine the contemporary medicine is facing the following major challenges (Moses, Martin 2011). The process of development of new treatments is slow and therefore there is a growing disappointment about the overall performance. On the other side the progress in molecular medicine, especially in the development of methods of genomics and fields of 'omics', provides unique opportunities to start the personalised medicine. There is a proposal to establish in parallel with the evidence based clinical practices hospitals that would work in the forefront of modern science and could, thanks to the application of modern methods of diagnostics, use different perspectives for the treatment of currently untreatable diseases (Moses, Martin 2011). The roles of academic health centres and teaching hospitals must be modified to improve their ability to conduct early-stage (proof-of-concept) clinical trials

(Moses, Martin 2011). Here, entirely new models for interaction are required, probably involving freestanding independent institutes or autonomous units within academic centres, where patients come specifically for access to such early-stage studies and where the mutual expectations for investigators, companies, and patients are clear and unambiguous. This change will hasten the divergence between institutions that offer routine care (and that are managed to provide low-cost, reproducible high quality) and those with capability for scientific innovation (where early-stage investigations occur) (Moses, Martin 2011).

E-Health is one of the cornerstones for the National Program for Promotion of Public Health. However, it is not entirely clear whether the authors of this plan have a wider understanding how to use this powerful approach. Therefore, it is probably feasible to follow the best practices available in the leading countries. One of those is under development in the USA. Namely, in 2009, the Food and Drug Administration (FDA) gave the Harvard Pilgrim Health Care Institute the lead role in fulfilling a 5-year contract to establish a system – the Mini-Sentinel – for developing and testing approaches and methods that could be used to inform the structure and operations of the full Sentinel System (Behrman et al. 2011). The institute is now leading a diverse partnership of approximately 200 epidemiologists, clinical content experts, statisticians, and data specialists from 27 institutions that are participating in this pilot system (Behrman et al. 2011; Mini-Sentinel). The FDA will soon begin to actively monitor the data, seeking answers to specific questions about the performance of medical products, such as the frequency of myocardial infarction among users of oral hypoglycemic agents (a topic selected because it has been difficult to identify drug-induced myocardial infarction through existing prospective surveillance mechanisms) (Behrman et al. 2011). The FDA will also monitor the occurrence of adverse events associated with select routinely administered vaccines. Using the Mini-Sentinel system, the FDA will also be able to obtain rapid responses to new questions about medical products and, eventually, to evaluate the health effects of its regulatory actions (Behrman et al. 2011). This monitoring portfolio will expand as the FDA and its collaborators acquire experience and develop operational efficiencies and as additional data resources become available. The distributed-database-and analysis model and the infrastructure of the Mini-Sentinel data network can be extended to other forms of evidence development. Provisions in the economic stimulus and health care reform legislation, and a recent report from the President's Council of Advisors on Science and Technology, envision expanded use of electronic health information for other types of public health surveillance, quality measurement, comparative effectiveness research, and biomedical research – all of which are essential to improving the country's health and health care delivery system (Behrman et al. 2011; Report 2010). There is no doubt that it is particularly challenging to establish appropriate governance for a distributed data network that can support multiple secondary uses for health information (Behrman et al. 2011).

CONCLUSIONS

Current overall development of medical sciences has been favourable. However, the existing funding of medical sciences is far from being enough for keeping sustainable development. The future of medical research will depend on the effectiveness of PhD studies, as well as on the finding of new models for PhD studies. The existing resources of Estonia do not allow strictly separated development of medical centres in Tartu and Tallinn. There is an ultimate need for stronger cooperation between these major centres in Estonia both in the medical education as well as medical research. For using the full capacity and for the improvement of medical services in Estonia the academic medicine, where it is necessary, has to find a way to the smaller medical centres in the country. The possible area of breakthrough for medical sciences in Estonia could be the effective implementation of specific areas from the drug discovery, like developing of disease models in laboratory animals, performing of toxicological and pharmacological studies and performing target validation. The promotion of scientific excellence will be another major task for medical sciences in Estonia in order to be the competitive in the European Union and to be attractive in establishing of the necessary collaborative links with the leading scientific centres around the world. The impact of the National Program for Promotion of Public Health for the overall development of medical sciences is not clear yet. The limited scope of this program toward the public health does not allow the participation of the majority of institutions involved in medical research in Estonia.

ACKNOWLEDGEMENTS

I would like to express my highest gratitude to members of the Estonian Academy of Sciences Ain-Elmar Kaasik and Raivo Uibo for their constructive criticism. Jürgen Innos has taken care for revising of the English language of manuscript.

REFERENCES

Behrman, R. E., et al. 2011. Developing the sentinel system – a national resource for evidence development. The New England Journal of Medicine, 364, 6, 498-499.

Kaasik, A.-E., Uibo, R. (eds.) 2005. Teadusmõte Eestis: Arstiteadus. Estonian Academy of Sciences (in Estonian).

Mini-Sentinel = http://www.minisentinel.org

Moses, H. III , Martin, J. B. 2011. Biomedical research and health advances. The New England Journal of Medicine, 364, 6, 567-571.

Publications 2007 =
http://www.esf.org/research-areas/medical-sciences/publications.html

Report 2010 = Report to the President Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward.
http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf

TerVE 2011 = http://www2.archimedes.ee/teadus/?leht=325

Uibo. R. 2007. Arstiteadus Eestis – saavutused ja probleemid.
http://www.akadeemia.ee/_repository/File/TEGEVUS/Esitlused20070221/Uibo.pdf

CHEMISTRY IN ESTONIA – REVIEW AND OUTLOOK

M. Karelson, I. Koppel

ANALYTICAL CHEMISTRY

I.Leito, M.Kaljurand

Introduction

Analytical chemistry is a very practical branch of chemistry and its approaches and methods are needed in many other areas ranging from geology to medicine and from ecology to materials science. As a result of this, it is quite logical that the research activities related to analytical chemistry are distributed among different universities and other institutions in Estonia. The leading role is held by University of Tartu and Tallinn University of Technology.

The implicit philosophy of science behind the idea of basic science is not very helpful in understanding the current dynamics of science in general as well as analytical chemistry in particular. In fields such as life sciences, parts of materials research, nanotechnology and most of the medical sciences we see that major investments in research are legitimised by promises that the research will result in economic and social benefits. Such promises are often much more articulate than the general claim that, by moving the frontiers of knowledge, in the end new territories of exploitable opportunities are created.

Frequently "analytical chemistry is presented as a tool, perpetuating the stereotype that analytical data are simply a support to other scientific efforts" (Raynie 2011). Analytical chemists are considered by other chemists as scientific partners with research teams developing new technology, products, or processes. Definitely, this approach is too negligent. In reality, analytical chemistry is a part of information science and looking at it in the context of the information theory may help to systematise attempt to peer into the longer-term future of analytical chemistry with the aim of identifying the areas of strategic research and the emerging of generic technologies likely to yield the greatest economic and social benefits.

Analytical Chemistry at University of Tartu

The research into development of novel analytical methods and fundamental studies in analytical chemistry at University of Tartu is led by the Institute of Chemistry. The highlights of these works are related to mass spectrometry, environmental analysis, enzyme biosensors and quality assurance of analytical results. For the first time a quantitative approach for characterising the ionisation efficiency of compounds in the electrospray ion (ESI) source of mass spectrometer has been developed and applied to a range of different molecules allowing elucidating the relationships between molecular structure and ease of ionisation. A series of methods have been developed for combating the matrix effects that is currently the main limitation of the otherwise very successful liquid chromatography electrospray ionisation mass spectrometry (LC-ESI-MS) analysis method. A novel construction approach of biosensors has been developed whereby the enzyme is applied to the sensor body as a coil of thread-like material containing immobilised enzyme. The most accurate micro-method available for dissolved oxygen concentration measurement has been developed. These are just a few examples of recent success.

A wide range of very practical analytical tasks are carried out at institutes of biochemistry, physics, technology, geology and many others. In those cases analytical chemistry usually serves as a tool in obtaining scientific results in other disciplines. An important trend is ever-increasing partnering of routine laboratories – such as health protection, environmental, as well as numerous industrial laboratories – with scientific research groups for setting up analytical methods at the laboratories and/or jointly solving analytical problems. Here the leading role is held by the ISO 17025-accredited Testing Centre of University of Tartu. In line with the ever increasing role of quality assurance in every sphere of human activity, an important part of the interactions between analytical chemists at universities and routine laboratories is focused on quality aspects of analytical chemistry – obtaining reliable results and convincingly demonstrating their reliability.

An important contributor to the above mentioned achievements has, no doubt, been the significant improvement of the research infrastructure during the recent years. The most visible of these improvements is relocation of the UT Institute of Chemistry to the brand-new Chemicum building opened in autumn 2009. Perhaps not less important are the different governmental research infrastructure programmes. These programmes have enabled purchase of such powerful analytical chemistry instruments as Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR MS), Fourier, Fourier transform infrared microscope (FT-IR microscope), advanced Raman spectroscopy systems, extensive electrochemical analysis installations, to name just a few.

The possibilities of these instruments are wide-ranging. For example, the FT-ICR MS is in principle able to determine the composition of highly

Figure 1.
Fourier transform ion cyclotron resonance spectrometer system.

Figure 2.
Chemistry students in the analytical chemistry laboratory.

complex objects of organic and biological nature or to discover trace level contaminants/impurities in such objects.

Aided by the top infrastructure the research has been highly fruitful. During the last five years roughly 30 analytical chemistry related peer-reviewed scientific papers have been published and six PhD dissertations have been defended at University of Tartu, Institute of Chemistry.

Analytical Chemistry at Tallinn University of Technology and Tallinn University

The introductory comments presented above explain well the basis for choosing research directions that analytical chemists at Tallinn University of Technology have taken. The selection of projects has been made with specific applications and uses in mind. Academic analytical chemistry is practiced in the Institute of Chemistry at Tallinn University of Technology and to a lesser extent in the Institute of Mathematics and Natural Sciences at Tallinn University. Research has been concentrated on developing analytical separation methods in biomedicine and environmental chemistry. Mostly, capillary electrophoresis (CE) is used as a separation method while the chromatography is used to a lesser extent. Application in biomedicine field has been related to the developing methods for determination of antioxidativity of natural products (plant extracts). As a result about 20 papers have been published during the last five years, mainly in Tallinn University of Technology, but growing activity of Tallinn University should be acknowledged here. In the field of environmental analysis the highlight is a proposal of the model of a field portable CE analyser for military use (detection of explosives, toxic industrial chemicals and chemical warfare agents). Third direction in research is related to the materials science: developing nano size and nanoporous materials, for the use in analytical chemistry as substrates for various applications (active particles, electrodes, chromatographic supports). Carbon aerogels developed in our institute have been applied as components in electrical actuators, membranes and paste electrodes (Palmre et al. 2011).

Recent support from EU has greatly improved instrumental capabilities of academic analytical laboratories in Tallinn. This has created a new line of research: the exploration of those instruments (mass, laser and diode array spectrometers) in connection with CE. When many existing solutions can be successfully implemented a surprising new problem of using those instruments starts to emerge: is the cost of analysis and maintenance justified by the results obtained? Maybe the information can be received cheaper and less harmfully for the environment. The green analytical chemistry is an emerging field in analytical chemistry (GAC) where Tallinn has been actively contributing by publishing a book in this field (Koel, Kaljurand 2010). GAC looks for environmentally benign alternatives to traditional analytical methods. GAC achieves its goal via reduction of use of solvents and energy, so miniaturisation of the instrumental analysis is the key issue in GAC. Portable field analyser is one of the practical achievements in this direction. An even more promising trend is microfluidics. The field is highly competitive and crowded but looking for unexplored areas analytical chemists in TTU have found droplet and digital microfluidics as a possible emerging field of investigation. Following is one typical example of possibilities of miniaturisation.

Fig. 3 comes from a research project on digital microfluidics carried out at Institute of Chemistry at Tallinn University of Technology. It shows stages of an oxidative reaction performed in droplets, implementing a process called electrowetting on dielectric (EWOD).

This phenomenon is used to produce very small droplets at nano- and microscale for transport and small-scale chemical processes. EWOD im-

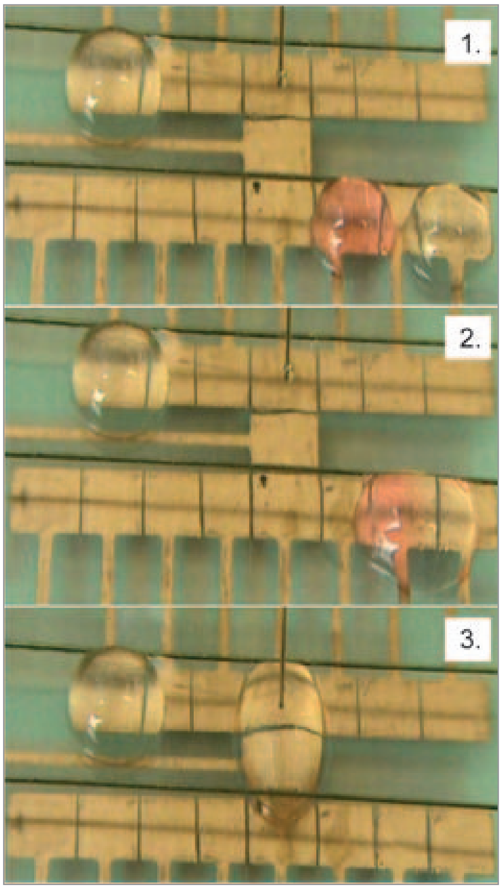


Figure 3.
Performing chemical reactions on the programmable EWOD platform. The platform is made from a printed circuit board which is covered by two thin layers (50 μm thick) of dielectric and hydrophobic materials correspondingly. Movement of a 3 μL droplet is programmed by software which powers the electrodes (visible as yellowish squares). The two droplets at the bottom merge (frame 2), the reaction proceeds (note the change of colour) and the reacting mixture is transported to the centre of the platform, underneath of CE capillary for analysis (frame 3). The droplet at the upper left corner is a CE buffer. This droplet will be transported under the CE capillary after the introduction of the sample is over.

proves the analysis of the contents of small droplets by being able to transport the droplet to the analyser. In general, after incubation, the resulting droplets can be analysed by mass-spectrometry. A high resolution separation method, CE, has been demonstrated for the analysis of droplet contents. The research was published in 2009 (Gorbatsova et al. 2009).

At the moment MALDI-MS and ESI-MS are intensively explored as analysers in Tallinn University of Technology.

Both methods are highly sensitive techniques that are well suited for analysing these microlitre droplets. Detection by NMR of droplets is a novel approach, anticipated at Tallinn University of Technology. NMR is a complementary technique which is highly informative.

Clearly this approach is not just a study of a natural phenomenon. Such research is typical for analytical chemistry. It is a performance that the researchers have created with specific applications and uses in mind: the platform can be used e.g. for an assay testing the activity of enzymes using µL-s of each enzyme in a droplet. Another application could be combinatorial chemistry. High yields in organic reactions require optimisation of the reaction conditions, including the choice of reagent, solvent, concentrations, temperature and reaction time. Such optimisation necessitates performing many reactions, and reactions on the scale of microlitres per reaction are no longer applicable to precious compounds such as advanced synthetic intermediates and organic molecules isolated from many natural sources. Many reactions can be conducted in parallel, and the consumption of samples can be reduced to nanolitres per reaction by using digital microfluidics of droplets. Multiple reactions can be set up.

The hypothesis here is that compartmentalisation of droplets (which have undergone either reactions or preprocessing in the digital microfluidic platform) into a capillary tube is possible and the sequence of droplets can be analysed by remote mass spectrometer or NMR (taking advantage of flow cells for the NMR probes). In this way the capability of this method will strengthen further.

Microfluidic paper-based analytical devices (µPADs) are a new class of point-of-care diagnostic and detecting devices proposed by Martinez et al. (2010). µPADs are essentially GAC devices that are inexpensive, easy to use (and designed specifically for use in developing countries). Attempt to look into the longer-term future of analytical chemistry locates µPADs as one of emerging generic technologies and analytical chemists in Tallinn see this as one of the areas of strategic research which likely yields economic and social benefits. µPADs rely strongly on colorimetric assays, so the different assay formats will be tested for their performance. Since the assay should be as general as possible we speculate that the use of functionalised gold nanoparticles will achieve very special role in the future of analytical and bioanalytical chemistry. Solution of colloidal gold is red due to local surface plasmon resonance. When functionalised, the nanoparticles of gold react with the target analyte and frequency of the resonance shifts which is detected photometrically. The list of possible applications seems endless. Also different substrates instead of paper are of interest.

Conclusions and outlook

As in the case of many other areas of science the whole science of chemistry is currently in the state of changes and increasing interaction with other natural sciences. There have recently emerged or are in the process of emerging such new fields as supramolecular analytical chemistry (i.e. application of selective molecular recognition in analytical chemistry), metrology in chemistry (application of the metrological and quality assurance concepts in analytical chemistry), the above described green analytical chemistry, etc. This trend of 'blurring' of the boundaries between different fields is perhaps the most remarkable development that is ahead.

Other foreseeable developments include development of sensors that enable obtaining analysis results faster and with less expense and effort than most of the today's instruments; increase of the importance of distance analysis – possibility to analyse objects that are inaccessible, dangerous, unique or valuable without taking samples or in many cases even without a physical contact with them; ever increasing complexity of the objects to be analysed; higher required accuracy and lower limits of detection – especially in fields such as health, forensics, environment; still increasing importance of quality assurance of analytical results.

In conclusion one of the future trends in analytical chemistry seems to be greening of the analytical process via miniaturisation and nanotechnology. What was not discussed here will probably be the extensive use of hyphenated separation-spectral instrumentation for analysis of proteome and metabolome of different classes of biological objects. Enormous amount of information that such combination yields will be processed by advanced chemometric techniques. The Holy Grail of such approach is definitely identification of biomarkers of various diseases.

THE STUDY OF CHEMISTRY OF SUPERACIDS AND SUPERBASES AT UNIVERSITY OF TARTU – BREAKING THE 'LIMITS OF GROWTH'

I. Koppel

The extensive introduction of heterogenous and homogenous superacid and superbase systems into chemical engineering has important technological implications for organic synthesis, petrochemical industry, the design and creation of novel, highly effective and environmentally friendly power sources, stereoselective catalysts, etc. (Olah et al. 2009; Ishikawa 2009; Gladysz 2002; Kagan 2002; Wakihara, Yamamoto 1998).

It has been demonstrated that the novel superstrong Brønsted acids or their salts can be used as effective components of the fuel cells, lithium-ion batteries, electrochemical supercapacitors, etc., catalysts for several important laboratory and industrial reactions, (Gladysz 2002; Kagan 2002; Wakihara, Yamamoto 1998) and as a new effective media (e.g., ionic liquids, etc.) for several reactions of synthetic and industrial importance.

Some time ago an experimental quantitative gas-phase scale of intrinsic superacidity scale of large number of strong NH, CH, OH and SH Brønsted acids was established (Koppel et al. 1994; Leito et al. 2009) which includes several novel superacids which are by far stronger acids than, e.g., the 'classical' sulphuric acid.

Even much stronger gas-phase acidities were recently predicted for a large number of different superstrong Brønsted acids (Koppel et al. 2000, 2002; Burk et al. 2008).

The practical demands and fundamental challenges have stimulated the still ongoing race for the creation of the least coordinating anions and for the design and preparation of the strongest possible neutral Brønsted acid (Koppel et al. 2000, 2002; Burk et al. 2008; Stoyanov et al. 2006; Lipping et al. 2009).

Also, a significant breakthrough was made (Issleib, Lischewski 1973; Schwesinger et al. 1996) due to the discovery and synthesis of a new class of neutral organic superbases – peralkylated phosphazenes, the basicity of which in solution seems to approach that of KNH_2. They are widely used in organic synthesis as selective, strongly basic reactants (Issleib, Lischewski 1973; Schwesinger et al. 1996). It was recently demonstrated (Koppel et al. 2001; Kolomeitsev et al. 2005; Kaljurand et al. 2007) that several organic superbases are expected to reach even higher intrinsic basicities (above 300 kcal/mol), whereas the respective ylides are expected to be much stronger bases than their imine counterparts.

However, despite a very significant progress in the field of practical use of superacidic and superbasic media and/or reagents the dynamics and detailed mechanism of heterogeneous and homogeneous superacid or-base catalysed processes have been only scarcely studied.

Also, rather seldom are the studies of the very important problem of the influence of the reaction medium on the superacidity and superbasicity of neutral Brønsted acids and bases.

Besides that, an attempt will be made to study the interactions of superstrong acids and bases with each other in order to study the feasibility of the spontaneous proton transfer between neutral partners in the gas phase (see also Koppel et al. 2000, 2002; Burk et al. 2008) which are of great interest.

The studies (Koppel et al. 2000, 2001, 2002; Burk et al. 2008; Kolomeitsev et al. 2005; Kaljurand et al. 2007) at UT Institute of Chemistry indicate that the resources of further increase of the superbasicity and superacidity of neutral molecules are far from being exhausted. Very significant progress in this field is at reach using some new principles (vide infra) of the design of superstrong acids and bases. It is reasonable and reliable to expect that, as a rule, the progressive increase of the superacidity or the superbasicity of novel superacids and superbases will be accompanied also by much higher catalytic power, and creation of more effective and environmentally friendly materials for electrical power sources. In that field tremendous progress is expected.

The simplest strategy to develop progressively more acidic neutral superacid systems (Koppel et al. 1994, 2000, 2002; Leito et al. 2009; Burk et al. 2008, Stoyanov et al. 2006; Lipping et al. 2009) would be the design of molecules which, along with the acidity site (C-H, O-H, S-H, N-H, etc.), include (several) highly dipolar superacceptor and strongly polarisable substituents which form very extensive, strongly conjugated system with the anionic protonisation center. Proceeding from the above-described princip-les, several new families of superstrong Brønsted acids could be anticipated.

The studies (Koppel et al. 2000; Stoyanov et al. 2006; Lipping et al. 2009) indicate that in case of the novel, potentially extremely strong superacids, conjugate acids of carboranes, the principles of design of superstrong acids still need further modification: besides the considerations described above one should take care not to introduce into the anion of the superstrong acid any features which could attract the proton – loosely bound electron pairs or π- electrons, etc. There is a solid basis to expect (Koppel et al. 2000; Stoyanov et al. 2006; Lipping et al. 2009) that some of the carborane-based superacids might exceed by their superacidity GA the classical mineral acids like H_2SO_4 by more than 80 (!) powers of ten.

Progressively more basic neutral superbase systems would involve,

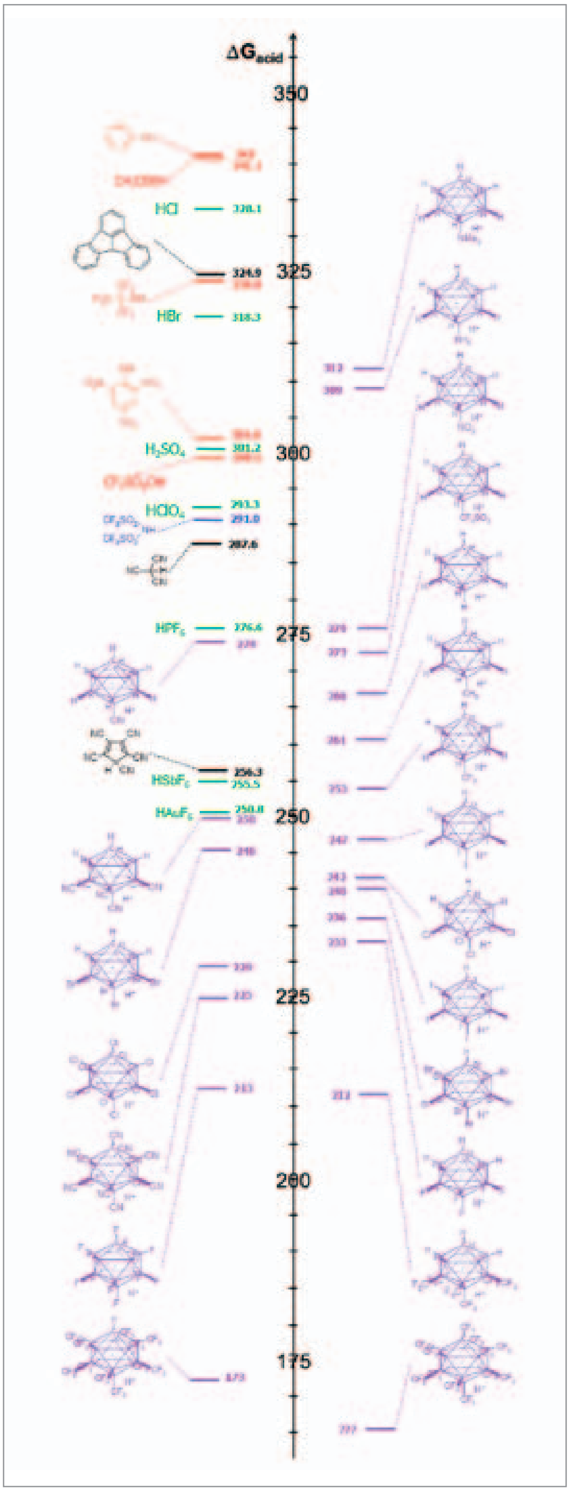


Figure 4.
Gas-phase superacidities ΔG_{acid} (kcal/mol, DFT B3LYP 6-311+G**) of some monocarbadode-caborate acid derivatives compared to a selection of Brønsted acids (Koppel et al. 2000; Stoyanov et al. 2006; Lipping et al. 2009).

along with the basicity site (e.g., = N-R) several highly electron-donor and polarisable substituents which can form extensive conjugated system with the protonated basicity center. Our studies in the gas phase and solution show that using at phosphorus atom the tetramethylguanidino- substituents instead of NMe2 substituents, results in guanidinophosphazenes which exceed by many powers of ten the basicity of the latter traditional phosphazene bases (Kolomeitsev et al. 2005; Kaljurand et al. 2007).

Our investigations (Koppel et al. 2001; Kolomeitsev et al. 2005; Kaljurand et al. 2007) indicate that the other families of superbasic compounds – phosphorous ylides (and some carbenes) are predicted to be even stronger neutral bases.

The above-described results indicate that the interaction of some of the currently existing and available neutral superacids and phosphazene or phosphorus ylide bases should make, for the first time ever, possible to observe in the gas phase the spontaneous proton transfer reaction between two neutral partners. The results described (Koppel et al. 2000, 2002; Burk et al. 2008; Lipping et al. 2009) strongly support this conclusion.

Very little attention has been given to the dependence of the acidity and basicity of strong and superstrong Brønsted acids and bases on the nature and solvation characteristics of the solvent. This problem was addressed in detail by experimental measurements of the acidity and basicity of strong and superstrong neutral acids and bases in the gas phase and in the nonaqueous s solvents CH3CN (Kaljurand et al. 2005), heptanes (Rõõm et al. 2003), THF, dichloroethane (Kütt et al. 2011). In the latter case, for the first time a very extensive (over 15 pKa units) superacidity scale in lower polarity aprotic solvent was established etc. Recently, the extensive gas-phase superacidity scale of cationic Brønsted superacids (conjugated acids of superweak neutral bases, e.g. H_2O, etc.) was established (Leito et al. 2010). The influence of some other factors like internal steric strain of cage-like acids (Abboud et al. 2003, 2004; Herrero et al. 2007), intramolecular H-bond (Rõõm et al. 2007), and relativistic effects, should also be seriously kept in mind.

ORGANIC AND BIOORGANIC CHEMISTRY

N. Samel

In mammals, lipid mediators, e.g. prostaglandins, thromboxanes and leukotrienes, are intimately linked to immune and inflammatory responses, cell proliferation and apoptosis. They have been shown to be major determinants in many pathologies, including diabetes, cancer, cardiovascular and neurodegenerative disorders. Lipid metabolising and regulating cascades have been targeted for drug development. Plant oxylipins, such as jasmonates, have antimicrobial effects, stimulate plant defence gene expression and regulate plant growth and development. While the occurrence of oxylipins has also been detected in certain invertebrate animals and lower plants, very little is known about their metabolic routes and biological role in these organisms. Since the early nineties, N.Samel's group research activities have mainly been focused on the elucidation of pathways of oxylipin biosynthesis in lower organisms, and the characterisation of fundamental catalytic, metabolic and regulatory abilities of the enzymes involved (fatty acid cyclooxygenases, lipoxygenases and peroxidases), as well as their evolutionary origins. Collaborative studies with A.R.Brash of Vanderbilt University on the synthesis of oxylipins in coral led to the discovery of the lipoxygenase-allene oxide synthase fusion protein pathway of the allene oxide synthesis (Koljak et al. 1997).

The unravelling of the 30-year-old enigma of marine prostaglandin biosynthesis, along with the cloning and characterisation of distantly related marine cyclooxygenase homologues, has cast new light on the structure-function relationship of cyclooxygenase genes and enzymes (Valmsen et al. 2001). This has also spurred the discovery of novel facets of the oxygenation stereo control in cyclooxygenase catalysis (Valmsten et al. 2007). Identification and characterisation of arachidonate 11R-lipoxygenase (Mortimer et al. 2006) gave a highly stable model protein for X-ray crystal studies and investigation of Ca^{2+}-mediated activation of lipoxygenases. The other current interests of the group are: mechanisms of cyclooxygenase catalysis and inhibition; oxylipin biosynthesis in arthropods and algae; metabolism and biological role of oxylipins and oxylipin signalling in non-vertebrate animals.

A.Lõokene's group investigates the molecular mechanisms that regulate lipoprotein metabolism. Lipoproteins are non-covalent assemblies of lipids and proteins, a major function of which is to transport lipids through body fluids. The impairment of lipoprotein metabolism leads to atherosclerosis, Alzheimer's disease, pancreatitis or diabetes. Lipoprotein metabolism is under the control of an intricate regulatory network that includes a number of proteins, lipids and proteoglycans. The research of A.Lõokene's group is specifically focused on lipoprotein lipase, the enzyme that plays a central role in the metabolism of triglyceride-rich lipoproteins. The activity, stability and transport of lipoprotein lipase is influenced by a number of ligands, such as apolipoproteins C-II, C-III and A-V, fatty acids, angiopoietin-like proteins 3 and 4, heparan sulfates and cell surface receptors. Of the above proteins, apolipoprotein A-V and the angiopoietin-like proteins were discovered only a few years ago and their action has been so far unknown. Together with our collaborators Dr Lõokene has proposed that apolipoprotein A-V mediates binding of lipoproteins to cell surface proteoglycans and low-density lipoprotein receptor family members (Lõokene et al. 2005). Interestingly, lipoproteins interact with cell surface receptors similarly to viruses (Nilsson et al. 2011). In collaboration with G.Olivecrona's group (Umea University, Sweden), A. Lõokene has elucidated the molecular mechanism of angiopietin-like protein 4 (Sukonina et al. 2006), apolipoprotein CII and dimerisation of lipoprotein lipase. The knowledge of these interactions, in combination with available structural information, provides the basis for the design of drugs that can be used for the treatment of lipid metabolic disease.

Platelets are small cells of great importance in the processes of thrombosis, haemorrhage and inflammation. Under normal physiological conditions, when a blood vessel is damaged, the task of platelets within the circulatory system is to arrest the loss of blood. This process involves the rapid adhesion of platelets to the exposed subendothelium, followed by platelet aggregation, which ultimately culminates in the formation of a platelet plug temporarily sealing off the damaged vessel wall. In contrast, in pathological conditions, such as atherosclerosis, arterial thrombus formation may limit the blood supply to nearby tissues, thus causing local ischemia and/or the progression of atherosclerotic lesions. The inhibitors of platelet activation (anti-aggregants) are of great interest for medicine since they can be used for the inhibition of thrombus formation. At present, medicine uses a number of anti-aggregants, such as aspirin. However, these anti-aggregants have unwanted side effects and are not always effective. G. Kobzar's research activities have been focused on the inhibitors of platelet activation. An ideal anti-aggregant should not result in bleeding complications. It is therefore essential to understand the exact mechanisms of the action of anti-aggregants. The research in G. Kobzar's group has revealed a number of natural weak anti-aggregants that together cause the inhibition of platelet aggregation much more effectively than does each of the compounds alone (Kobzar et al. 2009). The analyses of the synergistic effects of these anti-aggregants revealed the lipid mediators responsible for the transduction of signals in platelets. Another topic of our present interest is the mechanism of 'aspirin resistance', i.e. the inability of aspirin to inhibit the enzyme cyclooxygenase in platelets. The results of our study indicate that lactic acid – end product of anaerobic glycolysis in platelets might be a mediator of the effect of glucose on aspirin inhibition in platelets (Kobzar et al. 2011).

Enzyme-catalysed reactions are of great importance for living organisms. Studying such reactions *in vitro* provides useful procedures for synthetic chemistry, taking advantage of the selectivity of enzymes that provide the opportunity to produce stereochemically pure compounds. In addition, valuable information can be obtained to better understand the pathways taking place in nature. Lipases are enzymes that hydrolyse triglycerides into fatty acids and glycerol. They can be found in several organs and body fluids of living organisms. Commercially available immobilised lipases are widely used in organic synthesis, as they retain their catalytic activity also in organic media and accept artificial substrates.

O. Parve's group has focused on the investigation of the chemoenzymatic synthesis and lipase-catalysed modification of deoxy sugar esters (Villo et al. 2007), along with some natural compounds (sugars, prostaglandins and bile acids). Different derivatives of carbohydrates and glycoconjugates can be used as medicines, as well as in the food and cosmetic ingredients. The lipase-catalysed decyclisation of hemiacetals leading to their activation *via* aldehyde formation has been a recent topic of interest. The decyclisation mechanism and synthetic applications of different hemiacetals, deoxy sugars, sugars and glycoconjugates in particular, are under investigation. The scope and limitations of the developed stereoselective synthesis of (deoxy) sugar esters will be further studied. Lipases are also potential drug targets. Therefore, the influence of some compounds on the lipase activity is being tested as well. Synthetic work is supported by quantum chemical and molecular mechanical calculations. In order to clarify the arising conformational problems, some conformational search procedures have been proposed (Tamp et al. 2008).

ORGANIC CHEMICAL SYNTHESIS

M. Lopp

In his famous book "Introduction to the theoretical organic chemistry" that has been a basic textbook for Estonian students for decades Viktor Palm stated in 1974, "We hold the opinion that in spite of enormous historical, practical and methodological importance of chemical synthesis, it cannot be in the role of the fundamental basis of organic chemistry".

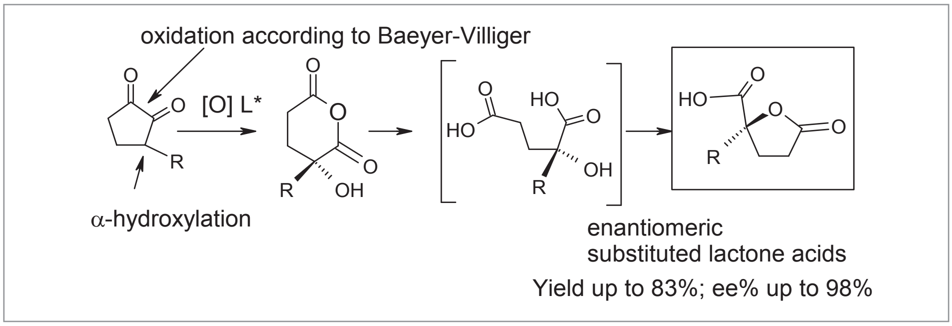
Since then considerable development of organic chemistry has occurred. It is obvious that organic chemistry science is built on quantum mechanics and firmly stands on general principles of physics. In spite of that, most of the new chemical reactions, new chemical structures and new conceptions – the constituencies of modern organic chemistry – are still discovered and rationalised by synthetic chemists, on the findings from synthetic lab. Organic chemical synthesis is still one on the most powerful driving forces in developing organic chemistry.

Organic chemistry uses terms like 'good' and 'bad' (e.g. leaving group); 'hard' and 'soft' (nucleophilic center); 'strong' and 'weak' (base); elegant (synthetic scheme) etc., which belong more to art and literature rather than a real science. We remember the words of Professor R.B. Woodward, who stated in 1956, "There is excitement, adventure, and challenge. And there can be great art in organic synthesis." In a search for new synthetic methodologies synthetic chemist has to balance between antagonistic qualities like 'general/selective' (reaction); 'active/inactive' (reagents); stable/labile (protecting groups), looking for harmony and stability. All that gives to synthetic chemistry science heuristic and intuitive character: it is something that lies on the border of science and art. Chemical synthesis – that is science which is hidden behind the art.

Because of that it is quite obvious that, when speaking about organic synthesis we very frequently think about organic chemistry, although these expressions are not synonyms. Every period in the history of organic synthesis has its mainstream problem exciting most chemists. If we look back to the near past, we may, with some reservation say that, fifties and sixties were the era of sterols; seventies-eighties – era of prostaglandins and other fatty acid metabolites, end of century and beginning of the 21st century – era of complex natural products etc. Synthesis is always developed together with organic chemistry – several outstanding synthetic chemists have been acclaimed as builders of theory of organic chemistry (e.g. Nobel Prize laureate Woodward). Modern ideology of organic synthesis is connected with the name of Nobel Prize laureate Elias Corey. Always, together with the synthesis of certain target molecules, new chemical reactions, new catalyst and reaction principals appear (e.g. Nobel Prize laureates Brown, Noyory, Sharpless, Grubbs, Heck, Negishi, Suzuki, etc.).

Already from eighties of last century the problems of asymmetric chemical synthesis arose, which brought intensive development in selective chemical synthesis methods, asymmetric catalysis, new stereoselective chemical reactions. In Estonia, the problems connected with asymmetry and stereoselectivity arouse together with the synthesis of prostaglandins that started in Tallinn in 1974. From that point the asymmetric oxidation topic starts. In these days we had a problem to transform cyclobutanones to lactones. To solve it, we addressed the Ti-catalysis, which was used to epoxidise allylic alcohols (Katsuki, Sharpless 1980). Indeed, the titanium tartaric ester is able to form with different ketones asymmetric intermediates which induce stereoselectivity of oxidation. By using this catalytic complex we succeeded in performing, among the first chemists, asymmetric Baeyer-Villiger reaction on cyclobutanones (Lopp et al. 1996; Kanger et al. 1998).

Further we turned to other carbonyl-containing substrates and discovered two highly stereoselective new transformation: 2-hydroxylation of 2-hydroxymethyl ketones and oxidative ring cleavage of 1,2-diketones (Paju et al. 2000, 2002, 2003ab). The last cascade reaction merits a special attention for yielding in one step from 1,2-diketones different substituted lactone acids in high enantiomeric purity and good yield (Scheme 1).



Scheme 1.
Asymmetric oxidation of 1,2-diketones.

The easy access to enantiomeric substituted lactone acids makes those compounds very valuable basic intermediates for many target chemical structures. So, on the basis of these compounds we have developed a synthetic route for homocitric acid and 2-alkylsubstituted 2-hydroxyglutaric acid lactones.

Lactone acids merit a special attention also because of being the milestone compound for a technology platform of nucleoside analogue synthesis. This means that by the use of similar starting compounds (substituted 1,2-dikeones), using one basic chemical process (asymmetric chemical oxidation cascade), applying similar universal key compounds (substituted lactone acids) it is possible to obtain a wide variety of nucleoside analogues of interest. Using that general approach a number of synthetic routes for the bioactive nucleoside analogues have been developed (Jõgi et al. 2008, 2009). The general platform enables to synthesise both, cyclic and acyclic nucleoside analogues, which all have been tested for anti-cancer and anti-viral activities.

What is the next challenge? In this respect two main directions may be laid down:

- Using environmentally friendly oxidisers like air oxygen, hydrogen peroxide for asymmetric oxidation.
- Use of metal free catalysts (organocatalysts).
- Finding new asymmetric reactions of 1,2-diketones.

To follow these direction organocatalysis may be the new challenge also for oxidation reactions. Indeed, organocatalysis where simple low molecular weight compounds act as catalysts in chemical bond formation is developing fast: just now is the 'golden era' of it. We have started with organocatalysis in 2005-2006 when Tõnis Kanger introduced new class of organocatalysts – 3,3'-bimorpholines (Kriis et al. 2006; Mosse et al. 2006). These organocatalysts act as HOMO lowering agents for aldehydes and, in the cases of monoammonium salts, also as LUMO lowering through hydrogen bonding. Since then, usefulness of that type of catalysts and new organocatalytic reactions has been demonstrated by T.Kanger's group on many chemical transformations (Kanger et al. 2007; Laars et al. 2009, 2010; Noole et al. 2010) .

Organocatalysis is still fast developing: new classes of substrates (other functionalities in addition to the carbonyl group) and new types of catalysts (in addition to secondary amines) are introduced; new type of reaction (e.g. radical reactions in addition to ionic reactions) is applied.

Nature has built synthetic pathways to obtain structurally diverse natural compounds *via* a relatively small number of transformations, using a small number of key building blocks. The general idea of organocatalysis is to follow the example of Nature and use organocatalysts (and their combinations) as enzymes. This has logically directed the researchers towards cascade reactions where multiple chemical reactions proceed simultaneously. That is a new challenge for chemists dealing with organocatalysis. We have been among the pioneers in this direction: recently we found a new multicomponent cascade reaction, resulting in stereoselective formation of five stereogenic centers in one reaction cascade and leading to aza-bicyclic structure (Kriis et al. 2010). These new structures show potential as agents to influence the neurotransmission system. Biological properties of these structures are presently under investigation.

Another organocatalytic asymmetric aza-ene-type cascade reaction led to highly functionalised 1,4-dihydropyridines in high enantioselectivities and good yields. 1,4-Dihydropyridines are a well-known class of biologically active heterocycles as well as analogues of NADH coenzymes and calcium channel blockers.

Organocatalysis in cascade transformations show good prospects in the synthesis of complex natural compounds (Grondal et al. 2010). It is only a beginning of undiscovered new chemical reactions and it is a long way to understand and to use the secrets of Nature in the synthesis of complex bioactive molecules.

On these selected examples it is obvious that organic synthesis is far of being 'finished' science. There are many new promising directions in developing, and the expectation of the society for new drug candidates as well other synthetic products remain high.

THEORETICAL CHEMISTRY AND MOLECULAR DESIGN

M. Karelson

The power of modern computer technology enables efficient use of the theoretical methods for the prediction of properties of matter. The theoretical approach has proven to be especially beneficial in chemistry and related technological areas, where the experimental study and synthesis of novel compounds and materials can frequently be time consuming, expensive or even hazardous. The contemporary *ab initio* quantum chemical computational methods can, in principle, predict the properties of isolated small molecules with accuracy comparable to the experimental precision. Historically, the first studies using quantum chemistry started in Estonia already in 1970s. The semiempirical methods were used to investigate the dependence of the ionisation potentials and proton affinities of compounds on the structure, later these studies were extended to more advanced nonempirical methods (I. Koppel, U. Mölder). Another direction involved the development of quantum chemical methodology and the respective software to describe the properties of molecules in dense polarisable media. It is well known that the majority of industrially important chemical processes, and all biochemical transformations in living organisms take place in such heterogeneous condensed media. The intermolecular interactions at close distances between molecules in condensed media may lead to significant changes in their electronic structure and subsequently in their physical and chemical properties. The self-consistent reaction field (SCRF) methods developed both at the semi- and nonempirical level of theory enabled to describe robustly the tautomeric behaviour, chemical reactivity and solvatochromic spectral shifts of organic compounds in different media (M.Karelson).

Presently, the quantum chemical methods are widely used both at the University of Tartu and Tallinn University of Technology. One direction involves the studies of the affinities of alkali metal ions, the property that is related to several interesting applications. For instance, the alkali metal cation exchanged zeolites have a great importance in processes where catalytic transformation is not needed, like adsorption and separation. The *ab initio* calculations have revealed the conditions on the binding of gas molecules with Li^+ cation in zeolites (Burk et al. 2009). The calculations have also enabled better understanding of the environmental fate of the major nuclear pollutant ^{137}Cs, its mobility and plant uptake, data on the interactions between Cs and carboxylic acids, which are mainly in anionic forms in soil (Sillar, Burk 2007). The quantum chemical methods have been also applied for the studies of chemical reactivity. A good example is the investigation of the structure and stability of Cu_xS_y-type clusters that serve as models for different copper thiolate clusters or for their intermediates in a variety of important copper proteins. Thermodynamic parameters calculated using density functional theory allowed to assess the effect of thermal contributions to formation energies of various copper thiolate clusters (Ahte et al. 2009). The structures of several carborane acid derivatives (HA) were calculated with DFT B3LYP method and it was established that the gas-phase acidities in these series of neutral acids $CB_{11}H_1H_{11}H$-based clearly depend on field-inductive and resonance effects of the substituent X (Lipping et al. 2009).

Table 1.

Successful QSPR models

Physical properties	Biomedical properties	Technological properties
Boiling points	Toxicity of aqueous pollutants	Glass transition temperature of polymers
Melting points of substituted benzenes	Toxicity of nitrobenzenes	Lithium cation basicities
Melting points of ionic liquids	Genotoxicity	Decarboxylation rates
Refractive index of organic compounds	Activity of α1-Adrenergic receptor antagonists	Charge-transfer constants
Refractive index of polymers	Activity of 5HT1A receptor antagonists	Uranyl complex formation
Viscosity of organic liquids	PDFGR inhibitor activity	Rare earth complex formation
Density of organic liquids	Activity of mosquito repellents	Cyclodextrine complex formation
Dielectric constants	Antimalarial activity (D6 strain)	Gas phase homolysis of C-X bonds
GC retention indices	Antibacterial activity	Rotation activation of amide bonds
GC response factors	Skin permeation	Polarisability of poly-aromatic hydrocarbons and fullerenes
Critical micelle concentrations of surfactants	Blood-brain partition coefficients	
UV spectral intensities	Tissue-air partition coefficients	
Aqueous solubility	Blood-breast milk partition coefficients	
Biphasic partitioning coefficients	Sweetness	
Soil sorption coefficients	Bioconcentration factors of polychlorinated biphenyls	
	HIV Protease inhibition	

The computer-aided molecular design based on quantum-mechanical methods has been suggestive in the development of new polymeric materials with better electrical conductivities. The calculations at different levels of theory with the inclusion of electron correlation effects have provided new information about the details of the geometrical and electronic structure of doped conductive heterocyclic polymers and their behaviour depending on the surrounding media and counterions (Tamm et al. 2005).

The high complexity and extensive size of many important chemical and biochemical systems still prohibits the use of *ab initio* quantum chemical methods and thus the relationship between the chemical and physical properties and the molecular structure in these systems is often poorly described and understood. Thus another approach, the direct development of empirical equations that are commonly referred to as the quantitative structure-property relationships (QSPR) offers an attractive alternative way for the computational prediction of properties of complex molecular systems on the basis of a large variety of molecular descriptors (Karelson 2000). Notably, the QSPR methodology has been widely used in pharmaceutical chemistry for a computer-assisted drug design. In analytical chemistry, QSPR equations are commonly used for the prediction of the spectroscopic, chromatographic and other analytical properties of compounds. In recent years, the QSPR approach has been rapidly expanding to diverse areas of industrial and environmental chemistry. The most useful feature of QSPR is that, in principle, it is applicable for any chemical, physical or biomedical property of chemical compounds, provided that the molecular structure of those are known and that a representative set of experimental data on the property of interest is available. In Table 1, various physical, technological, environmental and biomedical properties of compounds are listed for which successful QSPR models have been developed by groups in Tartu and Tallinn (Katritzky et al. 2010).

CHEMISTRY IN NATIONAL INSTITUTE OF CHEMICAL PHYSICS AND BIOPHYSICS

T. Pehk

The name of National Institute of Chemical Physics and Biophysics (NICPB) makes emphasis on research in physics at the Institute. Nevertheless it is impossible to draw borderlines between the physics, chemistry, biology, technology and other branches of science. It is in the best way illustrated by the work of NMR lab of our Institute, which started nearly 50 years ago by construction of NMR spectrometers and development of NMR spectroscopy for practical physics and chemistry research within the Institute of Cybernetics of the Estonian Academy of Sciences. Our NMR lab has been among the world pioneers of heteronuclear studies and high resolution solid state NMR spectroscopy. NMR lab was the basic unit at birth of our Institute in 1981. We have not only eye-witnessed by our instruments the progress in NMR instrumentation during these years from 40 to 800 MHz in proton resonance frequencies, from field swept to frequency swept and to pulsed Fourier transform (FT) spectrometers, but also given our contribution to the development of NMR spectroscopy.

Built in our laboratory probeheads for solid state studies with ultrafast magic angle spinning rates up to 70 kHz, high temperature probeheads up to 1100K and low temperature probeheads with working temperatures down to 10K give new important possibilities for the research of various materials and molecules, starting from trapped into fullerene cage hydrogen molecules to ceramic electrolytes, from high temperature superconductors to different flours for bread baking, from silicon NMR of nettle leaves to nitrogen NMR of biopolymers (I. Heinmaa). In high resolution NMR multinuclear approach has been always preferred, and main attention has always been targeted to stereochemical applications. Advantages of 13C NMR spectroscopy together with 2D FT correlations in the determination of absolute configurations have been demonstrated. Comparative NMR analytical method for the determination of very small differences in acid-base properties has been developed. Deuterium and even 13C isotope effects on pKa values can be measured. Present instrumentation in the lab corresponds to top world level, and most sophisticated NMR experiments can be performed. Cooperation with all Estonian universities and laboratories abroad from about 20 countries has been going on for decades. NMR results have frequently pointed to new research directions in laboratories of our partners. Cooperation with Estonian chemistry firms has resolved several problems connected with the quality of their production and their R&D problems. Better selectivity and sensitivity of modern NMR instrumentation and our long-time experience form solid basis for all kinds of future cooperation between the chemistry related research institutions and our NMR lab with simultaneously well equipped very strong magnetic field high resolution liquid and solid state instrumentation.

New methods for the determination and investigation of toxicity mechanisms of chemicals and nanosized materials have been developed. Microbial sensor systems inducing bioluminescence as a response to heavy metals including their nanoparticles have been constructed. Important role of nanosized materials in induction of intracellular reactive oxygen radicals has been demonstrated. Up to now research has been focused on ZnO, TiO_2 and CuO nanoparticles due to their increased use in consumer products and lack of relevant ecotoxicological information. Proposed approach was essential point out in recent (2008 and 2009) publications which belong according to Thomson Reuters ESI to past10-year most cited papers of Estonian scientists (A.Kahru, M.Heinlaan, A.Ivask, I.Blinova, V.Aruoja). At 2011 Kahru received the Estonian national science award for the 2007-2010 publications cycle ''Ecotoxicology and mechanisms of toxicity of synthetic nanoparticles". Expansion of research in this field is an important factor towards green environment on Earth.

Biomedical chemistry research in NICPB is based on proteomics studies with particular emphasis on structure-function relationships. Detailed characterisation of snake venoms components (proteases, disintegrins, phospholipases) and their isolation is important part of research in this field. Isolation of new potential thrombolytic and anticancer agents is target of these investigations. Sequencing, synthesis of oligonucleotides, electrophoresis, HPLC and other methods are applied for this research. Constructed and built in laboratory of chemical physics of NICPB unique MALDI-TOFF mass spectrometer is used in these studies (J.Siigur, E.Siigur).

Quantitative HPLC-ES-MS studies in chemical physics laboratory (R. Tanner) are performed on quite complicated mixtures. For example, produced in summer time by cyanobacteria nonribosomal peptides – several individual microcystins in Estonian big lakes have been quantitatively analysed, preparative separation of ^{13}C, ^{15}N, ^{2}H labelled peptides is connected with NMR research. Intensive cooperation with the department of biomedical engineering of Tallinn University of Technology has been going on for many years. Future cooperation in the analysis of metabolic disorders of kidney diseases and food grain fusarium mould toxins are planned.

Oil shale processing is accompanied with atmospheric pollution and solid wastes. Environmental chemistry research in NICPB is connected with the characterisation of hazardous components in fly ashes from different oil shale burning methods. New methods for the practical use of fly ash have been elaborated. Huge quantities of oil shale solid wastes have been produced in Estonia. They are contaminating surface- and ground-water resources with toxic levels of various organic and inorganic compounds. New proposals for cutting down the water contamination from oil shale semi coke have been elaborated (U. Kirso). From the ecological point of view the combustion problems of Estonian oil shale, argillite shales and bio fuels are analysed and their impact on global oxygen balance is continually in the focus of research (A.-T. Pihlak).

New analytical methods are applied for the analysis of Estonian oil shale and argillite shales. The existence of two different types of argillites has been shown and their potential technological utilisation for fuel production and separation of rare elements from them is estimated. New solutions to problems of modern clean energy, such as developing realistic second generation non-food bio fuels are investigated. This involves, for example the zeolite catalysed Fischer-Tropsch and reverse water gas reactions and thermal solution enhanced in situ fuel oil production from shales (E.Lippmaa). Zeolite studies have been important topic in NICPBP for many years by experimental multinuclear NMR studies (P. Sarv) and by theoretical computational studies (A. Trummal).

In inorganic chemistry Solid Oxide Fuel Cell Cerium oxide based electrolyte compositions with different dopants are fabricated and their potential for new effective Fuel Cell production are investigated (J.Subbi).

REFERENCES

Abboud, J.-L. M., et al. 2003. Stereoelectronic, strain, and medium effects on the protonation of cubylamine, a janus-like base. Angewandte Chemie-International Edition, 42, 2281-2284.

Abboud, J.-L. M., et al. 2004. The enormous apparent gas-phase acidity of cubylamine. Chemical Physics Letters, 2004, 398, 560-563.

Ahte, P., Palumaa, P., Tamm, T. 2009. Stability and conformation of polycopper-thiolate clusters studied by density functional approach. Journal of Physical Chemistry A, 113, 9157-9164.

Burk, P., et al. 2008. Feasibility of the spontaneous gas-phase proton transfer equilibria between neutral Brønsted acids and Brønsted bases. Journal of Physical Organic Chemistry, 21, 571-574.

Burk, P., et al. 2009. Computational study of cesium cation interactions with neutral and anionic compounds related to soil organic matter. Journal of Physical Chemistry A, 113, 10734-10744.

Gladysz, J. A. (ed.) 2002. Recoverable Catalysts and Reagents. (Chemical Reviews; 102, 10).

Gorbatsova, J., Jaanus, M., Kaljurand, M. 2009. Digital microfluidic sampler for a portable capillary electropherograph. Analytical Chemistry, 81, 8590-8595.

Grondal, C., Jeanty, M., Enders, D. 2010. Organocatalytic cascade reactions as a new tool in total synthesis. Nature Chemistry, 2, 167-178.

Herrero, R., et al. 2007. The intrinsic(gas-phase)acidities of bridgehead alcohols.An experimental(FT-ICR) and computational study. International Journal of Mass Spectrometry, 267, 302-307.

Ishikawa, T. (ed.) 2009. Superbases for Organic Synthesis. Wiley, Chichester.

Issleib, K., Lischewski, M. 1973. Dimethylaminoiminophosphorane. Synthesis and reactivity in inorganic and metal-organic chemistry. Synthesis and Reactivity in Inorganic and Metal-Organic Chemistry, 3, 255-264.

Jõgi, A., et al. 2008. Asymmetric synthesis of 4'-C-benzyl-2,3-dideoxynucleoside analogues from 3-benzyl-2-hydroxy-2-cyclopentane-1-one. Tetrahedron Asymmetry, 19, 628-634.

Jõgi, A., et al. 2009. Synthesis of 4 '-aryl-2 ',3 '-dideoxynucleoside analogues. Tetrahedron, 65, 2959-2965.

Kagan, H. B. (ed.) 2002. Frontiers in Lanthanide Chemistry. (Chemical Reviews; 102, 6).

Kaljurand, I., et al. 2005. Extension of the self-consistent spectrophotometric basicity scale in acetonitrile to a full span of 28 pK(a) units: Unification of different basicity scales. Journal of Organic Chemistry, 70, 1019-1028.

Kaljurand, I., et al. 2007. Experimental gas-phase basicity scale of superbasic phosphazenes. Journal of Physical Chemistry A, 111, 1245-1250.

Kanger, T., et al. 1998. Asymmetric oxidation of cyclobutanones: Modification of the Sharpless catalyst. Tetrahedron Asymmetry, 4475-4482.

Kanger, T., et al. 2007. Bimorpholine-mediated enantioselective intramolecular and intermolecular aldol condensation. Journal of Organic Chemistry, 72, 5168-5173.

Karelson, M. 2000. Molecular Descriptors in QSAR/QSPR, J. Wiley & Sons, New York.

Katritzky, A. R., et al. 2010. Quantitative correlation of physical and chemical properties with chemical structure: Utility for prediction. Chemical Reviews, 110, 5714-5789.

Katsuki, T., Sharpless, K. B. 1980. The first practical method for asymmetric epoxidation. Journal of the American Chemical Society, 102, 5974-5976.

Kobzar, G., et al. 2009. Effect of vitamin B6 vitamers on platelet aggregation. Platelets, 20, 120-124.

Kobzar, G., Mardla, V., Samel, N. 2011. Short-term exposure of platelets to glucose impairs inhibition of platelet aggregation by cyclooxygenase inhibitors. Platelets, 2011, 22, 338-340.

Koel, M., Kaljurand, M. 2010. Green Analytical Chemistry. Royal Society of Chemistry, Cambridge.

Koljak, R., et al. 1997. Identification of a naturally occurring peroxidase-lipoxygenase fusion protein. Science, 277, 1994-1996.

Kolomeitsev, A. A., et al. 2005. Guanidinophosphazenes: Design, synthesis, and basicity in THF and in the gas phase. Journal of Physical Chemistry A, 2005, 127, 17656-17666.

Koppel, I. A., et al. 1994. The gas-phase acidities of very strong neutral bronsted acids. Journal of the American Chemical Society, 116, 3047-3057.

Koppel, I. A., et al. 2000. Gas-phase acidities of some neutral superacids: A DFT and Ab initio study. Journal of the American Chemical Society, 122, 5114-5123.

Koppel, I. A., et al. 2001. Intrinsic basicities of phosphorus imines and ylides: A theoretical study. Journal of Physical Chemistry A, 105, 9575-9586.

Koppel, I. A., et al. 2002. Generalized principle of designing neutral superstrong bronsted acids. Journal of the American Chemical Society, 124, 5594-5600.

Kriis, K., et al. 2006. Enantioselective synthesis of wieland-miescher ketone through bimorpholine-catalyzed organocatalytic aldol condensation. Synlett, 11, 1699-1702.

Kriis, K., et al. 2010. A novel diastereoselective multicomponent cacade reaction. Organic Letters, 12, 2230-2233.

Kütt, A., et al. 2011. Equilibriuum acidities of superacids. Journal of Organic Chemistry, 76, 391-395.

Laars, M., et al. 2009. Enantioselective organocatalytic Michael Addition of aldehydes to β-nitrostyrenes. Journal of Organic Chemistry, 74, 3772-3775.

Laars, M., et al. 2010. Cyclic amino acid salts as catalysts for the asymmetric Michael reaction. Tetrahedron Asymmetry, 21, 562-565.

Leito, I., et al. 2009. Revision of gas-phase acidity scale below 300 kcal/mol. Journal of Physical Chemistry A, 113, 8421-8424.

Leito, I., et al. 2010. Gas-phase basicities around and below water revisited. Journal of Physical Chemistry A, 114, 10694-10699.

Lipping, L., et al. 2009. Gas-phase brønsted superacidity of some derivatives of monocarba-closo-borates: A computational study. Journal of Physical Chemistry A, 113, 12972-12978.

Looking beyond the Endless Frontier, ESF Forward Looks Scheme: Analysis and Recommendations.
http://www.esf.org/publications.html?tx_ccdamdl_pi1%5Bpointer%5D=3&cHash=5b0ad90c43bcc15c5e28bd6537ec47c1

Lopp, M., et al. 1996. Aasymmetric Baeyer-Villiger oxidation of cyclobutanones. Tetrahedron Letters, 37, 7583-7586.

Lõokene, A., et al. 2005. Apolipoprotein A-V-heparin interactions: implications for plasma lipoprotein metabolism. Journal of Biological Chemistry, 280, 25383-25387.

Martinez, A. W., Phillips, S. T., Whitesides, G. M. 2010. Diagnostics for the developing world: Microfluidic paper-based analytical devices. Analytical Chemistry, 82, 3-10.

Mortimer, M., et al. 2006. Identification and characterization of an arachidonate 11R-lipoxygenase. Archives of Biochemistry and Biophysics, 445, 147-155.

Mosse, S., et al. 2006. 3,3'-bimorpholine derivatives as a new class of organocatalysts for asymmetric Michael Addition. Organic Letters, 8, 2599-2562.

Nilsson, E. C., et al. 2011. The GD1a glycan is a cellular receptor for adenoviruses causing epidemic keratoconjunctivitis. Nature Medicine, 17, 105-109.

Noole, A., et al. 2010. Enantioselective Henry reaction catalyzed by CuII salt and bipiperidine. Journal of Organic Chemistry, 75, 1313-1316.

Olah, G. A., et al. 2009. Superacid Chemistry. Wiley, New York.

Paju, A., et al. 2000. Asymmetric oxidation of 1,2-cyclopentanediones. Tetrahedron Letters, 6883-6887.

Paju, A., et al. 2002. Asymmetric oxidation of 3-alkyl-1,2-cyclopentanediones. Part 1, 3-hydroxylation of 3-alkyl-1,2-cyclopentanediones. Tetrahedron Asymmetry, 2439-2448.

Paju, A., et al. 2003a. Asymmetric oxidation of 3-alkyl-1,2-cyclopentanediones. Part 2, Oxidative ring clevage of 3-alkyl-1,2-cyclopentanediones: synthesis of 2-alkyl-γ-lactone acids. Tetrahedron Asymmetry, 1565-1573.

Paju, A., et al 2003b. Asymmetric oxidation of 3-alkyl-1,2-cyclopentanediones. Part 3. Oxidative ring clevage of 3-alkyl-1,2-cyclopentanediones: synthesis of α-hydroxy -γ-lactone acids and spiro-γ-dilactones. Tetrahedron Asymmetry, 2393-2399.

Palmre, V., et al. 2011. Electroactive polymer actuators with carbon aerogel electrodes. Journal of Materials Chemistry, 21, 2577-2583.

Raynie, D. E. 2011. Book review of green analytical chemistry. Journal of the American Chemical Society, 133, 1146.

Rõõm, E.-I., et al. 2003. Acid-base equilibria in nonpolar media. 3. Expanding the spectrophotometric acidity scale in heptane. Journal of Organic Chemistry, 68, 7795-7799.

Rõõm, I., et al. 2007. Brønsted basicities of diamines in the gas phase, acetonitrile, and THF. Chemistry – A European Journal, 27, 7631-7643.

Schwesinger, R., et al. 1996. Extremely strong, uncharged auxiliary bases; monomeric and polymer-supported polyaminophosphazenes (P2–P5). Liebigs Annalen, 7, 1055-1081.

Sillar, K., Burk, P. 2007. Adsorption of carbon monoxide on Li-ZSM-5: theoretical study of complexation of Li+ cation with two CO molecules. Physical Chemistry Chemical Physics, 9, 824-827.

Stoyanov, E., Kim, K.-C., Reed, C. A. 2006. An infrared vNH scale for weakly basic anions. Implications for single-molecule acidity and superacidity. Journal of the American Chemical Society, 128, 8500-8508.

Sukonina, V., et al. 2006. Angiopoietin-like protein4 converts lipoprotein lipase to inactive monomers and modulates lipase activity in adipose tissue. Proceedings of the National Academy of Sciences of the USA, 103, 17450-17455.

Tamm, T., Tamm, J., Karelson, M. 2005. Complexes of oligopyrrole dications with inorganic anions: a comparative theoretical HF/post-HF study. Synthetic Metals, 149, 47-52.

Tamp, S., et al. 2008. A total conformational analysis of diastereomeric esters and calculation of their conformational shielding models. Journal of Molecular Structure: THEOCHEM, 851, 84-91.

Valmsen, K., et al. 2001. 15R-cyclooxygenase of invertebrate origin: Molecular cloning and expression from the caribbean coral *Plexaura homomalla*. Proceedings of the National Academy of Sciences of the USA, 98, 7700-7705.

Valmsen, K., et al. 2007. A critical role of non-active site residues on cyclooxygenase helices 5 and 6 in the control of prostaglandin stereochemistry at carbon 15. Journal of Biological Chemistry, 282, 28157-28163.

Villo, L., et al. 2007. Synthesis of deoxy sugar esters: A chemoenzymatic stereoselective approach affording deoxy sugar derivatives also in the form of aldehyde. Journal of Organic Chemistry, 72, 5813-5816.

Wakihara, M., Yamamoto, O. (eds.) 1998. Lithium Ion Batteries, Fundamentals and Performance. Wiley & WCH, Weinheim.

NANOTOXICOLOGY: SCIENCE AT THE INTERFACES. ESTONIAN PERSPECTIVE

A. Kahru, A. Ivask, K. Kasemets, I. Blinova

Nanotoxicology and nanoecotoxicology are research fields studying the safety aspects of nanotechnologies at human and ecological levels, respectively. Estonian scientists have contributed remarkably into studies concerning ecotoxicological effects and mechanisms of toxic action of engineered nanoparticles, especially metal oxides (e.g., TiO_2, ZnO, CuO).

INTRODUCTION

According to the definition by US National Nanotechnology Initiative (NNI; www.nano.gov) nanotechnology is science, engineering, and technology conducted at the nanoscale (1-100 nm). As a growing applied science, nanotechnology has considerable global socioeconomic value, and the benefits afforded by nanoscale materials and processes are expected to have significant impacts on almost all industries and areas of society. By 2015, the nanotechnology economy is estimated to be valued at 2.2 trillion Eur (Lux Research 2008).

Currently, in the conditions of the worldwide global economic recession, exponential growth of population, shortage of food, feed, fuel and raw materials and increasing environmental and societal problems, nanotechnologies have big expectations in almost every domain, from energy production to medicine. Moreover, nanotechnology has been referred to as the next industrial revolution (Kennell 2009). Therefore, dealing with uncertain risks from engineered nanomaterials to human health and the environment is an important, critical and challenging task given that some properties of the materials at nanoscale may present harm to human health and the environment. European Academies Science Advisory Council (EASAC) and EC Joint Research Centre (JRC) are currently finalising a joint report (to be available at www.easac.eu) entitled “Impact of engineered nanomaterials on health: considerations for benefit-risk assessment” (EASAC-JRC Report 2011) describing the state-of-the art knowledge on the safety aspects of engineered nanomaterials and identifying needs for further scientific investigations. According to this report there are big knowledge gaps in the safety aspects of engineered nanomaterials and need for further scientific investigations. The lack of adequate safety information may ultimately counteract any preliminary gains and hamper the sustainable development of nanotechnologies.

NANOPARTICLES: PRODUCTION, EXPOSURE, PROPERTIES AND POTENTIAL ADVERSE EFFECTS ON HUMANS AND THE ENVIRONMENT

Man-made or synthetic or engineered NPs (eNPs), i.e., particles with at least one dimension < 100 nm can be metals and metal oxides (e.g., nAg, nAu, TiO_2, ZnO, CeO_2), metal salts like CdS or CdSe (quantum dots) or carbon-based (e.g., C60-fullerenes, single-wall and multi-wall carbon nanotubes or dendrimers). In addition to synthetic NPs, there are also natural NPs that are by-products of various processes (exudation, weathering, combustion).

At nanoscale the materials have different or enhanced properties compared with the same ones at larger size, i.e., their conventional 'bulk' (micro-size) counterparts, due to an increased relative surface area that translates into higher reactivity: 1 nm NP would have ~76% of its atoms on the surface. In addition, the NPs tunable optical and electronic properties, as well as modulatable mechanical properties are beneficial for applications in composite structural materials. The changed physico-chemical properties, however, may lead to increased bioavailability and toxicity, mostly due to greater number of reactive groups on the particle surface (Nel et al. 2006). By the year 2009 the number of consumer products containing NPs/materials (mostly nanosilver) on the market already exceeded 1000 (www.nanotechproject.org).

In sunscreens NPs of TiO_2 are used to block UV but still look transparent on the skin. Nanosilver particles are added to various products as antimicrobials. Nano-solar cells are expected to increase efficiencies for transfer of energy. Carbon nanotubes are 100-times stronger than steel and used in tennis rackets and hockey sticks. The above given are just few examples. The nanotechnologies also hold great promise for medical therapies. As current cancer drugs are poorly target-specific, NPs can have specific compounds attached to them that recognise unique receptors on certain malignant cells. The world's first nanotechnological cancer drug is ABRAXANE® – a human-albumin-bound paclitaxel. Cancer therapy based on hyperthermia *via* localised heating by gold NPs has been suggested. Fullerenes can be useful for medical applications as they allow addition of chemical groups in 3D-orientation and nanostructured hydroxyapatite as bone substitute. Remarkable progress is also expected in early diagnostics: iron NPs can be loaded with imaging contrast agents and quantum dot markers for diagnostics are expected to be available in the near future (Kennell 2009).

> The greatest impact of nanotechnologies is from the materials and manufacturing sector, e.g., coatings and composites for cars and buildings, followed by electronics, e.g., displays and batteries and, finally, healthcare, driven by pharmaceutical applications (Lux Research 2008).

NANOPARTICLES AND NANOMATERIALS: BENEFIT-RISK ASSESSMENT

Given the investments to the nanotechnologies are already remarkable, the likelihood of occupational and environmental contact increases (Fig. 1).

Already at 2004 a large insurance company Swiss Re released a report “Nanotechnology ‑ Small matter, many unknowns” and the same year Royal Society and Royal Academy of Engineering published a report “Nanoscience and nanotechnologies: opportunities and uncertainties”. At 2007 EU Scientific Committee on Emerging and Newly Identified Health Risks (SCENIRH) issued a document on the risk assessment of nanomaterials. All these reports unanimously agreed on necessity of rigorous analysis of benefits and risks, to guarantee the sustainability of nanotechnologies. And our first defence against concerns toward nanotechnologies is relevant scientific knowledge (Barnard 2006).

> As long as we do not have enough scientific information on eNPs enabling proper hazard evaluation, the precautionary principle should apply. Better safe than sorry.

Already at 2004 a large insurance company Swiss Re released a report “Nanotechnology – Small matter, many unknowns” and the same year Royal Society and Royal Academy of Engineering published a report “Nanoscience and nanotechnologies: opportunities and uncertainties”. At 2007 EU Scientific Committee on Emerging and Newly Identified Health Risks (SCENIRH) issued a document on the risk assessment of nanomaterials. All these reports unanimously agreed on necessity of rigorous analysis of benefits and risks, to guarantee the sustainability of nanotechnologies. And our first defence against concerns toward nanotechnologies is relevant scientific knowledge (Barnard 2006).

As long as we do not have enough scientific information on eNPs enabling proper hazard evaluation, the precautionary principle should apply. Better safe than sorry.

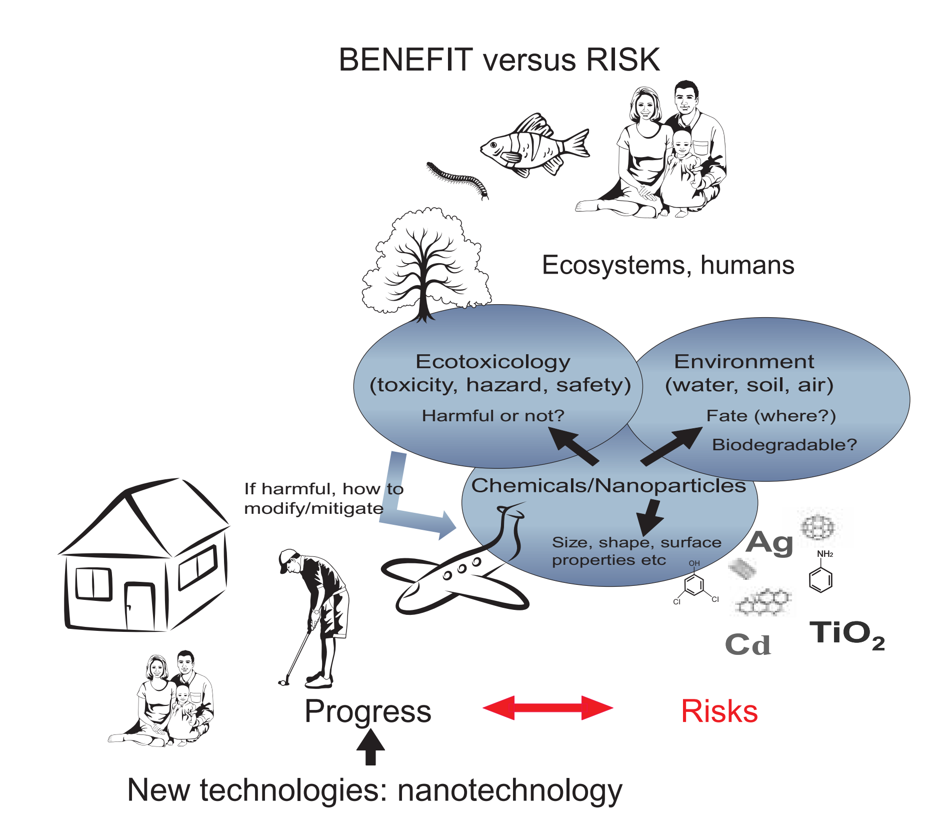


Figure 1.
Sustainability of a new technology is based on integrated development of technological breakthrough and analysis of its long-term effects on health and environment.

NANOTOXICOLOGY & NANOECOTOXICOLOGY – INTERDISCIPLINARY SCIENCES

Nanotoxicology (Oberdörster et al. 2005) is a new interdisciplinary field of research that emerged in the early 1990s and studies toxicological effects, risk assessment, and safety evaluation of nano-structured materials on human health (Maynard et al. 2011; Zhao, Nalwa 2006) (Fig. 2).

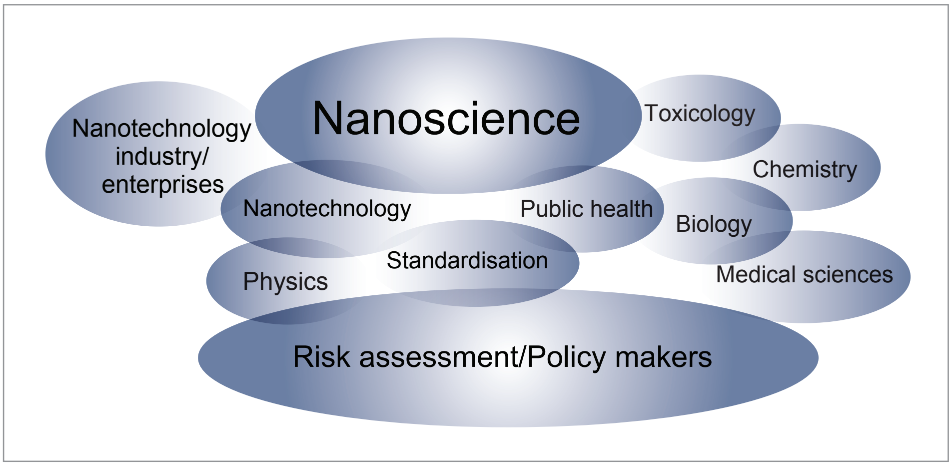


Figure 2.
Nanotoxicology – an interdisciplinary science.

In the human body there are three main potential routes of entry of (synthetic) NPs: the lung, the skin and gastro-intestinal (GI) tract. The first two exposure routes are prevailing in occupational settings and thus mostly studied. The body-entered NPs may further translocate *via* bloodstream, penetrate through cell membranes and accumulate in organs.

The intact skin is probably resistant to eNPs but damaged or sunburned skin may lead to the uptake of NPs and their translocation. Carbon nanotubes have physical characteristics similar to asbestos fibres and thus, their effects are mainly connected with respiratory tract. Carbon nanotubes can penetrate into the lungs avoiding inherent clearance mechanisms and thus, inducing cell malformations (Poland et al. 2008). Currently there is already remarkable toxicological information on effects of various eNPs obtained mostly on *in vitro* cell cultures (see below).

In fact all kinds of eNPs produced in high amount enough (as all industrial chemicals!) will inevitably end up in the environment *via* various industrial and household waste streams and thus may pose hazard at all levels of biological food-chain. For example TiO_2 or ZnO NPs applied in sunscreens may not only penetrate the human's sunburned skin epidermal cells, but can also be washed away from the skin ending up in the waterbodies and affecting aquatic ecosystems. When the whole ecosystem is concerned, one has to understand that while risk assessment for human health concerns one species, environmental risk assessment should ideally consider millions of species, with different morphology, physiology, and ecological habitats (Fig. 1). The adverse effects of chemicals/pollutants on the level of the ecosystem are studied by ecotoxicology that emerged in 1970s with increasing awareness about contaminants in the biosphere and their adverse effects on ecosystem health. Currently, ecotoxicology has become an important part in environmental and ecological risk assessment. Development of ecotoxicology was severely lagging behind toxicology. The same is true for nanoecotoxicology: the first papers on ecotoxicological effects of NPs appeared just in 2006) (Fig. 3) (see also Kahru, Dubourguier 2010).

Current scientific information on different aspects of nanoparticles: share of 'nano'safety information

As the number of published scientific papers is the early indicator for the development of the field, a bibliometric survey in various databases was made. In general three keywords were used: 'nanoparticles' to cover all research connected with nanotechnologies that was combined with 'toxic*' reseor 'ecotoxic*', to search for the volume of human- or environment-related papers on potential hazardous effects of NPs (Box 1). The obtained data were analysed either by subject area, publication time or country-wise.

Fig. 3 shows that the papers on 'nanoparticles' started to appear at 1999. Currently, there are >100 000 papers registered in ISI WoS and Scopus on this keyword: about 1/3 concerning structure and properties of NPs. Due to the priorities of nano-medical research, there is already substantial information on toxic effects of NPs: about 4500 papers, i.e. 4%. At the same time the total number of papers on safety aspects on NPs was only about 1000. Even more striking is the shortage of scientific information on environmental hazard of eNPs – just a few hundred papers published till now appearing since 2006 (Fig. 3).

Box 1. Keywords and databases used for the bibliometry

Keywords: nanoparticles, toxic*, ecotoxic*
Bibliographic databases used:
Thomson Reuters ISI Web of Science (ISI WoS)
(http://apps.isiknowledge.com/)
Thomson Reuters ISI Essential Science Indicators (ISI ESI)
(http://esi.isiknowledge.com/home.cgi) – updated as of May 1, 2011 to cover a 10-year period (2001-2011)
Scopus (now officially named SciVerse Scopus) (www.scopus.com)
SJR – SCImago Journal & Country Rank; http://www.scimagojr.com

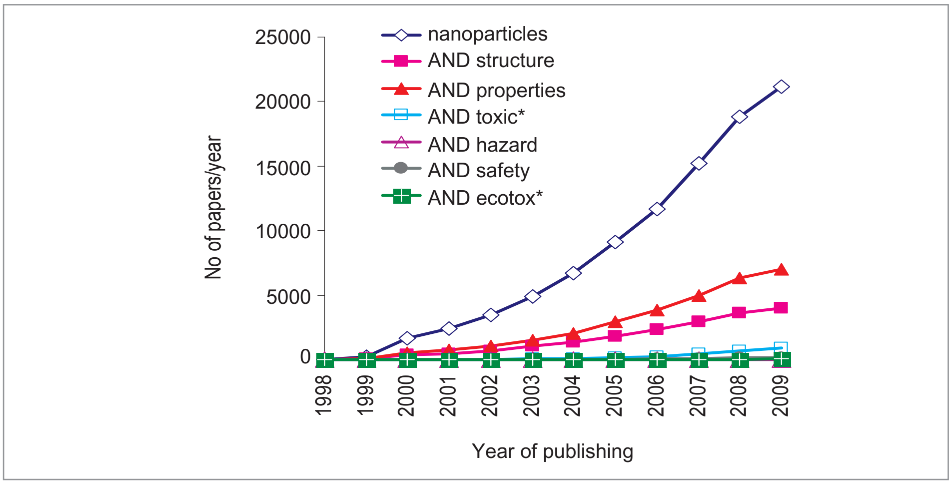


Figure 3.
Number of publications (per year) registered in ISI Web of Science.
(Timespan=1980-2009. Refined by: Document Type=(ARTICLE OR PROCEEDINGS PAPER OR REVIEW). Total nr of papers on key-word 'nanoparticles' : 96 474. Search made: Feb 1, 2009.)

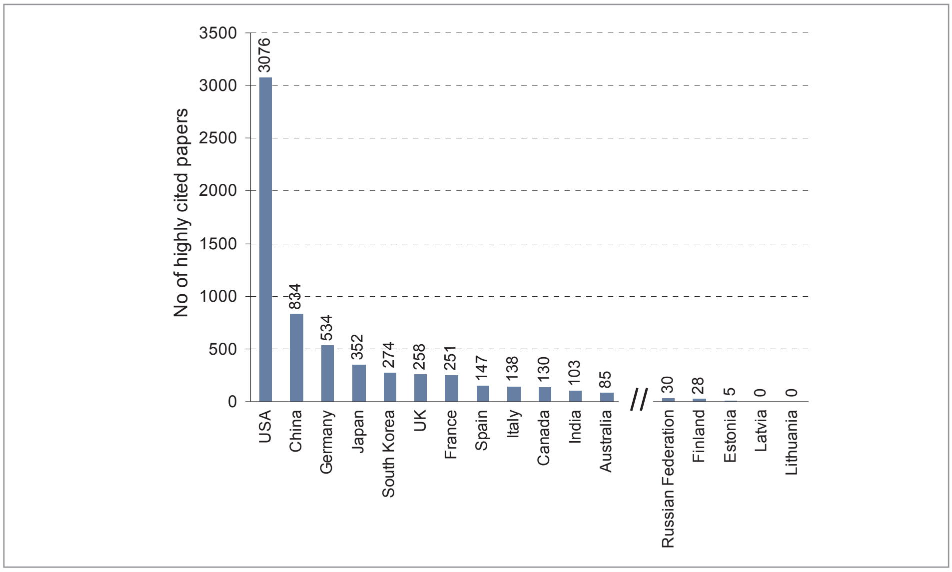


Figure. 4.
Number of highly cited papers for different countries*, 2001-2011, in Thomson Reuters ESI (keyword nano*). Total number of the highly cited papers on 'nano*' in the ISI ESI was 5 919. Search made: 16.05.2011.

* Countries from the USA till Australia were all within top 12 in Thomson Reuters ESI (1991-2000) for a Special topic "Nanotechnology" (http://esi-topics.com/nano/nations/d1b.html)

Search in ISI ESI for the number of highly cited papers in past 10 years (2001–2011) across all fields on the keyword 'nano*' for the countries that were about decade back in top 12 in Thomson Reuters ESI (1991–2000) for a Special topic "Nanotechnology" (http://esi-topics.com/nano/nations/d1b.html) showed that the outstanding leader by highly cited papers as by 2011 on keyword 'nano*' is US (>3000 papers), followed by China, Germany and Japan. In addition to these top 12 countries we also checked information for some other countries: Russia had currently 30 highly cited 'nano' papers, Finland 28, Estonia 5 and both Latvia and Lithuania none (Fig. 4). The Estonian research into nanotechnological topics is described in more detail below.

GLOBAL NANORACE & SAFETY ASPECTS OF NANOTECHNOLOGIES: STRATEGIC APPROACHES

By 2007, more than 60 countries had started national nanotechnology programs. Levels of funding in nanotechnology R&D in 2008 reached 13 billion Euros worldwide, with the US and Japan leading in this activity (http://www.bricnanotech.com/nanotechnology/). The benchmark event of the start of the global nanorace is generally considered the launch of the National Nanotechnology Initiative, NNI (www.nano.gov) in the US in 2001.

> The vision of the US NNI:
> Understanding & controlling matter at nanoscale → revolution in technology/industry (benefits for the Society)

The 2012 US Federal Budget provides for the NNI $2.1 billion (1.5 billion Eur). Major projects and activities in NNI are divided to 8 program component areas (PCAs). Funding of PCA7 "Environment, Health, and Safety" (EHS) at 2010 was about 5% of the total NNI budget and main contributors being National Science Foundation (NSF) (30%), National Institute of Health (NIH) (22%) and Environmental Protection Agency (EPA) (19%). As EHS has currently relatively small funding share, the NNI also continues to increase its investments aimed at implementing the Government's strategy for nano EHS research.

In Europe, the strategic document "Nanosciences and Nanotechnologies: An Action Plan for Europe 2005–2009" was published in 2005 for the "immediate implementation of a safe, integrated and responsible strategy for nanosciences and nanotechnologies". There are 7 key elements in this Plan, including “Public health, safety, environmental and consumer protection” (http://cordis.europa.eu/nanotechnology/actionplan.htm). Nanoresearch is also important in FP7 (2007–2013): 7% of the total FP7 is budgeted for NMP (Nanosciences, nanotechnologies, materials and new production technologies), i.e., altogether 3.5 billion Euros.

Contribution of regulatory frameworks to safety assessment of nanomaterials

The Organization for Economic Co-operation and Development (OECD) is a regulatory body that has published the set of most relevant internationally agreed testing methods to assess the safety of chemicals (in http://www.oecd.org/env/). Although it is recognised worldwide that regulation and labelling of nanoproducts is needed, there is currently no international regulation of nanoproducts or nanotechnology and no agreed protocols for toxicity testing or evaluating the environmental impacts of NPs. However, according to OECD most of the toxicity testing guidelines are suitable also for NPs (http://www.oecd.org/dataoecd/6/25/47104296.pdf). At 2006 OECD formed the Working Party on Manufactured Nanomaterials (WPMN) that defined a list of 14 representative manufactured nanomaterials: silver, iron, TiO_2, aluminium oxide, cerium oxide, zinc oxide, silicon dioxide, nanoclays, carbon black, fullerenes (C60), single-walled carbon nanotubes (SWCNTs), multiwalled carbon nanotubes (MWCNTs), polystyrene, dendrimers.

In US, the Environmental Protection Agency (EPA) and Food and Drug Administration (FDA) and in Europe the Health & Consumer Protection Directorate have already started dealing with the potential risks of NPs. The European Food Safety Authority (EFSA) launched in 2011 a public consultation on a draft guidance on risk assessment concerning potential risks arising from applications of nanoscience and nanotechnologies in the area of food and feed.

For regulating industrial chemicals in EU, a new chemical policy Registration, Evaluation, Authorisation and restriction of Chemicals (REACH) came into force at 2007. REACH requires that all substances on the European market, which are manufactured or imported in quantities of 1 tonne or more per year have to be evaluated for hazardous effects on humans and environment by the year 2018. Although REACH is not referring specifically to nanomaterials, this regulation deals with substances, in whatever size, shape or physical state. Thus, it follows that under REACH manufacturers, importers and downstream users have to ensure that their nanomaterials do not adversely affect human health or the environment.

Toxicological research in Estonia – more than 50 years old

The history of (industrial) toxicology in Estonia dates back to the 1950s. The first toxicological experiments were made in 1951 at the Institute of Experimental and Clinical Medicine (IECM; now restructured into National Institute for Health Development - NIHD) and concerned safety aspects of Estonian oil-shale, oil-shale phenols and different derivatives such as adhesives, resins, mastics, softeners and solvents. During 1997–1999 the IECM participated in the EU project "Prevention of asbestos-related diseases in Hungary, Estonia and Karelian Republic of the Russian Federation". Lead by T.Veidebaum, the research in asbestos-related occupational hazards, diagnostics of occupational cancers and health risks related to asbestos has been continued at NIHD. At present, toxicological research in Estonia is carried out mainly in the groups lead by A. Zharkovsky (Tartu University), M. Karelson (University of Tartu/Tallinn University of Technology), A. Kahru and J. Siigur (National Institute of Chemical Physics and Biophysics, Tallinn).

Notably, in the Estonian libraries a large collection of the Russian language toxicological literature has been preserved. Within the FP6 project OSIRIS (www.osiris-reach.eu) this Russian language toxicological information was collected by NICPB researchers into a web-database E-SOVTOX (http://kbfi-databases.eu/database) (Sihtmäe et al. 2009; Heinlaan et al. 2008).

The Estonian Society of Toxicology

The Estonian Society of Toxicology (ETS; www.kbfi.ee/ets) was created in 1997 and has 55 members, mostly scientists and students dealing with chemical safety, occupational health and environmental risk assessment. ETS has organised several national seminars and international conferences and organised several seminars for various Estonian stakeholders e.g., on Good Laboratory Practice and new chemical regulation REACH. During 2003-2004 ETS was involved in the Btox program initiated by T. Malmfors and University of Uppsala, Sweden, to promote the toxicology education in Estonia, Latvia and Lithuania. Currently (2011) ETS is creating web-available teaching materials "Basics of environmental toxicology" (Keskkonnatoksikoloogia alused) in the Estonian language. The project is funded by Estonian Environmental Investment Centre. ETS is a member of EUROTOX (Federation of European Toxicologists and European Societies of Toxicology) and IUTOX (International Union of Toxicology). The Chair-person of ETS since 1997 has been Dr. Anne Kahru.

Nanotechnological research in Estonia: from 'nano' to 'bio'

Estonian scientists studied nanomaterials already in early 1980s. As an example, the work on structural studies of silicates using solid-state high-resolution NMR published in "J. American Chemical Society" by Endel Lippmaa and his co-workers (JACS 1980, 102: 4889-4893; 1981, 103: 4992-2996) belonged to the most cited chemistry papers worldwide and have been cited by now > 1000 times. With some reservation, E.Lippmaa's and his co-workers' (Ago Samoson, Märt Mägi) research in National Institute of Chemical Physics (NICPB) can be considered as a start of nanomaterials science in Estonia.

Box 2.
10-year highly cited papers in Thomson Reuters Essential Science Indicators (ESI) for Estonia. Searched across all fields using keyword 'nano*'

Results of the search (5 papers) are presented in the order and form as they appeared in ESI (16.05.2011). Authors with Estonian affiliation are in italic, asterisk marks the corresponding author.

1**.** *Heinlaan, M., Ivask, A., Blinova, I., Dubourguier, H. C., Kahru, A.** Toxicity of nanosized and bulk ZnO, CuO and TiO_2 to bacteria *Vibrio fischeri* and crustaceans *Daphnia magna* and *Thamnocephalus platyurus*. Chemosphere, 71, 7, 1308-1316.

See also (http://sciencewatch.com/dr/fbp/2009/09augfbp/09augfbpKahr/).

2. Winkler, P. M., Steiner, G., Vrtala, A., Vehkamaki, H., *Noppel, M.,* Lehtinen, K. E. J., Reischl, G. P., Wagner, P. E., Kulmala, M.* Heterogeneous nucleation experiments bridging the scale from molecular ion clusters to nanoparticles. Science, 2008, 319, 5868, 1374-1377.

3. *Aruoja, V., Dubourguier, H. C.*, Kasemets, K., Kahru, A***.** Toxicity of nanoparticles of CuO, ZnO and TiO_2 to microalgae *Pseudokirchneriella subcapitata.* Science of the Total Environment, 2009, 407, 4, 1461-1468.

4. *Kahru, A.*, Dubourguier, H.C., Blinova, I., Ivask, A., Kasemets, K.* Biotests and biosensors for ecotoxicology of metal oxide nanoparticles: A minireview. Sensors, 2008, 8, 8, 5153-5170.

5. Ji, X. B., Kadara, R. O., *Krussma, J***.**, Chen, Q. Y., Banks, C. E.* Understanding the physicoelectrochemical properties of carbon nanotubes: Current state of the art. Electroanalysis, 2010, 22, 7-19.

As of May 16, 2011, in the Estonian R&D database ETIS (www.etis.ee) the number of ISI WoS papers (classification 1.1) was 20 533 and within these publications 272 papers were on 'nano'. The work in the Estonian Universities concerns mostly materials science. Notably, one of the eight Competence Centres in Estonia is dedicated to nanoresearch from materials science perspective. This Estonian Nanotechnology Competence Centre, ENCC (www.encc.ee) is a consortium of industrial and science partners, the scientific leader partner being Institute of Physics at the University of Tartu. The biological effects of synthetic NPs and nanosafety aspects are studied at NICPB.

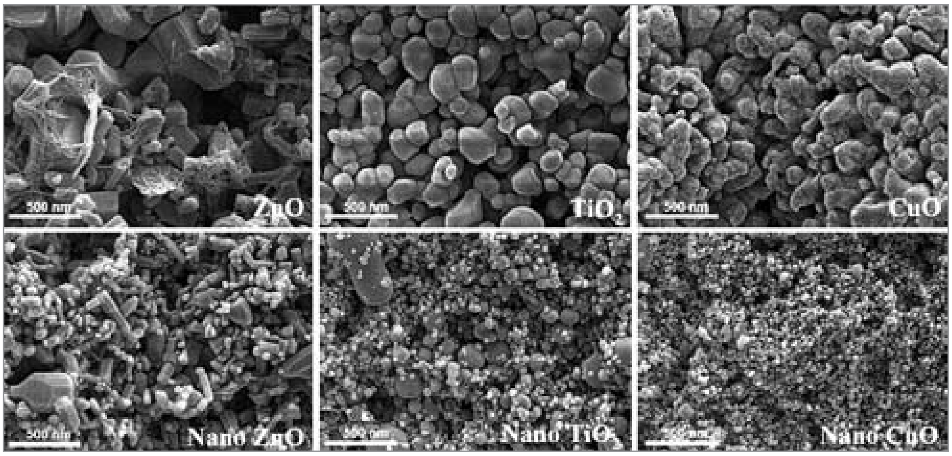
In Thomson Reuters ESI there are currently (years 2001–2011) five highly cited papers for Estonia retrieved by keyword 'nano*': three of them originated from NICPB and two were co-authored by scientists from the University of Tartu (Box 2). One of these papers (Heinlaan et al. 2008) was also rated by ESI as a 'fast breaking paper' and A.Kahru was the first Estonian scientist to be interviewed by Thomson Reuters Science Watch (http://sciencewatch.com/dr/fbp/2009/09augfbp/09augfbpKahr/). This paper has by now 107 citations in ISI WoS.

Nano(eco)toxicological research at National Institute of Chemical Physics (NICPB), Estonia

As shown above, a remarkable part of Estonian most cited 'nano'research concerns safety aspects of nanotechnologies and originates from NICPB, Group of In Vitro Toxicology and Ecotoxicology (currently equal to Laboratory of Molecular Genetics) lead by Anne Kahru. The ecotoxicological research of this group started in 1990s and was initially focused on pure chemicals followed by oil-shale industry-related pollutants (years 1993–2000), heavy metals (2000–2006) and since 2006, synthetic nanoparticles. The research started from TiO_2, ZnO and CuO NPs but also nano-silver and carbon-based NPs (fullerenes, dendrimers, MWCNTs) were studied. NPs analyses for toxic effects were also characterised by physico-chemical properties: size by transmission electron microscopy (TEM) and scanning electron microscope (SEM; Fig. 5A) and the specific surface area of powders by BET analysis. Hydrodynamic size and zeta-potential of NPs in different test media was characterised using Malvern Zetasizer Nano ZS (Fig. 5B). Most of NPs tend to agglomerate in aqueous media and the size of agglomerates depends on various factors.

Given eNPs are already produced in high production volumes without adequate scientific knowledge on their action on biota and ecosystems, there is an urgent need to fill the gaps in several areas addressing:

- toxicity of eNPs on key model organisms of different biological food-chain levels;



A

B

Figure 5.
Scanning electron microscopy of ZnO, TiO_2 and CuO particles: nanosized oxides on lower panels of part A and respective micro-sized or 'bulk' formulations on upper panels of part A (Kahru et al. 2008); Malvern Zetasizer Nano ZS (B).

- mechanisms of action of eNPs on different types of organisms;
- working out relevant (eco)toxicological and *in vitro* test formats for eNPs;
- need for integrated testing strategies;
- fate of eNPs in the environment.

In accordance with the former knowledge gaps, the nano(eco)toxicological research at NICPB has proceeded in four major directions :

- to reveal the dose-dependent effect of eNPs on different representatives of aquatic food chain: primary producers (algae), consumers (daphnids, protozoa) and destructors (bacteria) and establish the L(E)C50 (half-lethal concentration or half-effective concentration), NOEC (no-observed-effect concentration), MIC (minimal inhibitory concentration) values necessary for the environmental risk assessment of eNPs;
- to adapt and modify the existing *in vitro* and ecotoxicological test methods (i) for the analysis of synthetic NPs that are intrinsically difficult to study for their aquatic toxicity as these substances are

partially insoluble, agglomerating, settling, sorbing nutrients/viability dyes etc and (ii) towards increase of sample throughput, to increase the cost-efficiency and speed of testing;

- to study the mechanisms of toxicity of different NPs on organisms of different complexity (prokaryotic, eukaryotic, particle-ingesting and not ingesting) on different aquatic ecosystem food-chain level;
- to develop new biological screens for mechanistic studies of NPs, i.e., gene-modified sensor bacteria responding specifically to certain signals from the environment, such as bioavailable heavy metals, DNA-modifying chemicals, reactive oxygen species producing chemicals. Most of these screens are versatile also for mechanistic studies of 'conventional' chemicals.

Model organisms used were (i) from OECD and/or German Institute for Standardization (DIN) ecotoxicological tests, e.g., crustaceas *Daphnia magna* and bacteria *Vibrio fischeri*, algae *Selenastrum capricornutum = Pseudokirchneriella subcapitata,* in order to use the knowledge obtained also for the regulatory purposes; (ii) other particle-feeding organisms: crustaceans *Thamnocephalus platyurus* and protozoa *Tetrahymena thermophila*; (iii) versatile eukaryotic unicellular model organisms for mechanistic *in vitro* toxicology: yeast *Saccharomyces cerevisiae* with full library of single-gene knock-out strains available (EUROpean Saccharomyces Cerevisiae ARchive for Functional analysis, EUROSCARF) and protozoa *Tetrahymena thermophila* that has been widely used for testing of toxic effects of industrial chemicals and thus one of the main organisms used for QSAR-predictions but also for various mechanistic cell-biology investigations.

The results are presented in papers (Heinlaan et al. 2008, 2011; Kahru et al. 2008; Kasemets et al, 2009; Aruoja et al. 2009; Blinova et al. 2010; Mortimer et al. 2010, 2011; Ivask et al. 2010) and summarised in Review by Kahru and Dubourguier (2010). Briefly, TiO_2 NPs (no specific photoactivation applied) were not toxic on test organisms used (see above) even at 20 000 mg/l, with the exception of algae *Pseudokirchneriella subcapitata* (72 h EC50=9.7 mg/l). Entrapment of algae cells by agglomerates of TiO_2 in the test medium could be the one reason for the observed toxic effect (Aruoja et al. 2009). ZnO and CuO NPs were remarkably more toxic: one of the first in the world, we showed that for some aquatic organisms such as *algae P. subcapitata* and crustaceans *Thamnocephalus platyurus* ZnO NPs and for algae also CuO NPs were toxic already at sub ppm level (<1 mg/l) (Heinlaan et al. 2008; Aruoja et al. 2009) that refers to their high ecotoxicological hazard.

For mechanistic studies, we have developed a novel combined bioanalytical approach that allows for the profiling of toxicological properties of metal oxide NPs. For that, the traditional medium to high-throughput *in vitro* assays with pro- and eukaryotic organisms were applied to compare the toxicity of NPs to that of their 'bulk' counterparts. In parallel the contribution of dissolved metal ions to the overall toxicity of NPs was detected by gene-modified metal-specific biosensors (Ivask et al. 2009) (and also by analytical chemistry techniques) (Kahru, Dubourguier 2010; Heinlaan et al. 2008, 2011; Kahru et al. 2008; Kasemets et al. 2009; Aruoja et al. 2009; Blinova et al. 2010; Mortimer et al, 2010, 2011; Ivask et al. 2010). In most cases the toxic effect of CuO and ZnO NPs was caused by dissolved Cu- and Zn-ions. However, toxicity of CuO NPs to yeast *Saccharomyces cerevisiae* and crustaceans *Daphnia magna* was caused also by the oxidative stress and immune system imbalance, respectively (Kasemets et al. 2009; Heinlaan et al. 2011). The latter aspect was studied using TEM-visualisation of the sections of nCuO exposed gut of *D. magna*: we detected changes in gut epithelial ultrastructure and observed massive presence of bacteria in the gut (referring to immunological imbalance) but fortunately did not observe any internalisation of CuO NPs by gut epithelial cells (Heinlaan et al. 2011). However, by studying the membranes of protozoa *Tetrahymena thermophila* upon exposure to nCuO we observed changes in their fatty acid composition towards increase of rigidity, probably *via* inhibition of respective desaturases (Mortimer et al. 2011).

New test formats and approaches for the *in vitro* toxicity screening of NPs and mechanisms of toxicity were developed:

- the panel comprising of 13 recombinant sensor bacteria (Ivask et al. 2009) and oxidative-stress sensitive mutated luminescent strains (Ivask et al. 2010) have been constructed or adapted in our lab and used for the toxicological profiling of NPs (Kahru, Dubourguier 2010; Heinlaan et al. 2008, 2011; Kahru et al. 2008; Kasemets et al. 2009; Aruoja et al. 2009; Blinova et al. 2010; Mortimer et al, 2010, 2011; Ivask et al. 2010).
- test format for kinetic bacterial luminescence inhibition test (Flash Assay) that allows to study turbid and coloured suspensions was first used for NPs and adapted for 96-well microplates (Mortimer et al. 2008).
- test format of algae growth inhibition assay was modified, allowing to discriminate the potential effect of shading of light by NPs as well as increase the throughput of samples (Aruoja et al. 2009).
- for protozoan acute toxicity assay endpoints not interfering with turbidity (such as ATP content) were introduced and evaluated (Mortimer et al. 2010).

- the modulatory effect of environmental matrix on bioavailability and toxicity of metal oxide NPs was studied using natural surface waters from Estonian rivers that are rich in natural organic matter (Blinova et al. 2010).

> Our investigations have produced new scientific knowledge on potential hazard of NPs to living organisms and mechanisms of toxicity of NPs. The multidisciplinary infrastructure of NICPB and availability of more specific equipment by research partners in Estonia and abroad has proven very favourable for the biology-orientated nanotechnology research.

In a recent (2010) Review paper Kahru and Dubourguier "From ecotoxicology to nanoecotoxicology" (Kahru, Dubourguier 2010), part of the Special Issue of Elsevier journal "Toxicology" entitled "Potential Hazard of Nanoparticles: From Properties to Biological & Environmental Effects" (guest editors A.Kahru and K.Savolainen) (2010) existing data on toxicity of eNPs for environmentally relevant species were analysed, summarising also the work of the NICPB researchers. The focus was set on selected NPs (TiO_2, ZnO, CuO, silver, SWCNTs, MWCNTs and C60-fullerenes) and organism groups representing main food-chain levels. The analysis showed that the most harmful were nAg and nZnO that were classified 'extremely toxic', followed by C60 fullerenes and nCuO that were classified 'very toxic'. SWCNTs and MWCNTs were classified 'toxic'. TiO_2 NPs were classified as 'harmful'. These results should deserve thorough attention of environmental risk assessors for evaluation of the potential adverse effects of synthetic NPs on ecosystems. Of course, for a risk assessment, it is also necessary to keep in mind what dosage would be considered realistic and that not every observed biological effect automatically equates to a health risk. But first the dose-effect data are needed, to 'feed' the models. As nanoecotoxicological data are still rare, further studies are needed (Kahru, Dubourguier 2010). It is significant that this Review has been the most downloaded paper of "Toxicology" for the academic year Oct 2009-Sept 2010 and also currently (Jan-March 2011) (http://top25.sciencedirect.com/).

In 2010, the first nanoecotoxicological PhD thesis in Estonia was defended by Margit Heinlaan. Since March 2010 Angela Ivask from NICPB has been in PostDoc in the Center for Environmental Impacts of Nanotechnology at University of California (UC CEIN; http://www.cein.ucla.edu/). The NICPB's *In vitro* toxicology and ecotoxicology Group is also one of 35 European research partners in EU FP7 Project "Development of reference methods for hazard identification, risk assessment and LCA (Life Cycle Assessment) of engineered nanomaterials" (NANOVALID; 2011-2015).

Nano-industry in Estonia

As indicated above, Estonia despite of its small size has undergone considerable development in nanoscience and technology, both from the materials science as well as from the EHS perspective. In addition to research institutions, there are also two nanotechnological industries. One of them, 'NanoFormula' in Voka, Virumaa, is preparing nanotechnological products for household, car care and professional use (www.nanoformula.eu). Currently, NanoFormula exports its eNP-containing products to Europe and China and handles 12 metallic eNPs, including nano-Ag. Another Estonian nanotechnological company, Medicate Est OÜ, is offering a preparation NANO-HAP (nanohydroxyapatite; www.nanohap.eu).

Nano-outlook for Estonia focusing on environmental safety aspects

Estonia is a country of 1.34 million people. The small size of Estonia dictates limits to its R&D possibilities and forces to set priorities and the investment into R&D from the national budget has to be carefully weighed. Estonian Research and Development and Innovation Strategy for 2007–2013 "Knowledge-based Estonia" (www.etis.ee/Portaal/etap.aspx) sets the goal of Estonia to become an innovative top research and high-technology country with a competitive economy. The Strategy also determines priority fields which need special attention from the state: biotechnology, information & communication technologies, materials technology, energy technology, environmental protection and welfare services. However, the Strategy document (49 pages altogether), does not contain the word 'nano'. Although Estonian R&D strategy does not place a special emphasis on 'nano', materials sciences/technology is one of its research priorities.

The Estonian Science Foundation publication "Mobilitas Programme – Express yourself through mobility" (http://www.etf.ee/public/files/Mob%20booklet.pdf) states: "the fields in which Estonia is already globally competitive today are materials science, environmental sciences and ecology, pharmacology and toxicology, as well as botany, zoology and chemistry".

Most of all, the study of environmental safety aspects of nanomaterials is an interdisciplinary one (Ke, Lamm 2011) and needs cooperation of scientists working in biological, environmental and materials sciences, especially when 'safety-by-design' approach is applied (see below). The nano(eco) toxicology and *in vitro* nanotoxicology research incorporating aquatic-food-chain effects related basic research and mechanistic *in vitro* nano(eco)toxicological research needs also knowledge and laboratory equipment to study biomolecular and chemical interactions e.g., on subcellular level. Moreover, mastering of molecular biology, protein analysis and biophysical techniques is also necessary in that interdisciplinary field. Estonian researchers have remarkable competence in all these mentioned research areas.

Bibliometric analysis (Fig. 6) shows that in Estonia analogously to the US, UK, Canada, Australia and Finland the current research into the 'materials sciences' and 'environmental sciences' is well balanced. Differently, in Japan, China, Russia and South Korea there is a strong tendency towards development of 'materials sciences' compared with 'environmental sciences'. The latter is a warning sign as the big countries such as China are rapidly developing nanotechnologies but the research into environmental issues is not probably balancing that.

From the point of view of the research strategy into biological effects of synthetic nanoparticles worldwide, including hazardous effects on humans and environment, following moments could be considered important for Estonia:

- The nano-science (including nanotoxicology and risk assessment) as a perspective direction should be included in national R&D strategy.
- The current and future nanotechnological projects (incl. at state science roadmap level) should be conducted so that the development of new nanomaterials/composites are integrated with projects dealing with respective safety assessment. This allows already in the phase of the early development to 'discard' these nanomaterials/composites that show undesirable properties regarding health or environmental effects. The toxicological screening provides also feed-back and allows tailored modifications (coating etc) of NPs with desirable properties ('safety-by-design' approach).
- The national financial instruments (programs) should be created to fund joint projects involving different essential aspects of nanotechnological research (nanomaterials design and structural studies in parallel with safety-related studies), relying on existing competence and experience in interdisciplinary research. The evaluation of these proposals/projects should be done by 'nano' and 'bio' referees in parallel.
- Funding of larger targeted 'nano-bio' projects should be prioritised joining scientists of different Universities and Research Institutions (creating synergy from different infrastructure and research profiles,

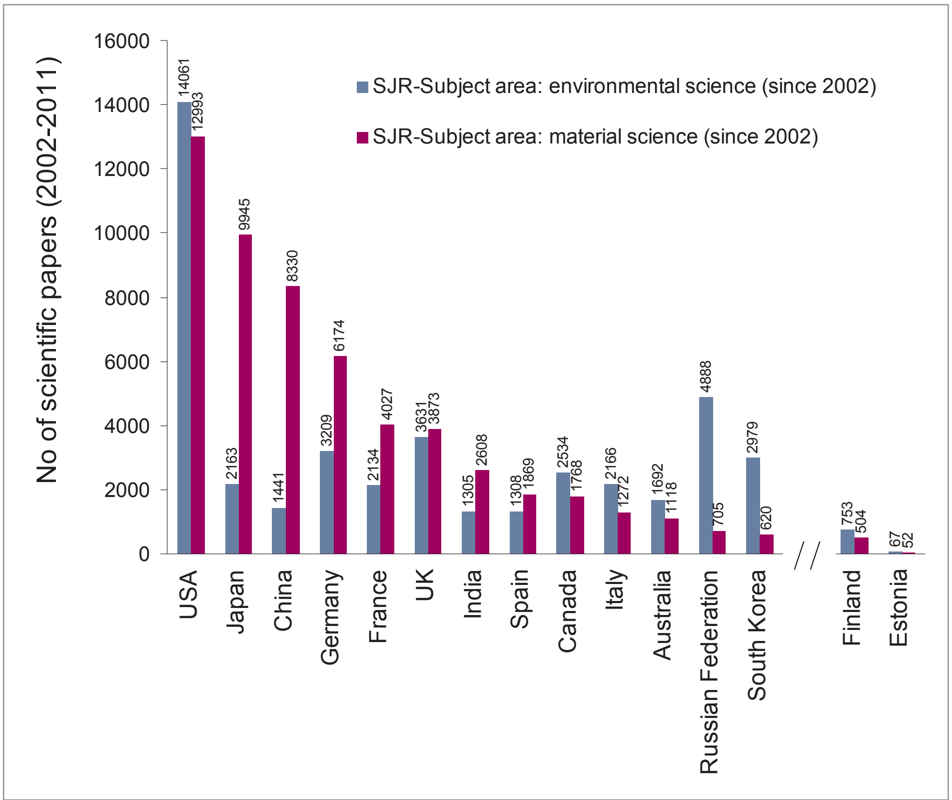


Figure 6.
Number of scientific papers (2002–2011) in two subject areas: 'environmental science' and 'materials science'.
Source: SCImago. (2007). SJR - SCImago Journal & Country Rank.
http://www.scimagojr.com. Retrieved: May 03, 2011.

different existing cooperation partners and contacts). This multi-institutional approach could be one of the prerequisites for funding of big strategic projects, Centres of Excellence and Scientific Roadmap objects. Multidisciplinary approach is also inevitable for 'safety-by-design' nanotechnology development where the competence on nano(eco)toxicology and nanotechnology should be combined. Multidisciplinarity will lead also to co-supervision of PhD students by supervisors from different Institutions and supports cooperation at the national level.

- Education of young people in all major disciplines that are essential for the competent understanding of main research *foci* concerning nanotechnologies, i.e. multidisciplinary training and integrated teaching should be prioritised and supported. That also involves (i) funding of exchange of the PhD students and PostDocs with high-level laboratories and (ii) funding of education of students at MSc level and PhD level abroad in leading Universities on (nano)toxicological disciplines

- As all scientific fields connected with ‘nano’ develop in high speed, due to the special priority national R&D programmes and funding but also due to the scientific curiosity and competition, it is of vital importance to generate and/or improve wide access for the Estonian scientists and students to the main bibliographic databases (Science Direct, Scopus, ISI Web of Science), incl. full access to key-journals. Also, it is important to teach students and researchers how to search the databases and encourage to read also older literature (digitalised former issues of Journals in Science Direct, for example, that contain very useful information for nanosafety research).

Last but not least: the general opinions emphasised in the current paper on the perspectives and feasible priorities of the Estonian nano-research are coherent with the general opinions on the current European priorities concerning environmental safety aspects of nanotechnologies presented in the final draft of the European Academies Science Advisory Council, EC Joint Research Centre (EASAC-JRC) report “Impact of engineered nanomaterials on health: considerations for benefit-risk assessment” (2011).

Acknowledgements

Authors thank their former colleague and co-author, late Prof. Henri-Charles Dubourguier for his invaluable contribution. AK wishes to thank Prof. Endel Lippmaa (NICPB, Estonia) for the discussions that stimulated the visionary and interdisciplinary thinking so precious for this area of research.

REFERENCES

Aruoja, V., et al. 2009. Toxicity of nanoparticles of CuO, ZnO and TiO_2 to microalgae *Pseudokirchneriella subcapitata*. Science of the Total Environment, 407, 4, 1461-1468.

Barnard, A. S. 2006. Nanohazards: knowledge is our first defence. Nature Materials, 5, 245 –248.

Blinova, I., et al. 2010. Ecotoxicity of nanoparticles of CuO and ZnO in natural water. Environmental Pollution, 158, 1, 41-47.

EASAC-JRC Report 2011. Impact of engineered nanomaterials on health: considerations for benefit-risk assessment. Final Draft EASAC-JRC Report. www.easac.eu; http://jrc.ec.europa.eu

Heinlaan, M., et al. 2008. Toxicity of nanosized and bulk ZnO, CuO and TiO_2 to bacteria *Vibrio fischeri* and crustaceans *Daphnia magna* and *Thamnocephalus platyurus*. Chemosphere, 71, 7, 1308-1316.

Heinlaan, M., et al. 2011. Changes in the *Daphnia magna* midgut upon ingestion of copper oxide nanoparticles: A transmission electron microscopy study. Water Research, 45, 1, 179-190

Ivask, A., et al. 2010. Profiling of the reactive oxygen species-related ecotoxicity of CuO, ZnO, TiO_2, silver and fullerene nanoparticles using a set of recombinant luminescent *Escherichia coli* strains: Differentiating the impact of particles and solubilised metals. Analytical and Bioanalytical Chemistry, 398, 2, 701-716.

Ivask, A., Rõlova, T., Kahru, A. 2009. A suite of recombinant luminescent bacterial strains for the quantification of bioavailable heavy metals and toxicity testing. BMC Biotechnology, 9, 41.

Kahru, A., et al. 2008. Biotests and biosensors for ecotoxicology of metal oxide nanoparticles: A minireview. Sensors, 8, 8, 5153-5170.

Kahru, A., Dubourguier, H. C. 2010. From ecotoxicology to nanoecotoxicology. Toxicology, 269, 105-119.

Kahru, A., Savolainen, K. 2010. Potential hazard of nanoparticles: From properties to biological and environmental effects. Toxicology, 269, 2-3, 89-91.

Kasemets, K., et al. 2009. Toxicity of nanoparticles of ZnO, CuO and TiO_2 to yeast *Saccharomyces cerevisiae*. Toxicology in Vitro, 23, 6, 1116-1122.

Ke, P. C., Lamm, M. H. 2011. A biophysical perspective of understanding nanoparticles at large. Physical Chemistry Chemical Physics, 13, 7273-7283.

Kennell, D. 2009. Nanotechnology: An industrial revolution? Monthly Review. http://mrzine.monthlyreview.org/2009/kennell111209.html

Lux Research 2008. Nanomaterials State of the Market Q3 2008: Stealth Success, Broad Impact. Report.
https://portal.luxresearchinc.com/research/document_excerpt/3735

Maynard, A. D., Warheit, D. B., Philbert, M. A. 2011. The new toxicology of sophisticated materials: nanotoxicology and beyond. Toxicological Sciences, 120, Suppl. 1, S109-S129.

Mortimer, M., et al. 2008. High throughput kinetic *Vibrio fischeri* bioluminescence inhibition assay for study of toxic effects of nanoparticles. Toxicology in Vitro, 22, 5, 1412-1417.

Mortimer, M., et al. 2011. Exposure to CuO nanoparticles changes the fatty acid composition of protozoa *Tetrahymena thermophila.* Environmental Science and Technology, 45, 6617-6624.

Mortimer, M., Kasemets, K., Kahru, A. 2010. Toxicity of ZnO and CuO nanoparticles to ciliated protozoa *Tetrahymena thermophila*. Toxicology, 269, 2-3, 182-189.

Nel, A. et al. 2006. Toxic potential of materials at the nanoscale level. Science, 311, 622-627.

Oberdörster, G., Oberdörster, E., Oberdörster, J. 2005. Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. Environmental Health Perspectives,113, 823-839.

Poland, C. A., et al. 2008 Carbon nanotubes introduced into the abdominal cavity of mice show asbestos-like pathogenicity in a pilot study. Nature Nanotechnology, 3, 423-428.

Sihtmäe, M., Dubourguier, H. C., Kahru, A. 2009. Toxicological information on chemicals published in the Russian language: Contribution to REACH and 3Rs. Toxicology, 262, 27-37.

Zhao, Y., Nalwa, H. S. (eds.) 2006. Nanotoxicology – Interactions of Nanomaterials with Biological Systems. American Scientific Publishers.

CHALLENGES IN GEOLOGICAL RESEARCH IN ESTONIA

T. Meidla, V. Kalm, K. Kirsimäe, R. Vaikmäe

INTRODUCTION

Geology is a remarkably wide and diverse area of natural sciences. A number of institutions and a large number of researchers have contributed to study of composition and structures of the sedimentary cover in Estonia and adjacent areas, to recording of related palaeoenvironmental changes and biota, to interpretation of early developments of the region based on the composition and structure of the crystalline bedrock, to study and step-by-step exploration of the mineral and groundwater resources. Results of these research activities are documented in the geological literature and unpublished archives, as well as in the geological maps which have been in continuous development since the early 19th century.

Visibility of fundamental research is often indicated by citation indexes. ISI Web of Science demonstrates a clear increase in publishing geosciences research results by authors working in Estonian R&D institutions. Over the last ten years, from year 2001 to 2010, the number of publications referred by ISI Thomson Reuters increased from 32 to 101 respectively. That makes 4.1% and 6.0% of all publications of Estonian researchers in respective years. From the total sum of citations to articles of Estonian researchers, the citations to geosciences publications amounted to 3.7% in 2001 and 6.5% in 2010. All this happened, against the background of a decrease in numbers of academic positions, from 70 in 2001 to 62 (full time equivalents) in 2010, in the fields of Earth sciences in Estonia. In 2010, according to Essential Science Indicators (ISI Web of Science, ESI), University of Tartu reached a position within 1% of most influential research institutions in Earth Sciences. Against the background of increasing funding of basic scientific research in Estonia (Schiermeier 2009), the proportion of funds from state budget for basic geoscientific research decreased from 83.2% in 2001 to 79.3% in 2010. Instead, proportion of foreign funds (EU FP, contracts, and grants) increased from 3.8% in 2001 to 7.5% in 2010, while income from domestic R&D contracts was stable around 13%. Along with foreign funding, international collaboration was remarkably extending since the 1990s when the science reforms started in Estonia. A trend of globalisation of research contacts has been characteristic to the first decade of the 21st century. In 2001 the list of countries that were most active cooperation partners through joint publications named Sweden, Finland, Canada, France and Germany. In 2010 active cooperation with Nordic countries was continuing but in top of the partners list are now USA, Canada, France, Germany and England.

Several classical subject categories of geology, like multidisciplinary geosciences, palaeontology, physical geography, geology, geochemistry and geophysics, have remained permanently important in the research publications of the last 10 years. However, in 2010 already we had publications referred to in ISI Web of Science in meteorology & atmospheric sciences, remote sensing, limnology, archaeology and geological engineering. In these areas the research results were much less internationally visible earlier, although not totally missing in Estonian science landscape.

Considering the aforesaid and also the relatively simple geological structure of the area, it is often suggested that our knowledge on geology of Estonian and adjacent areas is fairly complete. The generally accepted view on geological development of the area comprises very dynamic developments in the Proterozoic, early stabilisations of the area, very slow latest Proterozoic to mid-Palaeozoic marine deposition being mainly controlled by long-term climatic trends, a multi-stage erosional process from Late Palaeozoic up to Cenozoic, and mainly glacially-influenced sedimentation of the recent past.

As elsewhere (ICSU 2010), the Earth science in Estonia has reached a point of transition where the current priority of understanding of the functioning of Earth system is giving priority to research focused on detailed lab-based investigation of various geological materials, observation-based forecasting and global environmental sustainability. Definition of new challenges in such a dynamic situation could still be possible on the firm basement of previous investigations, and even a well studied area may be capable of surprising us with novel aspects, new possible directions and a variety of possible applications.

GEOFLUIDS IN THE GEOLOGICAL HISTORY OF BALTIC BASIN

The Baltoscandia, encompassing Fennoscandian Shield and the Baltic Basin, represents one of the most stable old cratonic areas of the world. Apatite fission track (AFT) ages of 500-800 Ma in the shielded areas of East European Craton in Finland are the oldest documented anywhere on Earth (Hendriks et al. 2007) recording, thus, an exceptionally stable tectonic position of the Basin. As the result of this stability, the Neoproterozoic to Late Palaeozoic sedimentary deposits within the Baltic Basin, representing a fill-up of a slowly subsiding epicontinental sea and subsequent infill of the developing Caledonian foreland basin, has preserved untectonised and unmetamorphosed. Organic material maturation indicators (reflectivity of vitrinite like material, chitinozoans, acritarchs and biomarker ratios) suggest maximum palaeotemperatures up to ~125 °C in south-western, 80–90 °C in central and 50–80 °C in northern part of the Basin (Lazauskiene, Marshal 2002), which is in agreement with the traditional interpretation of the geological/diagenetic evolution of the basin. That neglects any or at least any important, effect of (hydrothermal) fluid activity in the geological history of the basin.

However, immature organic material indicators in the northern part of the basin are in conflict with this understanding. Recent investigations on diagenetic grade and the palaeomagnetic remagnetisation of sediments (Somelar et al. 2010 and references therein) suggest either much deeper burial (>2 km) or a series of basin wide (hydrothermal) fluid intrusions/thermal perturbations during the basin development.

Based on a flexural subsidence and thermal modelling Middleton et al. (1996) suggested a 4-6 km deep foreland basin in front of a Himalayan type Caledonian orogeny, while several AFT models suggested the sub-Cambrian peneplane being buried to 4-6 kilometres depth in Caledonian foreland basin, causing a significant reheating event.

Erosion of the basin and its marginal areas took place only during the Carboniferous and earliest Permian with the magnitude of the erosion in southern part of the basin in order of 0.5-2 km (Molenaar et al. 2008). Nevertheless, Hendriks and Redfield (2006) have challenged the earlier AFT interpretations by re-interpreting the data in terms of non-thermal Radiation Enhanced Annealing. Consequently they argue the existence of a several kilometres deep foreland basin. Although their conclusion has been lively and hotly debated, it is quite clear that throughout the Fennoscandia there is no direct evidence that a deep, extensive foreland basin ever existed (at least at its eastern margins) and indeed some geological evidence points to the opposite (Hendriks et al. 2007).

Instead, there is a growing evidence of (diagenetic-hydrothermal) fluid activity in relation to Caledonian orogeny in the area. Högdahl et al. (2001) explained low P-T resetting of U-rich zircons to the east of Caledonian front with ~150 °C saline fluids. Kendrick et al. (2005) suggested a Caledonian mineralisation event for Cambrian sandstone hosted Pb-Zn ores, which was caused by mixing of hydrothermal basinal and pore fluids. Microprobe Ar/Ar dating of the zoned K-feldspar overgrowths from the same sandstone suggest two discrete events – early burial diagenesis (528-567 Ma) and a later tectonically induced fluid flow event related to the collapse of the Caledonian orogen (425–400 Ma) (Sherlock et al. 2005). The Pb/Pb and U/Pb data on calcite, fluorite and galena veins as well as Nd model ages of the fluorite bearing veins in Sweden and southern Finland suggest, though with large error, the Caledonian age (~400 Ma) of mineralisation (Alm et al. 2005).

Moreover, recent extensive integrated mineral, fluid and isotope studies for fracture controlled mineralisation in Palaeoproterozoic basement of Central Sweden (Sandström, Tullborg 2009) show four major events, of which two earliest are tied to Proterozoic development of the craton, sometime between 1.8 and 1.1 Ga and mostly during a tectonothermal event between 1.1 and 1.0 Ga, possibly in response to far-field effects of the Sveconorwegian orogeny. The third mineralisation event, characterised by quartz, calcite, pyrite and asphaltite precipitation temperatures between 60 and 190 °C is correlated to a mixing of a fluid emanating from an organic rich overlying sedimentary cover and a deep basinal fluid from the crystalline bedrock, possibly as a far-field response to the Caledonian orogeny and/or the development of the Caledonian foreland basin (Sandström, Tullborg 2009). The same event is possibly recorded in fracture mineralisation in southwestern Finland. The fourth mineralisation event described in Sweden is possibly related to episodic precipitation during a long time period from the Late Palaeozoic to the present at low temperature conditions (<50 °C).

Also in the Baltic Basin, the Rb-Sr ages of Cambrian claystones and K-Ar ages of Ordovician K-bentonites at northern margin of the BB suggest a major Late Silurian – Devonian (425-350 Ma) event of recrystallisation (Somelar et al. 2010). The Early Devonian episode of illitisation is suggested in southern Baltic Paleobasin and Pomerania (Srodon et al. 2009). Additionally, the Late Silurian-Devonian epigenetic regional and fracture controlled metasomatic dolomitisation of carbonate succession within the Baltic Basin was suggested by Pichugin et al. (1977). Hydrothermal sepiolite has been reported in crystalline basement fracture/fault zones (Kuuspalu et al. 1973) in northern part of Basin. Recently Somelar et al. (2010) suggested flushing of K-rich fluids to explain illitisation of bentonites in N and NW of the Basin.

Results of our pilot research support migration of the K-rich fluid by common findings of authigenic K-feldspar within the sandstone rocks of Ediacaran, Cambrian and Ordovician age suggesting that fluids were saturated enough with respect to K, Al and Si to induce the nucleation of authigenic K-feldspar and probably authigenic illite. Another evidence line is the multicomponent remanent magnetisation in southern Finland and in Baltic Basin. The secondary magnetisations pinpoint towards the (1) Caledonian orogenic and (2) Permo-Triassic remagnetisation events (Plado et al. 2008; Preeden et al. 2009). The same fluids resulting from the Caledonian tectonic reactivation of the Svecofennian basement were probably responsible for the Silurian-Devonian remagnetisation recorded in the shear zones in southern Finland (Preeden et al. 2009) and Late Devonian-Mississippian remagnetisation of Silurian dolomites in central Estonia (Plado et al. 2008). The remagnetisation behaviour of the crystalline rocks in southern Finland (Preeden et al. 2009) and Ordovician and Silurian carbonate successions in the northern part of the Baltic Basin (Plado et al. 2008; Preeden et al. 2009) points, however, toward a second remagnetisation event in Permian to mid-Triassic, and probably in the Cretaceous, that might be related to the Permian-Triassic uplift and extensive erosion all over the Svecofennian domain and BB. There is also an evidence for a late Variscian thermal event associated with numerous diabase intrusions identified in the Baltic Sea. Ar/Ar dating of two igneous bodies hosted by Silurian shales revealed a late Devonian-Carboniferous age of the diabases (Šliaupa et al. 2001).

This geological evidence suggests that the overall stable development of the Baltic Basin since its formation was interrupted by series of fluid/thermal events and possibly tectonic in scales that have not been earlier considered in Baltic Basin. The physical and chemical properties and the age of these hydrothermal fluids in the Proterozoic crystalline basement and Neo-Proterozoic (Ediacaran) – Phanerozoic sedimentary cover within Baltic Basin are virtually unknown, but complex vertical and lateral diagenetic gradients in basin rocks suggest that the fluid migration was fracture controlled and the most intense chemical reactions are tightly coupled to crystalline basement and bedrock fracture zones with enhanced permeability. The old fault structures of the basement may have been recurrently reactivated during the Phanerozoic tectonic stresses and they gave a rise to the fluid circulation systems.

Further studies would test the nature of fracture controlled alteration mineral assemblages, paragenesis and spatial structure of alteration, which would constrain the properties and origin of (hydrothermal) fluids of recorded events using mineralogical, geochemical and fluid inclusion observations coupled with geochemical modelling. Evaluation of fluid driven activity in geological section in Baltic Basin forwards also important analytical challenges that require subgrain level and *in situ* mineralogical, fluid-inclusion and geochemical-isotopic evidences for intrusion of the diagenetic or hydrothermal fluid/brines. Important aspect of these studies would include precise dating and correlation of fracture controlled thermal fluid activity with tectono-magmatic/geodynamic events within Baltic Basin and on the Craton as a whole.

It is quite evident that these processes are linked to large scale Palaeozoic geodynamic processes in Baltica micro-continent and its margins, though we cannot exclude local scale reheating events, originating either from sedimentary, magmatic, and tectonic/impact processes. We would forward a hypothesis that the main fluid-event responsible for fracture controlled mineralisation (incl. epigenetic dolomitisation and clay mineral transformation) is possibly related to Scandinavian Caledonian orogenic development in Late Silurian-Devonian, whereas the significant remagnetisation of rocks points towards second, low temperature and oxygenated fluid event resulted in Late-Palaeozoic – Mesozoic reactivation.

Last, but not least – these studies are not of only scientific importance. Metallogenic potential of the Baltic Basin has been considered relatively low so far, though small scale evidences of Pb-Zn mineralisation have been reported, mainly in relation to regional dolomitisation. Geological setting of Baltic Basin would allow existence of either meteoric or basinal hydrothermal systems. First type is the best known by oxide Cu and Zn deposits in combination with roll-front uranium ores, while basinal hydrothermal systems are responsible for Mississippi Valley-type sulfide ore deposits containing significant Zn and Pb concentrations. Character of sulphide mineralisation as well as preliminary fluid inclusion results in central Estonian strongly dolomitised sections would hint specifically to Mississippi Valley-type mineralisation in localised fracture zones where bacterially mediated sulphate has created conditions for sulphide ore deposition. The recognition of significant hydrothermal activity in the rocks could lead us to a better understanding of deep processes and to finding of (new) hidden resources.

CLIMATIC RECORD FROM GLACIERS AND DEEP TIME

Climate and environmental change is arguably the biggest challenge mankind is going to face in this century. Natural cold-warm cycles of glaciations exemplify full spectrum of processes, shaping currently existing although changing natural conditions, and the study of that provides an understanding of environmental variation on much larger scale than reflected in historic records of the last centuries. One of our tasks is to reconstruct natural variations and cyclicity in past temperatures and rates of changes reflected in variety of geological, isotopic and paleontological evidence.

It is now fully recognised that a thorough knowledge of past environmental and climatic conditions is one of the keys to better understanding of the changes currently affecting our climate and our environment, as well as for anticipating their evolution in the coming century and beyond (Intergovernmental Panel on Climate Change IPCC). This information is built on a large variety of palaeodata inferred from studies based on terrestrial, oceanic, ice and sediment records.

Based on those various climate 'archives', i.e. on different types of material which allow recording and interpreting the palaeoenvironmental signals, geology-related studies in the field of palaeoenvironments and palaeoclimates in Estonia are carried out at different institutions. In numerous studies (Saarse et al. 2009 and references therein), various proxy data (subfossils, pollen spectra and tephra) are used to evaluate early human impact to the development of vegetation. In combination with the data from sedimentary record, climatic and palaeogeographic record throughout the Late Pleistocene and Holocene is integrated in the present understanding on the geological history of Estonian and adjacent areas during Pleistocene and Holocene (Kalm et al. 2011 and references therein). This relatively traditional approach has gained further support from an important lineage of isotope-geochemical investigations which are concentrated on ice cores (Vaikmäe 1990). The Palaeozoic workers have been increasingly successful in deciphering the palaeoenvironmental signal from the Ordovician and Silurian strata (Brenchley et al. 1994; Ainsaar et al. 2010; Kaljo, Martma 2006 and references therein).

Most of the conclusions and also modelling attempts in the contemporary climate research are based on relatively short historical time series. This is one of the reasons why it has been (and likely will be) difficult to reach a high-level international consensus on protective measures. Climate studies are a field where experiments are and will be hardly possible, or hardly safe if planned. At the same time, long-time forecasts are not yet trustworthy, processes and features being hardly predictable. The only possibility of replacing experiments is extending our knowledge on climatic processes in more distant past, including the record of deep time, the vast geologic time scale. Working with deep time data has already delivered substantial success to various working groups in Estonia.

The history of international ice core research started in the 1960s largely thanks to the pioneering work of Willy Dansgaard, who documented the worldwide pattern of the $^{18}O/^{16}O$ ratio of precipitation. He showed and modelled the parallel decrease of this ratio (hereafter $\delta^{18}O$) and of the temperature of the site for mid- and high latitudes and proposed that this relationship could be used to reconstruct past climate changes from ancient precipitation accumulated in ice sheets (Dansgaard et al. 1969).

Since the early 1960s, the global ice core community has produced a wealth of scientific results from a still relatively limited number of deep drilling sites in Greenland and Antarctica with the longest record extending back to the last interglacial in Greenland and covering eight glacial-interglacial cycles in Antarctica (Jouzel et al. 2007). The Vostok, Dome Fuji, and Dome C long deep Antarctic ice cores offer a coherent picture of glacial-interglacial variations in Antarctic temperature and in global atmospheric composition (Jouzel, Masson-Delmotte 2010).

By now the longest record available from Dome C, covers eight climatic cycles and shows that the Antarctic climate is marked by glacial-interglacial variations reaching 8–10 °C. Glacial temperatures are relatively stable from one glacial period to the next. By contrast, Antarctic temperatures exhibit large variations from one interglacial period to the next. Prior to 400,000 years ago, lukewarm interglacials are systematically 1–3 °C cooler than the Holocene (EPICA 2004). No Greenland ice core record is yet available spanning more than a climatic cycle. The 123,000 years of climate history archives in the NorthGRIP ice core suggest that the end of the last interglacial period was 3–5 °C warmer than today, which is coherent with the large Eemian northern hemisphere summer insolation change (North GRIP, 2004). At the same time, the Greenland ice core records are marked by a succession of 25 abrupt events punctuating the last glacial period and the last deglaciation. The Dansgaard-Oeschger (DO) events are formed by a cold and rather stable phase, which can last for several thousands of years ('stadial'), followed by an abrupt warming, and a progressive cooling from this warm transient interstadial back to glacial conditions. The latest of those abrupt events, the cold Younger Dryas (YD) event, a climate stage which took place during the second half of the last deglaciation and gave the deglaciation two-step character, is well recognised in Western Europe, in Greenland and in North Atlantic (Rind et al. 1986). For example in Greenland, a gradual temperature decrease started at the Bölling (B) around 14.5 ky BP, spanned through the Alleröd (A) and was followed by the cold YD event, which terminated abruptly around 1.5 ky BP. Recent results suggest that this BA/YD sequence may have extended throughout all of the Northern hemisphere, but the evidence of a late transition cooling was for long poor for the southern Hemisphere.

Knowledge of links between Southern and Northern hemisphere climates throughout the deglaciation period is crucial for better understanding the global climate change mechanisms. A detailed isotopic record analysed in an ice core drilled at Dome B in east Antarctica, within a joint project between Estonia, France and Russia, fully demonstrated for the first time the existence of an Antarctic cold reversal (ACR) (Jouzel et al. 1995). These results suggest that the two-step shape of the last deglaciation has a worldwide character but they also point at noticeable inter-hemispheric differences. Thus, the coldest part of ACR, which shows a temperature drop about three times weaker than in the YD record in Greenland, may have preceded the YD by about 1 ka and the two warming periods started there before they started in Greenland.

With the completion of major projects in Greenland and Antarctica over the last 15 years, the international ice core community is making plans for the next decades. The costs and scope of future work emphasise the need for coordinated international collaboration. Developing this international collaboration is the charge of international partnership in Ice Core Sciences (IPICS), a planning group currently comprising ice coring scientists, engineers, and drillers of 18 nations, including Estonia (Brook et al. 2006). Through a series of international meetings of the IPICS, the four forthcoming projects were defined as follows:

- A deep ice coring program in Antarctica that extends through the mid-Pleistocene transition, a time period where Earth's climate shifted from 40,000 year into 100,000 cycles;
- A deep ice core in Greenland recovering an intact record of the last interglacial period;
- A bipolar network of ice core records spanning approximately the last 40,000 years;
- A global network of ice core records spanning the last 2,000 years.

The fifth, critical element of IPICS is the development of advanced ice core drilling technology.

Working with the palaeoenvironemental signal from the Palaeozoic is much more complicated. The advances of the last decades are mostly related to the palaeoenvironmental studies in the Ordovician and Silurian which are well developed, sufficiently exposed and of excellent preservation in Estonia. Palaeoenvironmental studies could not be carried out in a narrow area but require proxy data from different environmental and climatic settings to be analysed simultaneously.

The Ordovician and Silurian periods comprise an important part of the warm and long-lasting latest Pre-Cambrian-Early Devonian greenhouse cycle. This epoch was generally characterised by high pCO_2 level, warm climate, widespread tropical shelves, high eustatic sea level and increasing diversity of biota, being in many respects different of the present-day conditions, overall.

The long-lasting greenhouse epoch was interrupted by one or several short episodes of global climatic perturbation, the most important of them being the latest Ordovician Hirnantian glaciation (Brenchley et al. 1994; Brenchley 2004). The Hirnantian glaciation was a prominent event which is marked by tillites or glaciomarine diamictites in high latitudes (*ibid.*, Fig. 9.2).

It is noteworthy that the Hirnantian glaciation is rather similar to the last glaciation in several features. It has nearly equal duration (less than two millions of years), comprises several glacial stages, was similar in extent (the estimates of the Hirnantian glacioeustatic regression vary mostly between 50 and 100 m), similar influence to the palaeoenvironments (resulting in a fall in seawater temperatures by possibly as much as 10 °C – Brenchley et al. 1994). In a way, the Hirnantian glaciation could be taken as a model of the Quaternary glaciation which just ended or suspended.

The Hirnantian glaciation caused major two-stage extinction, one related to the sea level drop with temperature decrease and another to the deglaciation-related sea level rise. Local extinctions of populations or species in a particular area are commonly observed today. Extinctions on a larger scale are commonly studied in the fossil record. The distinction between these two phenomena is not always simple in the fossil record, because the local/regional environmental disturbances are quite often of a similar magnitude with the major extinction events, being expressed as elimination of a number of taxa from the continuous fossil record. At the same time, it is very clear that we are today facing a remarkable extinction event. Number of species who went extinct during the historical period may be counted in hundreds and may reach thousands in near future, but understanding consequences and related new trends is very much dependent on the background data – which are still very difficult to obtain.

Extinctions tend to have a complicated nature because of selectivity and biogeographical variation. Past extinction studies of today include a combination of methods – palaeontological, stratigraphic, sedimentological, geochemical, palaeoclimatological and others, whereas the integrated studies are today widely replacing an elementary biostatistical approach. The Hirnantian glaciation and related environmental changes could effectively be used for calibrating and developing the climate models of today, if certain conditions could be fulfilled:

- time series of the environmental data of the recent past must be extended and reliability increased;
- time resolution of the Palaeozoic environmental data needs to be increased and resolution of the proxy signals improved.

Our geological setting provides a unique opportunity to study the regional climatic changes by integrated sedimentological, palaeontological and geochemical methods in much more detail, and to test the influence of global events to the regional climate and biota. Palaeozoic studies are usually operated on a timescale calibrated in millions of years, but increasing resolution of investigations is possible in case of rapid environmental changes such as climate perturbations (glaciations, e.g. Meidla et al. 2011 and references therein), sedimentological changes (volcanic ash falls, e.g. Hints et al. 2003) and perhaps in case of changes of ocean chemistry and circulation (Ainsaar et al. 2010). Attempts of distinguishing Milankovič-scale cycles in the latest Ordovician sedimentary successions (Armstrong 2007), attempts of creating a temporal framework of about 20-100 ky for the early Palaeozoic sedimentary successions (Sadler et al. 2009) are changing limitations of time resolution at deciphering processes in deep time.

It is obvious that improving usefulness, resolution and application of the palaeoenvironmental data of deep time will be vital in future. As an instrument of climate studies, this approach would be a safer and easier (and less costly) way of improving quality of our climate models and reliability of our forecasts, compared to experiments which already are carried out or planned in various fields (e.g. CO_2, capture and storage, manipulation of the oceanic and atmospheric processes, etc.).

Case studies on the past biodiversity, extinctions and recoveries, are equally important Most of the species that have ever lived on Earth are extinct, whereas the extinctions of fossil taxa and subsequent replacement by other species are the processes which can be observed in the life history repeatedly. The understanding on processes of extinction and recovery are an important issue for understanding the evolution of the biosphere. Environmental case studies on the Early Palaeozoic are potentially capable of influencing the basic concepts of palaeontology, palaeoecology and Earth System Science, forming a platform for better understanding the global change of today.

ON A BORDERLINE BETWEEN 'GREEN' AND 'HARD' SCIENCES – THE ACTIVE FILTRATION OF PHOSPHORUS

Phosphorous is a vital element for existence of life as such being a major constituent of DNA, RNA, ADP and ATP and phosphorus in the form of Ca-phosphate mineral makes us stand upright. Apart of that, the phosphorous is an important nutrient of the biochemical cycle.

However, already slightly elevated concentrations of phosphorous can trigger the eutrophication. For more than 40 years, excess phosphorus has been recognised as one of the main nutrients responsible for eutrophication. Phosphorous, mainly in the form of soluble phosphate, is intensively used for agricultural and domestic purposes (fertilisers, detergents, etc.) and the run-off from agricultural landscapes and especially the domestic waste waters are enriched in respect to phosphates. Consequently, the effluents from human activities must be treated to remove the surplus P, in order to maintain the ecological balance of natural systems. Phosphorous removal and recycling from waste water is not a traditional research problem in geological sciences. However, as limits for discarded water quality are becoming more and more stringent, demand for efficient but cheap filter materials is increasing.

In recent years worldwide studies have shown effectiveness of active filtration through alkaline media for the removal of phosphorous (e.g. Kõiv et al. 2010 and references therein). Active filtration technique uses direct precipitation/immobilisation of the phosphates into low soluble forms and it is considered as a promising technique for small-scale waste water treatment, whereas the nutrient-loaded filter material can be used in agriculture as phosphorous fertiliser and soil conditioner. Potential materials for active filtration of phosphorus include the technological waste products as metallurgical slags (Pratt et al. 2007) and high Ca content industrial products, coal and oil shale ash, etc., where Ca occurs in CaO form (lime) and/or natural Ca-silicate minerals/rocks (e.g. Adam et al. 2007).

Active filtration is a complex hydrogeochemical-mineralogical process, the effectiveness of which depends on the ionic strength of the solution, the activity of Ca and other competing ions, the form and activity of P, and chemical-mineral composition and microstructure of the filter material. Recent studies (Adam et al. 2007) have shown that P removal is inhibited in the presence of competitive ions and/or inhibitors such as Mg^{2+}, Cl^{-}, SO_4^{2-}, CO_3^{2-} and also organic compounds (humic acids), whereas extensive supersaturation of pore-water with respect to Ca^{2+} and phosphate is required for the precipitation of stable Ca-phosphate phases (Liira et al. 2009). This means that effective use of these low cost materials would require optimisation of the binding reactions. At this point the geological know-how of meta-stable mineralogical and low-temperature geochemical systems meets the environmental technology.

Though phosphorus retention in Ca-rich filter materials, as hydrated oil shale ash, occurs through irreversible reactions of adsorption and precipitation on a solid phase it is still influenced by the complex physical-chemical properties of the heterogeneous material, especially its high content of different Ca- and Al-compounds, porosity and high pH. A typical problem is a chemical clogging of the filter material by calcium carbonate precipitates (e.g., Liira et al. 2009) that results possibly from some limitations in sorption kinetics or the chemical stability of the effluent, whereas the retention efficiency on calcite is not sufficient, because of its sensitivity to carbonate equilibrium.

Moreover, Kõiv et al. (2010) show that the removal rate of phosphorus in hydrated oil shale ash filters depends significantly on the concentration of P in solution and possibly all Ca-rich alkaline filter materials such as hydrated oil shale ash, C-type coal ash, crushed gas-concrete and Filtralite®-P type lightweight aggregates, which are cheap filter materials for P removal in different wastewater compositions, may suffer from incomplete removal of trace amounts of phosphorus at low inflow concentrations, limiting the usefulness of active filtration technique for wastewater purification to levels above 0.5 mg P L^{-1} while lower concentrations are required in regulation limits for waste water in many countries.

Another emerging question of the active filtration is the binding capacity of Ca-rich waste materials towards chemical analogues of phosphate (oxyanions) as arsenates, vanadates and molybdates, and possible immobilisation of heavy metals. Arsenic contamination, both from geological and anthropogenic sources is a major concern world-wide and stable binding of mobilised oxidised arsenic would be another application for these materials. Moreover, hydrated fly ash containing high Ca and sulphate sulphur are characterised by presence of ettringite, which can account for up to 30% of crystalline phases. The geochemical behaviour of ettringite can be compared to natural Ca-phosphate, apatite. Similarly to apatite, ettringite allows a suitably high and stable pH of the solution and good kinetics for secondary Ca-phosphate crystallisation by providing a high number of nucleation sites on its needle like crystallites. Importantly, ettringite has been shown to control the activity of several contaminants, both cations and oxoanions, and it is concomitant with the reduction of leachate heavy metal trace elements. This would further enhance the environmental potential of the Ca-rich waste materials.

CONCLUDING REMARKS

Balance between 'classical' discipline-based research and interdisciplinary research will change in favour of the latter even further. Landsliding process and probability in varved clays (Kohv et al. 2010), diatom-based water quality estimation (Heinsalu, Alliksaar 2009), hydrogeological impact of underground mining on wetland hydrology, CO2 storage in the bedrock (Shogenova et al. 2011), chemical properties and possible usage of oil-shale ash (Kõiv et al. 2010), medical aspects of groundwater chemistry (Indermitte et al. 2009) are worth mentioning.

All the above aspects and challenges have diversified basic research, and together with growing interest of industry in utilising science outcomes have initiated a number of new directions in the applied research in Estonia since the beginning of this century. Participation in state-financed technology programmes, namely the Energy Technology Programme and the Environmental Technology Programme, both running until 2015, are increasing the proportion of applied research of geoscientists even further. Advanced study of climate-related Earth surface processes should unite the disciplines of climatology, geology, palaeontology, glaciology, and atmospheric sciences, which traditionally have been somewhat disparate in Estonia.

In addition to these increasingly important research focuses, new science-related issues, as management and financing of trans-disciplinary research, facilities for metadata interfaces and joint standards across many disciplines, relevance of research outcomes to decision makers, research-based education of students, are the challenges alongside with fostering basic study in Earth science and related sciences.

REFERENCES

Adam, K., et al. 2007. Phosphorus retention in the filter materials shellsand and Filtralite P® – batch and column experiment with synthetic P solution and secondary wastewater. Ecological Engineering, 29, 200-208.

Ainsaar, L., et al. 2010. Middle and Upper Ordovician carbon isotope chemostratigraphy in Baltoscandia: A correlation standard and clues to environmental history. Palaeogeography, Palaeoclimatology, Palaeoecology, 294, 3-4, 189-201.

Alm, E., Huhma, H., Sundblad, K. 2005. Preliminary Paleozoic Sm–Nd ages of fluorite-calcite-galena veins in the south eastern part of the Fennoscandian Shield. Svensk Kärnbränslehantering AB, R-04-27.

Armstrong, H. A. 2007. On the origins of the Late Ordovician glaciation. Williams, M., et al. (eds.) Deep-time Perspectives on Climate Change; Marrying the Signal from Computer Models and Biological Proxies. Geological Society of London, 101-121. (The Micropalaeontological Society Special Publications).

Brenchley, P. J. 2004. End Ordovician climate. Webby, B. D., et al. (eds.) The Great Ordovician Biodiversification Event. Columbia University Press, New York, 81-83.

Brenchley, P. J., et al. 1994. Bathymetric and isotopic evidence for a short-lived Late Ordovician glaciation in a greenhouse period. Geology, 22, 295-298.

Brook, E., et al. 2006. The future of ice coring: International Partnerships in Ice Core Sciences (IPICS). PAGES News, 14, 1, 6-10.

Dansgaard, W., et al. 1969. One thousand centuries of climatic record from Camp Century on the Greenland ice sheet. Science, 166, 377-380.

EPICA community members, 2004. Eight glacial cycles from an Antarctic ice core. Nature, 429, 623-628.

Heinsalu, A., Alliksaar, T. 2009. Palaeolimnological assessment of the reference conditions and ecological status of lakes in Estonia – implications for the European Union Water Framework Directive. Estonian Journal of Earth Sciences, 58, 4, 334-341.

Hendriks, B., et al. 2007. A fission track data compilation for Fennoscandia. The Norwegian Journal of Geology, 87, 143-55.

Hendriks, B. W. H., Redfield, T. F. 2006. Apatite fission track and (U-Th)/He data from Fennoscandia: An example of underestimation of fission track annealing in apatite. Earth and Planetary Science Letters, 236, 1-2, 443-458.

Hints, O., et al. 2003. Biotic effects of the Ordovician Kinnekulle ash-fall recorded in northern Estonia. Bulletin of the Geological Society of Denmark, 50, 115-123.

Högdahl, K., Gromet, L. P., Broman, C. 2001. Low *P-T* Caledonian resetting of U-rich Paleoproterozoic zircons, central Sweden. American Mineralogist, 86, 534-546.

ICSU 2010 = Earth System Science for Global Sustainability: The Grand Challenges. International Council for Science, Paris. http://www.icsu-visioning.org/wp-content/uploads/Grand_Challenges_Nov2010.pdf

Indermitte, E., Saava, A., Karro, E. 2009. Exposure to high fluoride drinking water and risk of dental fluorosis in Estonia. International Journal of Environmental Research and Public Health, 6, 2, 710-721.

Jouzel, J., Masson-Delmotte, V. 2010. Deep ice cores: the need for going back in time. Quaternary Science Reviews, 29, 3683-3689.

Jouzel, J., et al. 1995. The 2-step shape and timing of the last deglaciation in Antarctica. Climate Dynamics, 11, 3, 151-161.

Jouzel, J., et al. 2007. Orbital and millennial Antarctic climate variability over the past 800,000 years. Science, 317, 793-796.

Kaljo, D., Martma, T. 2006. Application of carbon isotope stratigraphy to dating the Baltic Silurian rocks. GFF, 128, 123-129

Kalm, V., et al. 2011. Chapter 8, Pleistocene Glaciations in Estonia. Ehlers, J., Gibbard, P. L., Hughes, P. D. (eds.) Quaternary Glaciations – Extent and Chronology – A Closer Look. Elsevier, Amsterdam, 95-104.

Kendrick, M. A., et al. 2005. Noble gas and halogen evidence for the origin of Scandinavian sandstone-hosted Pb-Zn deposits. Geochimica et Cosmochimica Acta, 69, 1, 109-129.

Kohv, M., et al. 2010. Analysis of a retrogressive landslide in glaciolacustrine varved clay. Engineering Geology, 116, 1-2, 109-116.

Kõiv, M., et al. 2010. Phosphorus removal using Ca-rich hydrated oil shale ash as filter material – the effect of different phosphorus loadings and wastewater compositions. Water Research, 44, 10, 5232-5239.

Kuuspalu, T., Utsal, K., Vanamb, V. 1973. On the hydrothermal sepiolite from the crust of weathering of the Estonian crystalline basement. Eesti NSV Teaduste Akadeemia Toimetised, Keemia, Geoloogia, 22, 2, 169-173, (in Russian).

Lazauskienė, J., Marshall, J. 2002. Chitinozoan reflectance and the thermal history of the Lower Palaeozoic sediments of the Baltic Basin. Satkūnas, J., Lazauskienė, J. (eds.) 5th Baltic Stratigraphic Conference. Basin Stratigraphy - modern methods and problems. Vilnius, 93-97.

Liira, M., et al. 2009. Active filtration of phosphorus on Ca-rich hydrated oil shale ash: does longer retention time improve the process? Environmental Science and Technology, 43, 10, 3809-3814.

Meidla, T., Ainsaar, L., Truuver, K. 2011. Ostracods in Baltoscandia through the Hirnantian crises. Gutierrez-Marco, J. C, Rabano, I., Garcia-Bellido, D. (eds.) Ordovician of the World. Instituto Geologico y Minero de Espaňa, Madrid, 353-357.

Middleton, M. F., et al. 1996. Modelling of a Caledonian foreland basin in Sweden: Petrophysical constraints. Marine and Petroleum Geology, 13, 407-413.

Molenaar, N., et al. 2008. Lack of inhibiting effect of oil emplacement on quartz cementation: Evidence from Cambrian reservoir sandstones, Paleozoic Baltic Basin. Geological Society of America Bulletin, 120, 9-10, 1280-1295.

Pichugin, M. S., et al. 1977. Regional metasomatic dolomitization associated with tectonic disturbances in Lower Paleozoic deposits of the northern Baltic region. International Geology Review, 19, 903-912.

Plado, J., et al. 2008. Paleomagnetic age of remagnetizations in Silurian dolomites, Rõstla quarry (Central Estonia). Geological Quarterly, 52, 213-224.

Pratt, C., et al. 2007. Phosphorus removal mechanisms in active slag filters treating waste stabilization pond effluent. Environmental Science and Technology, 41, 2007, 3296-3301.

Preeden, U, Elminen T., Plado, J. 2009. Secondary magnetizations in shear and fault zones in southern Finland. Tectonophysics, 479, 3-4, 203-213.

Rind, D., et al. 1986. The impact of cold North Atlantic Sea surface temperatures on climate: Implications for the Younger Dryas. Climate Dynamics, 1, 3-33.

Saarse, L., et al. 2009. Development of the late glacial Baltic basin and the succession of vegetation cover as revealed at Palaeolake Haljala, northern Estonia. Estonian Journal of Earth Sciences, 58, 4, 317-333.

Sadler, P. M., Cooper, R. A., Melchin, M. 2009. High-resolution, early Paleozoic (Ordovician-Silurian) time scales. Bulletin of the Geological Society of America, 121, 5-6, 887-906.

Sandstrom B., Tullborg, E.-L. 2009. Episodic fluid migration in the Fennoscandian Shield recorded by stable isotopes, rare earth elements and fluid inclusions in fracture minerals at Forsmark, Sweden. Chemical Geology, 266, 3-4, 126-142.

Schiermeier, Q. 2009. Beyond the Block. Nature, 461, 590-591.

Sherlock, S. C., et al. 2005. High spatial resolution $^{40}Ar/^{39}Ar$ laserprobe analysis of complex polygenetic K-feldspar overgrowths: a record of sandstone burial and tectonically-driven fluid migration. Earth and Planetary Science Letters, 238, 329-341.

Shogenova, A., et al. 2011. CO_2 geological storage capacity analysis in Estonia and neighbouring regions. Energy Procedia, 4, 2785-2792.

Šliaupa, S., et al. A. 2001. Petrophysic properties of Cambrian deposits of central and western Lithuania. Research in Petroleum Technology, 5, 157-180.

Somelar, P., et al. 2010. Illitization of Early Paleozoic K-bentonites in the Baltic Basin: decoupling of burial- and fluid-driven processes. Clays and Clay Minerals, 58, 3, 388-398.

Środoń, J., et al. 2009. K-Ar dating of the Lower Palaeozoic K-bentonites from the Baltic Basin and the Baltic Shield: implications for the role of temperature and time in the illitization of smectite. Clay Minerals, 44, 3, 361-387.

Vaikmäe, R. 1990. Isotope variations in the temperate glaciers of the Eurasian Arctic. Nuclear Geophysics, 4, 1, 45-55.

THE ESTONIAN LANGUAGE AND LINGUISTICS: STATE-OF-THE-ART AND FUTURE PERSPECTIVES

H. Õim, K. Pajusalu

The paper gives an overview of the research in the Estonian language and linguistics and goals for the future. It does not tackle the issues in the framework of traditional linguistics only, but in close connection with the problems and tasks of the language technology. In the contemporary information society, language as the main means of communication has increasingly become mediated by information technology and its techniques. In the paper we will give an overview of the main results and future plans in this cooperative research in Estonia.

INTRODUCTION

This paper will give a general overview of research in the Estonian language and linguistics, and outline directions and goals for the future. The paper does not only deal with linguistics in the traditional sense. In the contemporary information society language as the main means of communication has increasingly become mediated by information technology in its multiplicity of techniques. This area of research and development is generally known as language technology but in some areas, e.g. text understanding and human-computer interaction it is tightly connected with artificial intelligence. Technical realisations are putting linguists under constant pressure to present their research results in the form which could be used in realisation-oriented projects. Thus, in the present paper the focus will be on language technology, because it is nowadays intertwined with various branches of linguistics. Nevertheless, the bulk of the paper will be devoted to providing an overview of the research in the Estonian language and linguistics: the present state and future directions. Finally, the paper will outline the directions and future research plans, which are linked to the official research program put together by the Estonian Ministry of Education and Research – the Development Plan of the Estonian Language (2011–2017) (see EKA 2011).

AN OVERVIEW OF EARLIER RESEARCH

The scientific study of Estonian is relatively short. The first doctoral dissertation on Estonian was defended by Mihkel Veske in 1872 at the University of Leipzig. During the time of the Russian Empire, several well-known linguists worked at the University of Tartu (e.g. Jan Baudouin de Courtenay (1883–1893) who is considered to be one of the founders of structural linguistics), but their work had very little to do with the Estonian language.

Linguistic research in the Estonian language started in earnest in the first half of the 20th century, when the professorship of the Estonian language was established at the University of Tartu. The main research areas were the description of features of Estonian as a Finno-Ugric language, the status of Estonian in this language family, and the systematic study of Estonian dialects.

When looking at the development of linguistics in Estonia, at least two critical periods should be pointed out which in many ways influenced current research themes and directions. The first one occurred during the second half of the 1960s and the first half of the 1970s, and resulted in the advent of contemporary linguistics. The second one was evidenced in the middle of the 1990s, opening up new contacts with European linguistics more broadly, and with language technology projects more specifically.

In the middle of the 20th century in the Soviet Union, several scientific research areas, which hitherto had been banned as 'idealistic' (e.g. cybernetics), received an official status. Among these was a subfield called structural and mathematical linguistics. At the time, at the University of Tartu, Huno Rätsep (member of the Academy of Sciences) compiled a study program which included lectures on modern trends of linguistics. That program was extended in accordance with the development of linguistics, starting with structural linguistics, to follow with generative grammar and its developments. The most important result of the program was the cropping up of a new generation of Estonian linguists who shared a conception of language as a structured whole, and at the same time had a good knowledge of the approaches used to describe different language phenomena (for instance, in morphology, syntax, semantics) in different and often competing theoretical concepts. Today there is a 'critical mass' of researchers to cover all layers and different branches of language description, from phonetics and phonology, morphology and syntax to semantics and pragmatics, from purely theoretical linguistics to computational linguistics and its applications. A new generation of young linguistics, already at work, will determine the contents and form of linguistic research in Estonia in the years to come.

After abandoning generative grammar in the very beginning of the 1990s, the prevailing framework has been functional, more specifically 'cognitive' (see e.g. the proceedings of the two conferences "Theoretical Linguistics in Estonia" (Pajusalu et al. 2002; Tragel, Õim 2006)). Research in formal (mathematically describable) properties of language has not been excluded either, and there have been some outstanding results in this field (e.g. Luuk 2010; Luuk, Luuk 2011).

The second decisive period in Estonian linguistics was in the middle of the 1990s when it became possible for Estonian linguistics, above all in the field of language technology, to participate in research projects of the European Union in the framework of the Copernicus program (with the Framework programs to follow). Estonian linguists had been carrying out work on language processing already since the 1980s, e.g. in morphological and syntactic analysis or building a thesaurus for an information retrieval system. There was some experience, but that was limited and sketchy. A major project that started in the end of the 1980s was the compilation of the computer corpus of Estonian. Today, Estonian linguists are involved in various international projects where they have access to knowledge and experience of the whole of Europe.

New research projects have dealt with the critical sub-areas of language description in the context of their time: speech analysis-synthesis, morphological and syntactic analysis, building computer lexicons (and dictionaries using computers), semantic databases. Those projects laid the basis for contacts within technological regional programs. At present, for instance, Estonian linguists are in regular contact and participate in several kinds of events, including regularly organised conferences, both with other Baltic States (conferences named "Human Language Technologies: The Baltic Perspective"; (see e.g. Baltic HLT 2010)), and with Nordic countries represented by NEALT (Northern European Association for Language Technology), which organises regular NODALIDA conferences (in 2007 one of those took place in Tartu, see http://omilia.uio.no/nealt). Finally, those projects gave a decisive impulse to start 'organisational' activities here in Estonia, from recruiting students to activities on the state level and launching language technology state programs for the periods 2006–2010 and 2011–2017 (see www.keeletehnoloogia.ee).

The following section will focus on two research fields of Estonian – language variation and change, and language technology. The overview of the studies in language variation and change is intentionally chosen to demonstrate how the oldest branches of Estonian linguistics have modernised over the last decades. The treatment of language technology will characterise the main challenges of the further research of Estonian.

Studies in language variation and change

The study of dialects and language history are often considered to be the most conservative branches of linguistics. Yet, it is from the study of language variation and change that several new fields of linguistics have developed. Dialectology forms the basis of modern quantitative sociolinguistics as well as several branches of corpus linguistics such as dialectometrics. Historical linguistics has provided input for the development of the study of language contacts and the comparative typology of languages. Both internationally and in Estonia, these two disciplines of linguistics converge in the study of the dynamics of languages.

The study of variation in Estonian

Estonian linguistics and the study of the Estonian language developed in the 19th and the beginning of the 20th century largely as comparative-historical linguistics and the study of dialects. Almost all the professors of the Estonian language at the University of Tartu have investigated the history of Estonian until quite recently despite the fact that the 1960s witnessed the advent of generative grammar and other trends of modern linguistics in Estonia (see Erelt 2003; Kasik 2011; Pajusalu 2003ab). Even the language reforms of the first half of the 20th century had direct connection with language history and dialectology. Estonian language innovators considered the closely related Finnish language as a natural source for developing Estonian. In the early 20th century the neologistic movement for modernising Estonian paid much attention to dialects which were seen as a treasure trove of the written language.

It is hard to overestimate the role of Andrus Saareste in the study of the history and dialects of Estonian. Saareste, who worked as Professor of the Estonian language at the University of Tartu in 1924–1941, laid the foundation of several directions in the study of Estonian. His research, however, focussed on dialectology, and it was he who introduced the methods of language geography to Estonian linguistics. At the same time, Saareste was interested in the deeper causes of language variation, and followed and introduced the standpoints of psycholinguistics of the time, and developed the conceptual analysis of Estonian vocabulary.

A new breakthrough in Estonian dialectology occurred in the end of the 1970s and in the 1980s when on the one hand, Estonian dialects came to be studied using the quantitative methods of dialectometrics, and on the other hand the study of the levelling of the dialects in Estonia united dialectology with sociolinguistics. One of the characteristics of Estonian dialectology has also been the close connection between the study of dialects and place names. Valdek Pall, who was the head of the dialectology division at the Institute of the Estonian Language in 1975–1990, made his name in Estonian linguistics as a moderniser of the study of Estonian place names. Marja Kallasmaa, who in the past decade published monographic studies about place names on the islands of western Estonian and in West Estonia (Kallasmaa 2003, 2010; etc.), and other researchers of place names have also collaborated with dialectologists.

In addition to the strengthening of the general sociolinguistic angle, the past couple of decades have witnessed the emergence of ethnolinguistics and the study of on-going linguistic changes. An important place is occupied by the updating of dialect databases and the creation of several new corpora and databases representing the language usage of different periods, modern colloquial language, Internet language, and other language variants, which enable studies applying automatic searches and extensive quantitative analysis. In these research areas, only preliminary work has been carried out so far; further development of these areas will be the work of the coming years.

Today, one of the priorities of Estonian dialectology is finishing of the compilation of the dictionary of Estonian dialects. The work has been complicated by the generation change among researchers at the Institute of the Estonian Language. The next work to be undertaken is the comparative grammar of Estonian dialects, which would provide an overview of the historical richness of Estonian vernaculars.

As a prerequisite of this work a dialect corpus has been under compilation since the end of the 1990s in collaboration of the University of Tartu with the Institute of the Estonian Language. The corpus is based on recordings whose transliterations serve as a basis for a general corpus in phonetic transcription. Dialectal texts in phonetic transcription have automatically been transferred into simplified syntactic transcription and transcribed on the word level. The transcribed corpus makes it possible to search for lexemes, grammatical forms and other information about different dialects. The dialectal corpus has already enabled the compilation of phonological overviews of the Estonian dialects, and the study of several grammatical phenomena in Estonian dialects. In the following years it is planned to compile systematic descriptions of the main phonetic and grammatical traits of Estonian dialects on the basis of the dialectal database.

In the last decade, several experimental phonetic studies of the sound systems of Estonian dialects were carried out. Topics investigated include the realisation of phonological quantities in different regional variants, and vowels and consonants in insular and South Estonian dialects. In addition to acoustic-phonetic studies, perception tests were conducted which show, for instance, that there are considerable differences in the quantity perception of western and eastern Estonians: besides durational differences western Estonians also rely on the pitch in the perception of second and third quantity degrees (Lippus, Pajusalu 2009). Two doctoral theses on related topics are currently in the making.

The speech of middle-aged Estonians from different parts of Estonia was recorded in the phonetic corpus of spontaneous Estonian. Previously, research in phonetics was mainly based on test-sentences which were easier to analyse, but nowadays technological developments have made it possible to analyse spontaneous colloquial speech. A better knowledge about various language varieties and registers facilitates the implementation of more sophisticated language technological applications.

It thence transpires that after a low point in the 1990s due to economic conditions Estonian phonetics is again on the rise. There are three active working groups that are based in the Institute of Cybernetics at the Tallinn University of Technology, the Institute of the Estonian Language, and the University of Tartu. The work at the Institute of Cybernetics and the Institute of the Estonian Language is more technological in its nature focusing on speech synthesis and recognition, and the development of practical aids. Additionally, at the Institute of the Estonian Language the compilation of a corpus of emotional speech is led by Hille Pajupuu, and the acoustic correlates of different emotions of Estonian speech are being studied in order to improve applications of Estonian speech synthesis. At the phonetics laboratory of the University of Tartu, the acoustics of dialects and spontaneous speech, and speech perception are investigated. The results of this work will be used for the compilation of general description of Estonian prosody and sound system.

The study of language variation is currently closely related to the study of spoken Estonian. The study of spoken colloquial language, which is carried out by the working group for spoken language and computational linguistics at the University of Tartu, led by Tiit Hennoste, has provided important new data about the structure of Estonian in general. This working group also collaborates closely with speech technologists at the University of Tartu in order to develop technological applications for Estonian such as for instance dialogue systems with automatic response.

An important task is the description of the full set of Estonian variants, i.e. the system of sub-languages. Language forms characterising the social status of the speaker – the so-called sociolects – are not clearly developed in the Estonian society. On the other hand, the standard language holds a very strong position. Still, several general cases of variation occur in the colloquial language usage, such as the interchanging use of *nud*- and *nd*-suffixes, e.g. variation of *olnud* ~ *olnd* 'been', or the word-initial *h*, which occur in principle in the speech of all native speakers of Estonian, but whose extent and conditions of usage depend on the dialectal background of the speaker as well as on social factors and those characterising the speech situation. Such cases of variation need to be described in a new type of grammar of colloquial Estonian.

Close ties between sociolinguistics and the study of language variation become apparent in the sociolinguistic investigations of the language as well as in the analysis of different types of texts. Since the 1990s a rapid development of text linguistics and discourse analysis has taken place in Estonia, led by Reet Kasik, associate professor at the University of Tartu. Here, critical linguistics, which deals with the ideological use of language, has played an important role. A series of volumes called 'exts and backgrounds', which was created upon initiative of Kasik in 2002, presents work in this field. Six books have appeared in this series so far, and monographic overviews are in preparation.

A new area which links language variation with ethnolinguistics is the so-called 'folk linguistics', which investigates how people themselves perceive and describe different varieties, borders between dialects, how they name different language forms, etc. In the past years, work has been published which analyses these aspects in certain areas of Estonian (e.g. Koreinik, Pajusalu 2007), although such work should be extended including the whole of Estonia in the coming years.

The linking of research in language variation and language change is facilitated by the comparison of different types of language data from different periods. In the case of Estonian this has been possible thanks to the extensive corpora of old written language and the written language of the 20th century which have been compiled since the 1990s at the University of Tartu and the Institute of the Estonian Language. These two large areas of linguistic study are linked best in the study of real time changes. An example of this is the dissertation of Mari Mets (2010), which follows the change in language use of the inhabitants of a region in Võrumaa over fifteen years, relating linguistic data with the analysis of social networks. Similar work is planned about other areas of Estonia.

The core of dialectology is still the language geography and dialectometrics which uses complex statistical analyses. Both language geographic and dialectometric studies, which in addition to the Estonian language area also include neighbouring countries, have offered valuable information about the relationship of Estonian with other Finnic languages, and enabled the drawing of new historic conclusions about language history. A landmark for the whole study of Finnic languages is the three-part Finnic language atlas (ALFE 2004–2010), which was compiled by linguists from Finland, Karelia, and Estonia (from the Institute of the Estonian Language and the University of Tartu), and where the editor-in-chief of the second volume was Tiit-Rein Viitso. Our knowledge about the relationships between Estonian and Livonian has been furthered by dialectometric studies. For instance, Arvo Kirkmann and Karl Pajusalu together with professor Eberhard Winkler from the University of Göttingen have published a study in relationships between the vocabulary of Estonian dialects and Salaca Livonian (Pajusalu et al. 2009), which showed that the South Estonian linguistic enclaves situated in Latvia, in particular Leivu, are much more strongly linked to Salaca Livonian than other Estonian dialects. This suggests the long-term links of Estonian and Livonian varieties in the territory of Northern and Central Latvia. In addition to Salaca Livonian, dialectometric comparisons with Courland Livonian and Votic are being carried out in order to clarify the historical development of Estonian in the context of closely related languages.

The study of language change

The first scientific studies in the history of Estonian were conducted in the 19th century after the birth of comparative-historical linguistics. Initially, the development of Estonian vocabulary and sound structure were studied. In that work the role of Finnish linguists was vital. In the 20th century, the study of the history of the Estonian language was increasingly focused on the development of phonological and grammatical structure, but general treatments were published only in the 1960s and 1970s (e.g. Kask 1967; Rätsep 1977-1979), and the first comprehensive overview of the phonological and morphological development of Estonian has appeared only recently (Viitso 2003, 2008). There is, however, no general monographic treatment of the history of the Estonian language which would present a systematic overview of the phonological and grammatical development of Estonian. According to the development plan of the Estonian language its compilation is planned for 2017.

A characteristic feature of Estonian linguistics is that there has been a working link between typological and historical language treatments, despite the fact that in other countries these areas of study have often been considered as opposites. A study by Tiit-Rein Viitso about typological change in Estonian (1991) was one of the first in that field.

The study of the history of Estonian has always taken into account data from related as well neighbouring languages. Typological studies, started by

Mati Erelt in the 1990s, resulted in five volumes in the series “Estonian: Typological Studies” which appeared in 1996–2001. Those books analysed the typological characteristics of Estonian as a language in the Baltic language area, paying more attention than before to traits in common with the neighbouring Indo-European languages, in particular to the morphosyntactic phenomena shared with the Baltic languages. Although Estonian belongs to a language family different from that of Latvian and Lithuanian, a millennia-long contact has caused interesting common developments and typological similarities. Therefore, even the Estonian three-way quantity system could be interpreted as a unique result of language contacts, where both Germanic and Baltic languages have played their role. The typological description of Estonian, which so far has remained at the level of single studies, also needs to be presented as a monograph in the nearest future.

In formal linguistics of the 1990s, linguistic optimality theory emerged after the research methodology of generative grammar somewhat lagged. That theory developed several approaches of generative grammar to language analysis, for instance, the distinction of deep and surface structure, but was better suited than earlier generative methods to the creation of different typologies as well as to the explanation of language change. In Estonian, optimality theory has successfully been used for the explanation of historical sound changes (e.g. Prillop 2006), and a monograph is in preparation. Internationally, optimality theory has also become important in describing grammar and semantics. In Estonian such studies will be conducted in the coming years.

In explaining language change it is not sufficient to observe internal regularities of the language structure and possible language contacts. It is vital to take into account changes in culture and in the world view as expressed in language. Such cultural influences are today studied by several schools of linguistics. A cognitive and ethnolinguistic approach was introduced in Estonian linguistics by Haldur Õim in the beginning of the 1990s, and language anthropological approach by Urmas Sutrop in the end of the 1990s (see Sutrop 2002). In both of these fields, extensive studies are being carried out.

In the study of the history of Estonian in the past decade, traditional branches were also thriving. There has been an increase in the etymological study of words, and an etymological dictionary meant for a wider audience is being compiled at the Institute of the Estonian Language. In the coming years an academic etymological dictionary of several volumes will be compiled. There are many new investigations of the history of the Estonian written language: several major publications have appeared, produced by the working group of the old written language at the Institute of the Estonian Language, led by Kristiina Ross, as well as by the University of Tartu working group, led by Külli Habicht. Researches of the old written language have also given their contribution to the study of the latest language change. In the coming years a comprehensive treatment of the history of the Estonian written language will be compiled, followed by treatments of older language history, and the language of folk songs.

Language technology

The work on the computer processing of Estonian started in the 1980s. There were two clearly distinguishable branches: one focusing on the language processing proper (e.g. morphological or syntactic analysis), and the other on the modelling of understanding language as a whole. The latter included morphological and syntactic analyses, but the main focus was on the processes of understanding the meaning of texts as the final result. In reality, that was the field of artificial intelligence. Despite poor resources and little experience work in both of these branches was fruitful. Nowadays, it is only the modelling of human-computer dialogue that represents the branch of artificial intelligence in Estonian computation linguistics (see e.g. Õim, Saluveer 1985; Koit 2010). For the rest, work on computational linguistics and language technology concentrates on developing programs for the processing of Estonian speech, words, phrases and sentences, and creating language resources (corpora, lexicons, and databases) required for this work.

The main goal in developing language technological applications, as formulated in official documents (e.g. EKA), is to reach the level which would enable the Estonian language to function successfully in the contemporary information society. To achieve this goal the National Program for Estonian Language Technology (2006–2010) and its follow-up program Estonian Language Technology (2011–2017) were launched. The work of the Centre of Excellence of Computer Science (2008–2015) at the Institute of Cybernetics of Tallinn Technical University also includes researchers from the University of Tartu under the state financed theme "Creating formalisms and effective algorithms for natural language processing and their application to the Estonian language" (led by Mare Koit), speech technology projects (led by Einar Meister) carried out at the speech technology laboratory of the Institute of Cybernetics. Directly connected with language technology is the state financed theme "Basic research in Estonian in the service of language technology" (led by Meelis Mihkla) at the Institute of the Estonian Language. The language technology work in general is coordinated by the recently founded Estonian Language Resource Centre where the results and resources of all language technology research in Estonia will be archived. The present language technology research in Estonia covers practically all levels of language from phonetics to pragmatics.

Research in speech technology is carried out at the Institute of Cybernetics and the Institute of the Estonian Language. In particular, in speech recognition some successful applications have been created, e.g. a system for transcribing recorded speech (http://bark.phon.ioc.ee/webtrans/). In the area of speech synthesis active research has been going on for years at the Institute of the Estonian Language. There exists, for instance, a system for reading texts to the blind – ELTE (http://elte.eki.ee/).

The work on grammar (syntax), semantics and pragmatics is mainly carried out at the University of Tartu. In the area of grammar, a system has been built which integrates morphological disambiguation, surface syntactic analysis and a program for building dependency-based syntactic trees (http://math.ut.ee/kaili/grammatika/). On the basis of the syntactic parser a preliminary version of grammar checker has been compiled. In semantics, there are three main directions of research, intended to form one whole: semantic disambiguating (i.e. building a semantically disambiguated corpus), compiling a lexical-semantic database, or the Estonian WordNet (Orav et al. 2011; http://www.cl.ut.ee/ressursid/teksaurus/), and a semantic analysis of Estonian (simple) sentences (for a general overview see Õim et al. 2010ab). In pragmatics, the main topic is modelling natural dialogue and building on this basis interactive information retrieval systems where a computer program functions as an interactive agent, e.g. an expert in a certain domain (Koit 2010; see also Treumuth 2006, 2010).

As to the applications one of the frontier areas is machine translation. At the University of Tartu, a system of Estonian-English translation has been developed, for some years. It is based on the use of parallel corpora and statistical methods of processing (see http://masintolge.ut.ee/ and http://masintolge.ut.ee/mt-for-kids/). In addition, some information extracting programs have been created, one of which is in use in the state information portal: http://eesti.ee/. Listing all the application systems available is out scope of the present paper.

Summing up, language technology is in many ways responsible for the realisation of the whole program – the ability of Estonian to function adequately in all areas of language use in the context of modern information society. Therefore, the most important task in the next decade is to enable the use of the Estonian language in all its written and spoken forms on the same level as it is possible in languages with highly developed language technology, for instance, in the Nordic countries. This means a need to create programs for speech synthesis and speech recognition, interactive information retrieval systems, and machine translation on the level of prototype systems. A prerequisite for this task is that there should be well-working analysis-synthesis programs on morphological, syntactic, semantic and pragmatic levels of Estonian.

FUTURE RESEARCH TASKS FOR THE STUDY OF THE ESTONIAN LANGUAGE

The goals and research tasks for the study of Estonian are formulated in the Development Plan of the Estonian Language (EKA 2011). The plan emphasises that in addition to the study of language structure it is important to investigate language use in different environments and functions, different variants of Estonian, contacts with other languages, and the influence of these contacts on the changes taking place in Estonian. The adequacy of the results of these studies depends on the quality of the data used, and therefore it is vital to continue work on creating and updating relevant databases.

The main goal of this work is to reach a certain level of knowledge which would make it possible to understand and direct linguistic processes in the changing society, and create necessary aids for language users. A prerequisite for this is the existence of databases and other language resources which would have the quality, quantity and diversity needed to guarantee the effectiveness and reliability of the study of Estonian, and the successful functioning of the users of Estonian in the continuously changing information society.

The most important task in the study of the structure of written and spoken Estonian is the compilation of comprehensive scientific overviews of Estonian phonetics and phonology, morphology and syntax, and a new description of the Estonian grammar and an explanatory (semantic) dictionary, which would be based on the 'real' use of Estonian in different contexts and in different social situations (ranging from official documents to the Internet).

In the study of Estonian dialects, the main task will be the compilation of a comparative grammar of the dialects and the continuation of the work on "The Dictionary of the Estonian Dialects". In the area of language contacts and language change, the task will be to analyse the language use and attitudes in multilingual communities, and to compile comprehensive overviews of the earlier history of the Estonian language and the history of the Estonian written language as well as an extended edition of the etymological dictionary of the Estonian language, and an etymological dictionary of Estonian place names.

In the development of databases and other language resources (corpora, dictionaries, etc.) it is vital to continuously update and technologically modernise the existing databases (e.g. the corpus of Estonian dialects, the corpus of old written Estonian, and others), to create new resources depending on the needs of the study of Estonian, and to continue the integration of different databases. All the application-oriented programs described in the language technology overview are and will be under development in the next decades. This is an area which will require long-term close collaboration of linguists and language technologists.

REFERENCES

ALFE 2004-2010 = Tuomi, T. (ed.) Atlas Linguarum Fennicarum I-III. Suomalaisen Kirjallisuuden Seura, Kotimaisten Kielten Tutkimuskeskus, Helsinki.

Baltic HLT 2010 = Skadina, I., Vassiljevs, A. (eds.) Human Language Technologies: The Baltic Perspective. Proceedings of the Fourth International Conference Baltic HLT 2010. IOS Press, Amsterdam etc.

EKA 2011 = Eesti keele arengukava 2011-2017 [The Development Plan of the Estonian Language]. Eesti Keele Sihtasutus, Tallinn.

Erelt, M. 2003. Study of Estonian. Erelt, M. (ed.) Estonian Language. Eesti TA Kirjastus, Tallinn, 379–398.

Kallasmaa, M. 2003. Läänemurde loodus- ja viljelusnimed. [Nature and Cultivation Names in the Western Dialect]. Eesti Keele Sihtasutus, Tallinn.

Kallasmaa, M. 2010. Hiiumaa kohanimed. [Place Names of Hiiumaa]. Eesti Keele Sihtastus, Tallinn.

Kasik, R. 2011. Stahli mantlipärijad. Eesti keele uurimislugu. [Successors of Stahl. The Story of the Study of Estonian]. Tartu Ülikooli Kirjastus, Tartu.

Kask, A. 1967. Eesti keele ajalooline grammatika. Häälikulugu. [Historical Grammar of Estonian. Sound Structure]. Tartu.

Koit, M. 2010. Eesti dialoogikorpus ja argumenteerimisdialoogi arvutil modelleerimine. [Estonian dialogue corpus and the modelling of argument dialogue on the computer]. Keel ja Kirjandus, 4, 241-262.

Koreinik, K., Pajusalu, K. 2007. Language naming practices and linguistic identity in south-eastern Estonia. Blokland, R., Hasselblatt, C. (eds.) Language and Identity in the Finno-Ugric World. Maastricht, 192-204. (Studia Fenno-Ugrica Groningana; 4).

Lippus, P., Pajusalu, K. 2009. Regional variation in the perception of Estonian quantity. Vainio, M., Aulanko, R., Aaltonen, O. (eds.) Nordic Prosody. Proceedings of the 10th Conference, Helsinki 2008. Peter Lang Verlag, Frankfurt, 151-157.

Luuk, E. 2010. Nouns, verbs and flexibles: implications for typologies of word classes. Language Sciences, 32, 349-365.

Luuk, E., Luuk, H. 2011. The redundancy of recursion and infinity for natural language. Cognitive Processing, 12, 1, 1-11.

Mets, M. 2010. Suhtlusvõrgustikud reaalajas: võru kõnekeele varieerumine kahes Võrumaa külas. [Social networks in real time: variation of spoken Võro in two villages of Võru County]. Tartu Ülikooli Kirjastus, Tartu. (Dissertationes Philologiae Estonicae Universitatis Tartuensis; 25).

Orav, H., Kerner, K., Parm, S. 2011. Eesti Wordnet'i hetkeseisust. [The current state of the Estonian WordNet]. Keel ja Kirjandus 2, 96-106.

Pajusalu, K. 2003a. Eesti keele ajaloo uurimine. [The study of the history of Estonian]. Emakeele Seltsi Aastaraamat 48. Tallinn, 166-196.

Pajusalu, K. 2003b. Eesti murrete uurimine. [The study of Estonian dialects]. Emakeele Seltsi Aastaraamat 48. Tallinn, 197-216.

Pajusalu, K., Krikmann, A., Winkler, E. 2009. Lexical relations between Salaca Livonian and Estonian dialects. Linguistica Uralica XLV, 4, 283-298.

Pajusalu, R., et al. (eds.) 2002. Teoreetiline keeleteadus Eestis. [Theoretical Linguistics in Estonia]. Tartu Ülikooli Kirjastus, Tartu.

Prillop, K. 2006. Optimaalsusteooria. Kuidas eesti keeles klusiilid nõrgenesid ja kadusid. [Optimality Theory. How the stop consonants weakened and disappeared in Estonian]. Keel ja Kirjandus, 12, 955-972.

Rätsep, H. 1977-1979. Eesti keele ajalooline morfoloogia I-II. [Historical Morphology of Estonian]. Tartu Riiklik Ülikool, Tartu.

Sutrop, U. 2002. The Vocabulary of Sense Perception in Estonian: Structure and History. Peter Lang, Frankfurd am Main.

Tragel, I., Õim, H. (eds.) 2006. Teoreetiline keeleteadus Eestis II. [Theoretical Linguistics in Estonia II]. Tartu Ülikooli Kirjastus, Tartu.

Treumuth, M. 2006. A natural language interface to a theater information database. SPECOM'2006 Proceedings: XI International Conference Speech and Computer, St. Peterburg, 25-29 juuni, 2006. St. Peterburg, 179-181.

Treumuth, M. 2010. A framework for asynchronous dialogue systems. Skadina, I., Vasiljevs, A. (eds.) Human Language Technologies: The Baltic Perspective. Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam etc., 107-114.

Viitso, T.-R. 1990. Eesti keele kujunemine flekteerivaks keeleks. [The development of Estonian into an inflecting language]. Keel ja Kirjandus, 8, 456-461; 9, 542-548.

Viitso, T.-R. 2003. Rise and development of the Estonian language. Erelt, M. (ed.) Estonian Language. Eesti TA Kirjastus, Tallinn, 130-230.

Viitso, T.-R. 2008. Liivi keel ja läänemeresoome keelemaastikud. [The Livonian Language and the Linguistic Landscapes of Finnic]. Eesti Keele Sihtasutus, Tallinn.

Õim, H. 2000. Eesti keeleteadusliku mõtte areng XX sajandil. [The development of Estonian linguistic thought during the 20th century]. Keel ja Kirjandus, 7, 468-479.

Õim, H., et al. 2010a. Main trends in semantic research of Estonian language technology. Skadina, I., Vasiljevs, A. (eds.) Human Language Technologies: The Baltic Perspective. Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam etc., 201-207.

Õim, H., et al. 2010b. Semantic analysis of sentences: the Estonian experience. Skadina, I., Vasiljevs, A. (eds.) Human Language Technologies: The Baltic Perspective. Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam etc., 208-213.

Õim, H., Saluveer, M. 1985. Frames in linguistic description. Quaderni di Semantica, 6, 2, 282-292.

Õim, H., Tragel, I. 2007. Teoreetilise keeleteaduse arengust mujal ja meil XXI sajandi alguses. [The development of theoretical linguistics here and elsewhere in the beginning of the 21st century]. Keel ja Kirjandus, 2, 98-115.

THE HUMANITIES IN ESTONIA: CURRENT TRENDS AND FUTURE PERSPECTIVES IN ARCHAEOLOGY, HISTORY, ETHNOLOGY, FOLKLORISTICS, LITERATURE AND SEMIOTICS

V. Lang, A. Selart, A. Leete, Ü. Valk, A. Merilai, K. Kull

INTRODUCTION

The continuous tradition of research and higher education in the humanities in what is now Estonia goes in many branches back as far as the early 19th century when the University of Tartu (UT) was reopened (1802). Even today that university plays the most significant role in the development of the humanities, although some other institutions (Tallinn University (TLU) in particular) have followed suit. In the UT, the humanities discussed in this paper are grouped under the umbrella of the Faculty of Philosophy; that is, in the institutes of History and Archaeology (IHA), Philosophy and Semiotics (IPS), and Cultural Research and Fine Arts (ICR). In the TLU they are in the Institute of History (IH), Estonian Institute of the Humanities (EIH), and in the institutes of the Estonian Language and Culture (ELC), and Fine Arts.

Not all the humanities beyond linguistics can be reflected in this paper. As a core, we chose four disciplines, which in 2008 were selected as part of the Centre of Excellence in Cultural Theory, financed by the European Union through the European Development Fund: archaeology, ethnology (and anthropology), folkloristics, and semiotics. In addition, research into history and literature as essential for the humanities in Estonia was also included. From the authors, Valter Lang is responsible for archaeology, Anti Selart for history, Art Leete for ethnology/anthropology, Ülo Valk for folkloristics, Arne Merilai for literature, and Kalevi Kull for semiotics.

CURRENT SITUATION

Archaeology

There are two main research centres in the field of archaeology: (1) the Chair and 'Kabinet' of Archaeology at the UT (established 1920; currently as a department in the IHA) and (2) the Department of Archaeology of the IH at TLU (est. in 1947). In addition, several archaeologists are also employed in both central and local museums (e.g. the Estonian History Museum in Tallinn) and in the Estonian Heritage Board (in Tallinn and the counties).

In the UT there are two full professorships, one in (general) archaeology (est. in 1920; since 1999 held by Valter Lang) and the other in laboratory archaeology (est. in 2007, held by Aivar Kriiska). The 'Kabinet' of Archaeology, which is taking care of archaeological collections, library and archives, is headed by Heiki Valk (since 1993). The main research is concentrating on target financed theme "Social, economical and cultural processes in Estonia in the prehistoric, medieval and modern times", which aims at analysing the whole archaeological record gathered in what is today Estonia on a new, contemporary and internationally acknowledged level. That means that the actual everyday research is carried out on themes involving the earliest prehistoric times until the modern times in the whole country. Remarkable in-depth investigations are currently in progress in the fields of the Stone Age settlement patterns, palaeogeographical environments and burial customs, social and cultural developments in the Bronze and Early Iron Ages, the Iron Age power centres in southern Estonia, urban and industrial archaeology of the medieval and modern times, prehistoric and historical field systems, textiles and weaponry, the Late Iron Age burial customs, medieval rural culture, the secondary use of prehistoric artefacts, ethical questions in archaeology, etc.

The Archaeology Department at the IH at TLU consists of three units. The Archaeology Museum (headed by Ülle Tamla) possesses the largest archaeological collections in Estonia (*ca* 1.3 million inventory units counted) and has a permanent exhibition on prehistory. The research topics of the museum are mostly connected with artefact studies (e.g. ancient silver, post-medieval hoards, amber artefacts). The unit of archaeology (headed by Erki Russow) together with the chair of archaeology (Marika Mägi) is mostly dealing with historical archaeology (particularly urban archaeology and archaeology of buildings), landscape archaeology, and social and gender archaeology. The Laboratory of Geoarchaeology and Ancient Technology (headed by Lembi Lõugas) has the target financed theme "Reflection of the development of human society in natural environment, ancient technology and archaeobiological material, and its tracing by interdisciplinary methods". This lab is mostly dealing with environmental archaeology, palaeo-anthropology, palaeozoology, palynology, ancient metallurgy and technology, etc.

On the national level, the research groups of both universities and main museums are closely collaborating when excavating sites and studying different scientific problems. The 'Kabinet' of Archaeology (UT) has for many years dealt with the composing of all-Estonian archaeological place-related database, the use of which is open for academic purposes for all researchers via the server in National Heritage Board. Another database concerning archaeological finds is being composed in the Archaeology Museum of the IH (TLU). All-Estonian cooperation is also going on in the state program "Natural Holy Places" lead by the 'Kabinet' of Archaeology.

History

The main centres of historical research in Estonia are the IHA of the UT, and the IH of the TLU. In the 1990s, the long-established academic tradition of historical studies in Tartu adapted to the international standards and developed into the central academic institution of historical research in Estonia. The formerly soviet-style IH of the Academy of Sciences of the Estonian SSR in Tallinn functioned as an independent research organisation in 1997–2005; in 2005 it became a part of the newly launched TLU. In addition to those two universities there are groups of historians employed in archives (Estonian Historical Archives, Estonian State Archives, Tallinn City Archives), and several museums (Estonian History Museum, Art Museum of Estonia). The Estonian Institute of Historical Memory was established in 2008 to study the Soviet period (1940/1941, 1944–1991) in Estonia, concentrating on Soviet repressive policy.

In the UT, the research into the 20th century history is quantitatively dominating. The focal points are the interwar Estonia and the Soviet period in Estonia, especially in the 1940s and 1950s based on mostly local and Russian archival sources. The target financed theme "Estonia in the Era of the Cold War" assembles a number of scholars of the relevant research. Notable research is provided into Early Modern cultural and social history within the target financed themes "Imperial pressure in the Baltic provinces from the early 16th to the middle 19th century" and "The diffusion and reception of ideas and theories in the Baltic provinces of Swedish and Russian empires (17th C. – early 20th C.) ".

Well established during the last decades have been the medieval studies. Most of the Estonian medievalists participate at the work of inter-institutional and interdisciplinary Centre of Medieval Studies founded in 2005 at the TLU (currently headed by Juhan Kreem). The medieval research focuses mainly on topics like urban or ecclesiastic history. Several active scholars collaborate with the project "Christianization, Colonization and Cultural Exchange: The historical origins of the European identity of Estonia (13th–17th C.)".

The formerly (in the 1960s-1980s) popular agrarian history has lost its central position on the research landscape. Some decline is also evident in the study of social and economic history of the late 19th C. Main respective activities are concentrated in the target financed theme "Adapting to modernity.

The Estonian society's response to political, social, economical and cultural challenges in times of transformation (16th–20th centuries)" (TLU). Some research is also carried on in the topic of agrarian migration of Estonians to Siberia around 1900.

Ethnology and anthropology

Major institutions of ethnology and anthropology in Estonia are the Estonian National Museum (ENM) (est. in 1909), the Department of Ethnology (assistant professorship in 1924, and professorship since 1939, currently held by Art Leete) at the ICR (UT), and the EIH at TLU, where the Department of Cultural and Social Anthropology (DCSA; Lorenzo Cañás Bottos) was established in 2006.

Regional areas of expertise of the Estonian ethnologists and anthropologists are Estonia, Russia (predominantly the Finno-Ugric peoples but also the peoples of Siberia), South and Central America, Caribbean, West Africa, India, England, Latvia, Ireland, Germany, and Scandinavia.

Research themes have revolved around issues such as landscape and culture, indigenous studies, colonialism and post-colonialism, religion, political anthropology, performance studies, non-human anthropology, cinema, ethnic and cultural identity, everyday culture in diachronic and synchronic perspectives, life-stories, subsistence technologies, social and culture change, heritage politics, and museum communication.

The ENM's recent research grants from the Estonian Science Foundation (EstSF) have been concentrated on the ethnicity, research and politics of Finno-Ugric peoples and developing museums in the 21st century information environment. The ESF grant of the DCSA is related to research in nation-building and state-making. That department is also to submit a target financed project that will examine the complex relationships between space, body, politics, and culture. The Department of Ethnology (UT) carries out several ESF grants on religious change and post-colonialism, knowledge production in the context of heritage scholarship, and memory practices in post-Soviet Estonia. The department also makes team effort in the framework of target financed project "Dynamic Perspectives of Identity Politics – Analysis of Dialogue and Conflict".

Folkloristics

Folklore research emerged in the 19th century Estonia but the academic training with professorship started as late as in 1919 (UT). The Estonian Folklore Archives (founded in 1927) have collected the oral traditions of Estonians and the local ethnic minorities. In 1947 the Institute of Literature and

Language was established together with the department of folklore, which became the main centre of research in the Soviet time. Those three institutions prepared the later landscape of folklore research, nowadays consisting of the Department of Estonian and Comparative Folklore (UT), and the Estonian Literary Museum in Tartu with the Department of Folkloristics and the Estonian Folklore Archives. The Estonian Literary Museum also includes the Department of Ethnomusicology, whose field of studies is closely related to folkloristics.

Whereas early scholars saw folklore as a survival of the traditional peasant culture and predicted its rapid disappearance, contemporary research has followed the international developments and conceptualises folklore as an expressive vernacular culture, shaped by tradition but always involving creativity. Hence, folklore has not disappeared with the emergence of written literature, but thrives in contemporary culture, including mass media, internet, and film; and in institutional forms such as schools, clubs, societies and professional circles. Therefore, the tasks of folklorists have expanded from research in obsolete forms of folklore to documenting and analysing its contemporary expressive forms. Preparing academic editions of folk songs, folk narratives, short forms and other genres, published as "Monumenta Estoniae Antiquae", has continued since the late 19th century, when Jakob Hurt started the series. The last volume of that grand project is the anthology of Estonian fairy-tales (2009).

During the last decades the importance of analytical research and folklore theory has been gradually growing and become prevalent. The Estonian Folklore Archives coordinates the target financed research project "Folklore and Folklore Collections in Cultural Changes: Ideologies, Adaptation and Application Context". The Department of Folkloristics, Estonian Literary Museum, is responsible for the target financed project "Narrative Aspects of Folklore: Power, Personality and Globalization". The Department of Estonian and Comparative Folklore (UT; Ülo Valk) carries out research, funded by the target financed project "Folklore and Society: Tradition Memory, Creativity, Applications". Academician Arvo Krikmann (Estonian Literary Museum) is one of the leading paremiologists of the world who has been coordinating projects on proverbs, riddles and other short forms of folklore. The Estonian Literary Museum is carrying out research in phrases, folk calendar, ancient Estonian religion, folktales, urban legends and rumours, traditions of Estonian settlements in Russia, charms, beliefs and folk medicine, contemporary folk humour, the folklore and mythology of Udmurtians and other Finno-Ugric peoples, the dynamics of Estonian folk song traditions, Baltic-Finnic folklore, folk customs, the history of research in folk medicine. The research team at the UT has done research in Estonian epic songs and Vepsian folklore, Setu traditions, folksongs and oral history, cultural heritage and representation, belief narratives and memory, charms and history of British folklore research, belief genres in Estonia and India, Votian folk religion and its historiography. The researchers at the TLU have been working on Estonian death culture and folk beliefs.

Literature

The Estonian literary research institutions collect, study, and provide instruction in the interpretation of literary phenomena, thus enriching our national culture and presenting it to the world. There are four main centres of disciplinary expertise: the ICR at the UT, the ELC at TLU, the Estonian Literary Museum (ELM) in Tartu and the Under and Tuglas Literature Centre (UTLC) of the Estonian Academy of Sciences in Tallinn. In addition, literary research is also carried out at several other institutions of philology, semiotics and cultural studies at the UT or TLU, as well as by independent scholars.

As for the research at the UT, the world literature and its reception in Estonia are studied in a comparative context under the supervision of Jüri Talvet. The history and poetics of Estonian literature, as well as literary philosophy (based on an innovative pragmapoetics), are the domain of Arne Merilai and his junior colleagues. The research in theatre studies (Anneli Saro et al.) focuses on methods of performance analysis, local theatre history and reception. Liina Lukas, in cooperation with UTCL, is responsible for Baltic-German written heritage and the online publishing project EEVA. The department receives financial support from the target financed project "World Literature in Estonian Culture. Estonian Literature in World Literature (a Historical and Modern Paradigm)".

Traditionally, literary scholarship constitutes an area of emphasis at the Institute of Germanic, Romance and Slavonic Languages and Literatures, as well as at the IPS (UT). Main fields of research in the Department of Classical Philology (Anne Lill) are 17th century Estonian poetry, classical metrics, rhetoric, drama studies and narratology, and the comparative study of ancient and modern poetry. Research projects are normally funded by grants of the EstSF and national research programs. Principal subjects of the Institute of Slavonic Philology (Ljubov Kisseljova) are the Western periphery of the Russian Empire, the formation of a Russian literary canon, Estonian–Russian cultural connections, the interrelations between literature and ideology, and the semiotics of Russian literary poetics. Research in this area is funded by the target financed project "Reception of Russian Literature in Estonia in the 20th century: from Interpretation to Translation", as well as grants provided by the EstSF.

In addition to Russian philology, cultural semiotic and formal analysis of poetics complements the intellectual landscape of the Department of Semiotics. In the Department of Philosophy, literary phenomena are occasionally part of the disciplinary discourse in phenomenological, aesthetical or analytical philosophical terms.

All Estonian publications, encompassing "Estica", "Baltica" and "Fenno-ugristica", as well as literary and cultural source materials (manuscripts, documents, photos, etc.) are collected and stored in the Estonian Literary Museum. The Research Group of Cultural Sources and Literature studies manuscripts, history of Estonian literature, culture and education, Estonian–Finnish literary relationships, women's literature, and develops online databases (ELLEN, ERNI). The Research Group of Cultural and Literary Theory focuses on comparative and semiotic poetics of Estonian culture. The activities of ELM include organising conferences on national literature, and publishing research results in multiple issues. It relies mainly on two target financed projects, "Sources of Cultural History and Contexts of Literature" and "Rhetorical Patterns of Mimesis and Estonian Textual Culture", as well as on the grants of the EstSF and national research programs.

The interdisciplinary aim of the UTLC (headed by Academician Jaan Undusk) is to study Estonian literature, drama, historiographical, philosophical and public discourses within the context of the multilingual Baltic space. The Museum of the UTLC oversees the cultural property of writers F.Tuglas, M.Under and A.Adson, prepares critical editions of their oeuvre, and provides support to guest researchers. Besides the grants of the EstSF and Estonian Cultural Heritage, the UTLC runs the target financed project "Autogenesis and Transfer: The Development of Modern Culture in Estonia". A Network of Science and Literature Studies has been established in collaboration with the University of Tartu Library.

The main area of research of the Department of Literary Research in the ELC at Tallinn University (Piret Viires and Tiina Aunin) is Estonian literary history in a comparative and theoretical context. The Institute publishes its proceedings and, in cooperation with ELM, the interdisciplinary series of monographs "Heuremata". The operations are funded by the target financed project "Rhetorical Patterns of Mimesis and Estonian Textual Culture" in collaboration with the Research Group of Cultural and Literary Theory at ELM.

The philologists of foreign languages at the TLU, inspired by former rector Rein Raud, represent also linguistic and semiotic literary scholarship studying Italian, French, English, German, Slavic, Near and Far Eastern and other literatures. Broadly international translation research is fostered in the Translation Centre of the Institute of Germanic and Romance Languages and Cultures.

Semiotics

Semiotics as a broader field of science, covering the study of meaning-making mechanisms everywhere in culture and life, started to be institutionalised since the 1960s when semiotics got a strong development in Tartu, first in the Chair of Russian Literature with Juri Lotman, and since 1992 as a separate semiotics department in the University of Tartu, led initially by. Igor Černov, then by Peeter Torop, and since 2006 by Kalevi Kull. In 2007, the Department of Semiotics became a part of the IPS. The Department of Semiotics in Tartu has got a high international reputation both due to its scholarly work and its attractive international teaching program, and has become one of the semiotics centres on the world scale.

Besides the Department of Semiotics in Tartu, semiotics is taught in the EIH (TLU), and as a special course in some other universities of the country. Of other semiotics-related institutions, there exists Jakob von Uexküll Centre as a unit of the Estonian Naturalists' Society (est. in 1993), the Estonian Semiotics Association (1998), Juri Lotman Fund (2001), and the Estonian Semiotics Repository (2007). Today all these institutions form a co-working network.

Recent theoretical work in semiotics carried on in Estonia includes a wide variety of subjects, among these general methodological problems, semiotics of culture, sociosemiotics, and biosemiotics. Studies in general semiotic theory include the analysis of general phenomena of translation and autocommunication, as well as the lower semiotic threshold zone, and the relationship between semiotics and other sciences (the differentiation into sigma-sciences and phi-sciences). It concerns also the distinctions between major levels of semiosic processes, or relationships between semiotics of culture and biosemiotics, between languages and other sign systems.

Studies in cultural semiotics have in large part been focused on explication and systematisation of methodology and results of Tartu-Moscow school. This also includes the inquiry, commenting and republication of Juri Lotman's heritage. Additional topics include semiotics of literature, semiotics of naming; semiotics of verse; analysis of the concept of emblem; political semiotics; semiotics of film, theatre, performance, and media; semiotics of translation, semiotics of landscape, and study of cultural autocommunication. Special mention deserves semiotics of translation, including studies of intercultural translation and translatability.

Remarkable development goes on in the field of sociosemiotics, which has focused particularly on semiotics of space and semiotics of city, as well as semiotics of ideological systems.

Tartu is a well-established centre of the contemporary biosemiotics. Departing from the works of Jakob von Uexküll and Thomas A. Sebeok and developing the field within the intellectual atmosphere of Tartu school, Tartu biosemiotics emphasises perspective of an animal subject, organismic and ecological levels in semiotic processes as well as mutually enriching relations between biology and cultural semiotics, especially via ecosemiotic studies. The biosemiotic research includes semiotics of biological mimicry, theory and history of zoosemiotics, legacy of Jakob von Uexküll and Thomas A.Sebeok, semiotic mechanisms of evolution and speciation. In addition to J.Lotman's models, R.Jakobson's and J.Uexküll's models are widely used in the contemporary research carried out by Estonian semioticians.

From 1960s to 1980s, Russian culture served as the main focus for analysis by the Tartu cultural semiotics group. Since the 1990s, the situation has radically changed – the major object of study is now Estonian culture. The objects of semiotic analysis have included Estonian poetry, monuments, food traditions, films, tourist performance, poetic speech, nature writing, etc. Many student projects in Tartu semiotics department include analysis of empirical material, e.g., study of film, naming, maps, literature, theatre, etc. Ecosemiotic studies have dealt with some examples of nature management, historical plant use and landscape design. Zoosemiotic work includes observations in Tallinn Zoo, as well as dog training analysis.

One of the strong points of infrastructure is the rich and valuable semiotics library resource that Estonia provides and develops. Jakob von Uexküll Centre in Tartu has the collection (almost complete set) of Uexküll's works, plus some archive materials. The Estonian Semiotics Repository owns Juri Lotman's personal library. Tartu University Library stores the epistolary archive of Juri Lotman and Zara Mints. The Department of Semiotics stores the memorial biosemiotic library of Thomas A. Sebeok, which includes his complete personal collection of zoosemiotic literature.

INTERNATIONAL COLLABORATION

International collaboration in the field of the humanities has stepwise become rather intensive, particularly after the collapse of 'iron curtain' in the late 1980s. In ARCHAEOLOGY, the Chair of Archaeology (UT) launched in 2003 the Baltic Archaeological Seminar (BASE), which unites the archaeologists from Estonia, Latvia, Lithuania and Finland who gather once every two years in one of the abovementioned countries for a theoretical archaeology conference. The conference presentations are later published in a more extended form in the journal "Interarchaeologia". Outside the BASE-movement there are tight cooperation contacts with the scientists from the universities of Helsinki, Turku, Vilnius, Latvia, St. Petersburg, Warsaw, Wroclaw, and several others. For the archaeologists at UT, the following cooperation projects deserve to be highlighted: 1) the research of the early settlement of north-eastern Savo in Finland (universities of Tartu and Helsinki), 2) the detecting of the early settlement of the basins of the rivers Narva and Luuga (Tartu and St. Petersburg), 3) the problematics of textile-impressed pottery in Eastern and Northern Europe (Tartu and Helsinki), 4) the detecting and research of the fossil field systems in Western Lithuania (Tartu and Vilnius), 5) international seminar of history and archaeology "The Medieval Towns of the Baltic Sea Region", which have taken place in Tartu 6 times (1997-2000, 2002, 2005), 6) "The Ecology of Crusading" project for 2010-2013 (UK, Poland, Latvia, Estonia), 7) international scientific cooperation between Earth Sciences and Archaeology takes place within EU COST action "Submerged Prehistoric Landscapes and Archaeology of the Continental Shelf" (2010–2013), within IGCP project 526 – Risks, Resources, and Record of the Past on the Continental Shelf (2007-2011), within the project: "Coastline Changes of the Southern Baltic Sea – Past and Future Projection" (CoPaF; 2009–2012), and within EstSF project "Development of the Baltic Sea coastline in Estonia through time: palaeoreconstructions and predictions for future" in cooperation with Herzen State Pedagogical University.

The archaeologists of the IH at the TLU have a joint project "Man, Environment, Technology: A Role of the Archaeofauna in Ancient Societies" with the Institute of Evolution and Systematics of Animals at the Polish Academy of Sciences. They also participate in the project "The wild horses during the Pleistocene and Holocene: palaeo-ecology and DNA", coordinated by the Zoology Museum in Dresden and the Max-Plank Institute of Evolutionary Anthropology. They are cooperating with the archaeological and osteological laboratories, Stockholm University, for the studies of stable isotopes based on seal and fish bones from the Baltic basin, and participating in the European Monitoring Programme, an INQUA working group within the Commission on Palaeo-ecology and Human Evolution.

Estonian HISTORIANS develop wide international cooperation including collaboration between institutions and personal scholars. A group of scholars of the UT participate at the joint Scandinavian-Baltic research program "Nordic Spaces: Formation of States, Societies and Regions, Cultural Encounters, and Idea and Identity Production in Northern Europe after 1800" (Eero Medijainen). Perspective international cooperation includes participation at the projects financed by the European Science Foundation and other non-resident institutions. The Centre for Medieval Studies, for instance, is already involved in prospective research programs "Symbols that Bind and

Break Communities: Saints' Cults as Expressions of Local, Regional, National and Universalist Identities", and "Cuius Regio. An analysis of the cohesive and disruptive forces destining the attachment of groups of persons and the cohesion within regions as a historical phenomenon". Both of the projects focus on comparative research of various European regions in medieval and early modern period. The IHA (UT) participates at the joint doctorate program "Baltic Borderlands" with the universities of Greifswald (Germany) and Lund (Sweden). Promising cluster of international research is associated with the yearbook "Forschungen zur baltischen Geschichte" having close institutional links with the University of Latvia (Riga), Nordost-Institut (Lüneburg, Germany), and Baltische Historische Kommission (Göttingen, Germany).

Estonian ETHNOLOGISTS and ANTHROPOLOGISTS contribute to various research projects and academic events in Estonia, Russia, and almost all other European countries, but also the USA, Puerto Rico, and India.

The Estonian National Museum cooperates with nine museums across Europe through the project "A Taste of Europe". The project is about European food production and food consumption and concentrates on national differences and similarities, changes over time, climate and environmental issues. The project is carried out with support of the Culture Programme of the European Union. The ENM also develops the Baltic Heritage Network project "Cultural Heritage of the Estonian Diaspora" that is financed by the Ministry of Education and Research.

The major international cooperation effort of the Department of Ethnology (UT) is related to FP7 international project "European National Museums: Identity Politics, the Uses of the Past and the European Citizen" that covers almost all European countries. Several recent projects (financed by INTAS and EUROCORES BOREAS programmes) of the department have been related to development of contemporary religiosity and identity changes among different marginal ethnic groups. This earlier cooperation has served as a basis for the most recent international collaborative effort under the Centre for the Study of Pentecostal and Charismatic Movements in Russia (financed by the John Templeton Foundation through the Pentecostal and Charismatic Research Initiative).

The centres of FOLKLORISTICS are closely connected to international scholarship. In 2005 the 14th congress of the International Society for Folk Narrative Research (ISFNR) was held in Tartu, co-organised by the folklorists of the UT and the Estonian Literary Museum. Prof. Valk has worked in the executive committee of the ISFNR since 2005 (until 2009 as the president) and in this connection the folklorists of the UT have been involved in organising several forums of this leading international society. Kristin Kuutma belongs to the board of the International Society for Ethnology and Folklore (SIEF). Jonathan Roper (UT) has been leading the ISFNR committee on charms, charmers and charming, which focuses on the genres of verbal magic. International conferences and seminars have annually been organised in Estonia, such as “Stories about India and Indian Stories” (2009) and “From Language to Mind 3” (2009). Several leading folklorists of the world have been lecturing at the Department of Estonian and Comparative Folklore, including Regina Bendix (Germany), Seppo Knuuttila (Finland), Diarmuid Ó Giolláin (Ireland/ USA), Alexander Panchenko (Russia), Leander Petzoldt (Austria), Anna-Leena Siikala (Finland), Timothy Tangherlini (USA), Vilmos Voigt (Hungary), and many others.

In the LITERATURE studies, much international credit is owed to the Estonian Comparative Literature Association with its biennial conferences (No. 9 in 2011), organised in collaboration with the International Comparative Literature Association (ICLA) and the European Network of Comparative Literary Studies (ENCLS). Marina Grishakova coordinates the participation in the Nordic Network of Narrative Studies. Typically based on bilateral agreements, outreach and international cooperation of the classical philologists include numerous universities and research centres in Germany, Scandinavia, Switzerland and the Baltic states.

International cooperation also includes Slavists from Russia, the Ukraine, Finland, Germany, Israel, USA and the Baltic states. During the last 5 years, 15 international conferences have been held, five volumes of the series “Humaniora: Litterae Russica” and five collections authored by junior philologists have been published.

In SEMIOTICS, several applied projects initiated by Estonian governmental organisations have been implemented by the Department of Semiotics (UT) in recent years. These have been related to Estonian brand, Estonian national customs (Estonian cuisine), ideological and national conflicts. Our semioticians have been repeatedly asked to provide semiotic expert opinions in legal disputes, particularly concerning textual or pictorial analysis.

Many world leading semioticians have visited the semiotics department and given some lectures or lecture courses – Thomas A. Sebeok, Paul Bouissac, Terrence Deacon, John Deely, Umberto Eco, Jesper Hoffmeyer, Gunther Kress, Roland Posner, Winfried Nöth, et al. Recent visiting professors who have taught longer courses in Tartu include Myrdene Anderson, John Deely, Roger Parent, and Ahti-Veikko Pietarinen.

Estonian semiotics group is currently well-known as one of leading groups in the contemporary world semiotics. This position is not only due to the results of research, but also due to the semiotics program taught in the University of Tartu. Thus, further emphasis should certainly be accorded to support the international semiotics study program, which has a task of preparing the professors of semiotics for the universities of the world. The latter is supported by the English language doctoral and master programs in semiotics that are provided in Tartu and that have already attracted much attention worldwide.

FORESIGHTS

The ARCHAEOLOGISTS of both universities are acting for the sake of achieving better conditions for the preservation of archaeological collections, preparation of new prehistoric exhibitions (both in the Institute of History at the Tallinn University and Estonian National Museum in Tartu), better financing of everyday research, and resultant completion of ongoing scholarly projects.

In coming years the Archaeology Museum of the IH together with the University of Warsaw is planning to publish manuscript archives of Marta Schmiedehelm – an Estonian archaeologist who worked with previous East-Prussian materials in the 1920s and 1930s. The Lab of Geoarchaeology and Ancient Technology is aiming to initiate research into genetic ancestries, breeding strategies and utilisation of Estonian indigenous sheep and cattle from prehistoric to modern times in collaboration with the colleagues from the UT and Finland. This lab also plans to start the investigations with using radiology in physical anthropology and ancient technology.

In collaboration with the Institute of Material Culture of the Russian Academy of Sciences, the Pskovian Archaeological Centre, University of Latvia, and several other institutions, the 'Kabinet' and Chair of Archaeology (UT) is applying for a Est-Lat-Rus project "Archaeology, Authority, Community" (to be financed by the European Structural Fund). The application – where the UT is the leading partner – has successfully passed the first selection phase.

Important for the future of HISTORICAL study in Estonia is to preserve the multi-centred and multi-perspective research. The collaboration between institutions needs support and simultaneously a reasonable competition as well. The instability of financing caused by short term project-based funding avoids initiation of large and long-standing endeavours like source publications. It will be possible to realise this kind of major research in the framework of the Estonian archival system. Archives are making a great job in publishing digitalised sources and copying Estonia-related materials in foreign archives.

There is a need to develop the scope of international cooperation in the future. The close collaboration with colleagues in Germany, Scandinavia, and Finland should be complemented with contacts in other countries. Thus, the very close historical contact between Estonia and Poland or Russia has not found adequate institutionalised form of collaboration of historians yet.

In ETHNOLOGY and ANTHROPOLOGY, the Estonian National Museum's perspectives of development are related to its new building. It will put new challenges to researchers in developing studies in Estonian everyday culture and a situation of the Finno-Ugric peoples in Russia, but also such interdisciplinary fields as museology, including museum audience studies. For the ENM, there exists obvious need to increase the number of research staff in the fields of Estonian and Finno-Ugric ethnology (currently 16 staff members).

Due to the current needs of the museum development, the ENM temporarily increases research in the field of museology and museum communication, including applied research and networking with European research teams. The development of new permanent exhibitions also needs targeted ethnological research in specific areas, such as heritage preservation, ethnicity issues among the Finno-Ugric groups in Russia, sociocultural processes in contemporary Estonian society, childhood studies, cultural processes in the digital age, but also research in everyday life in the recent past – in Soviet times.

The DCSA (TLU) is continuing to develop and foster research links with other departments and institutes by designing target financed projects that will centre on interrogating the relationship between 'culture and place'. The future research plans are targeted to understanding the different dimensions of movement, such as migration, transnationalism, displacements, diaspora, borders, etc. and trying to find new idioms to articulate movement. The researchers of TLU and UT continue to cooperate in the Centre of Excellence in Cultural Theory during the next few years. Mutual cooperation agreement between Centre of Excellence and the ENM is signed in 2011. In addition to the current research projects and themes new directions of research are envisioned in the field of development anthropology in the UT. There are also perspectives to develop a series of applied projects in cooperation with regional institutions in Estonia in the field of preservation of cultural heritage.

The ENM, the DCSA and the Department of Ethnology (UT) have become involved in international networks of scholars in multiple fields of research. Several projects will continue in the near future and constant search for new cooperation opportunities in these fields goes on. Additionally, scholars look for new international collaborative projects that will widen the scope of research (for example, in museological research by studying development of ethnic/national ideology in different regions of Russia; and in the development of new methodologies to address experiential aspects of social life).

The future of Estonian FOLKLORISTICS is closely connected with the Estonian Literary Museum and the University of Tartu that form the institutional foundation of the discipline in Estonia. Hopefully, academic folkloristics will also be continued in Tallinn; that will mainly depend on the developments in Tallinn University. An important role will be played by the Academic Folklore Society (established in 1925, re-established in 1996) to provide common forums of discussion, such as the regular winter conferences of Estonian folklorists that have been held annually from 2006. Folklorists will also participate in the work of the Estonian Society for the Study of Religions (ESSR), which will foster their dialogue with other disciplines on international scale. (In 2006 ESSR was affiliated with the European Association for the Study of Religions and in 2010 with the International Association for the History of Religions.)

As regards the future, the development of the electronic databases and documentation of contemporary folklore will be important works to be carried on. Also, the preparation of academic editions, based on collections of the Estonian Folklore Archives will continue – both as electronic publications and in book format. Research will be continued in the poetics, contents, social functions and regional traditions of folk songs. Work on typology of folktales will be continued and the oral narrative tradition will be studied in relationship with literary culture. Also, legends and other belief narratives from the Estonian Folklore Archives will be studied in their social context, expressing the ideologies and dominant beliefs of the tradition carriers. Research in short forms of folklore (riddles, proverbs, phrases, poetic speech) and folk humour will be carried on. Among the important fields of research will be oral history, personal narratives, life-stories and place-lore. The latter focuses on the relationship between traditional communities and their living environment as it is expressed in narratives and personal memories. Research will continue in folk beliefs and vernacular religion – both from historical perspective, based on the collections of the folklore archives, and the developments in the context of New Spirituality, which shapes the current religious landscape of Estonia.

Theoretical grounds and conceptual framework of Estonian folkloristics is closely connected with international developments. The current research has strong international dimensions – ranging from the Baltic Finns and other Finno-Ugric peoples to countries, such as England and India. Also, research in ethnic minorities of Estonia has started and will be continued.

Contacts and cooperation will be developed with research centres in Finland, Germany, Hungary, India, Ireland, Latvia, Lithuania, Russia, UK, the USA, and in other countries. Ties with the leading professional organisations, such as the American Folklore Society (AFS), the ISFNR, the International Society for Ethnology and Folklore (SIEF) will also be fostered. The Department of Estonian and Comparative Folklore (UT) and its PhD program in folkloristics has a strong international orientation, and has attracted PhD students from many countries. All these developments will strengthen the international positions of Estonian folkloristics, at the same time erasing the borders between local and national research traditions.

In LITERATURE, the institutions involved in university teaching and professional research proceed from educational and scientific funding which directs and charts the course of the discipline and ensures a solid point of departure.

However, some obvious shortcomings need to be addressed. (1) The mediating role of the TLU between Estonian and world cultures calls for increased coordination and a joint fiscal structure. (2) Classical philology at the UT is in need of additional recurring funds. (3) Although the principal training of literary scholars and teachers takes place at the UT, the branches of Estonian literature and theatre research are nevertheless left without a single major grant. At the same time the potential of the ICR as a centre of expertise is remarkable. (4) To enrich its profile, the ICR should also integrate some extra-institutional scholars or researchers abroad whose scholarly interests include Estonian literature or theatre: Thomas Salumets (Canada), Cornelius Hasselblatt (the Netherlands), Maire Jaanus and Mardi Valgemäe (USA), Satu Grünthal (Finland), et al.

Coordinated by the ELM, the interdisciplinary Centre of Excellence of Literature and Arts Research is about to be applied for. The centre will focus on dynamics of cultural changes in different art media in comparative sociohistorical context. The evolution of Estonian textual, visual and musical patterns will be studied from their beginnings to the postmodern transnational era. Integrating archives and creating digital databases, the project will have an applicable output. This centre *in spe* will consist of eight targeted projects from ELM, UT, Estonian Academy of Arts and UTLC.

In SEMIOTICS, the major aims towards the development of international projects are focused, besides the ongoing project, on the development of doctoral studies curricula in semiotics, on applied semiotics (together with some commercial semiotics groups in the United Kingdom), on translation semiotics (together with Italian semioticians), on biosemiotics (with the University of Copenhagen), on zoosemiotics (Charles University of Prague), and on semiotics of literature (ELTE University Budapest), on history of semiotic ideas (University of Lausanne, Switzerland), on sociosemiotics (Metropolitan University of London), etc.

The representation of Estonian scholars in the regional and world institutions of semiotics is certainly one factor in the further development. In addition to the former Tartu-Moscow school, additional links can be pointed out, as reflected in the names like Copenhagen-Tartu nexus or Tartu-Bloomington-Copenhagen school.

MAIN JOURNALS AND SERIES IN THE HUMANITIES

The official system of research evaluation in Estonia undervalues the humanitarian research in general and publications in Estonian in particular. The tendency results amongst others in growing isolation between academic community publishing intensively in foreign languages and non-professional audience. However, the Estonian researchers have managed in establishing and managing of a number of academic journals and series, which are acknowledged regionally or even worldwide. In addition to that, many of our scholars belong to the editorial boards of other international journals. A recently established book series of the Centre of Excellence in Cultural Theory "Approaches to Culture Theory" is also worth pointing out.

The archaeologists of the universities of Tartu and Tallinn are jointly issuing the "Estonian Journal of Archaeology" (since 1997; indexed by ISI WoS and ERIH 'B') and the series "Muinasaja Teadus" (since 1991; ERIH 'C'). The third journal, "Archaeological Fieldwork in Estonia", is annually published in cooperation between the two universities and the National Heritage Board.

There are several historical periodicals edited with economic support of archives and research institutions. "Tuna. Ajalookultuuri ajakiri" (since 1998) combines popular and scholarly presentations in local and general history. "Ajalooline Ajakiri" (1922–1940, 1998–2002, 2007; ERIH 'C') publishes academic papers mostly in Estonian, but also English. The German-language yearbook "Forschungen zur baltischen Geschichte" (since 2006) is a joint project of academic and research institutions in Estonia and outside focused on study of the whole Eastern Baltic area. "Acta Historica Tallinnensia" (since 1997; ISI WoS and ERIH 'B') contains primarily but not exclusively papers by the members of the Institute of History, TLU.

"Yearbook of the Estonian National Museum" is the oldest regular ethnological series in Estonia (published since 1925). A journal "Pro Ethnologia" was also published by ENM between 1993 and 2005. ENM has published also a series of monographs since 1998. From 2003 to 2005, the departments of

Ethnology and Estonian and Comparative Folkloristics at UT published joint series “Studies in Folk Culture”. Another joint publishing project of the departments is a series “Studia Ethnologica et Folkloristica Tartuensia” (since 1998). Ethnologists at the Institute of History (TLU) have published a series “Scripta Ethnologica” since 1996.

Since 1996 the folklorists of the Estonian Literary Museum have been publishing the journal “Folklore”. An “Electronic Journal of Folklore”, which is nowadays indexed by the ISI Web of Knowledge. Another electronic journal, “Mäetagused”, was also established in 1996 for Estonian readership and has published articles from folkloristics and the related fields, including translations of foreign scholarship. Since 2007 the Estonian Literary Museum, Estonian National Museum and University of Tartu has also published the international peer-reviewed “Journal of Ethnology and Folkloristics”. Estonian Literary Museum publishes the series “Reetor” (mainly dedicated to short forms), “Sator” (folk belief and customs), “Tänapäeva folkloorist” (about contemporary folklore), “Pro Folkloristica” (including articles by young scholars) and “Eesti Rahvaluule Arhiivi toimetised” (Proceedings of the Estonian Folklore Archives). Estonian Literary Museum has also specialised in e-publications of folkloristics (http://www.folklore.ee/pubte/) and has prepared several free-access data bases of folklore (http://www.folklore.ee/ebaas/).

The Department of Literature and Theatre Research at the ICR (UT) publishes the theoretical series “Studia litteraria Estonica” (since 1996) and two peer-reviewed periodicals: internationally acclaimed “Interlitteraria” (since 1996) and “Methis” (in cooperation with ELM, since 2008). The UTLC publishes four academic series: “Collegium Litterarum” for literary history, “Oxymora” for theory, a series for commentary of modern key texts, and the cultural collections of the Museum. The main tribunes in Estonian for publishing essays of literary research and criticism have been the monthly journals “Keel ja Kirjandus”, “Looming”, “Vikerkaar”, and “Akadeemia”.

The oldest semiotics journal worldwide, “Sign Systems Studies”, established by Juri Lotman in 1964, is published by Tartu University Press and is continually one of the leading journals in the field. Other publication series in semiotics published in Estonia include, currently, the international book series Tartu Semiotics Library (since 1998), “Bibliotheca Lotmaniana” (since 2010), a periodical “Acta Semiotica Estica” (publication of the Estonian Semiotic Society, since 2001), and an electronic journal “Hortus Semioticus” (since 2006). In addition, Estonian semioticians are co-editors of the international journal “Biosemiotics” (published by Springer), and the world major book series on semiotics “Semiotics, Communication, and Cognition” (published by De Gruyter Mouton).

PSYCHOLOGY IN ESTONIA

J. Allik

A SHORT HISTORICAL TOUR

The history of psychology in Estonia demonstrates clearly that the remote corner of the world as Estonia used to be perceived constituted an integral part of the German speaking academic community in the 19th century at the most crucial moment when the scientific psychology was just emerging (Allik 2007). Hermann von Helmholtz in his inaugural lecture as Rector of the University of Berlin talked about academic freedom of German universities and stressed that one manifestation of it was a complete freedom of movement for the students and professors between all universities of the German tongue, from Dorpat to Zurich, Vienna, and Graz (Helmholtz 1896). As a confirmation of these words, the first rector after the reopening of the University of Dorpat (Tartu) in 1802, Georg Friedrich Parrot (1767–1852) was interested in optical phenomena which he attempted to explain introducing the concept of unconscious inferences, anticipating a similar theory proposed by Hermann von Helmholtz almost 20 years later (Allik, Konstabel 2005). One of the following rectors, Alfred Wilhelm Volkmann (1800–1878), who is known as the inventor of the tachistoscope, was regarded by Edwin Boring – the dean of all historians – as one of the founding fathers of the experimental psychology (Allik 2007). Georg Wilhelm Struwe (1793–1864) played an essential part in solving the problem of personal equations, the very concept from which the experimental psychology was started in the Wilhelm Wundt's laboratory at Leipzig. Physicist Arthur Joachim von Oettingen (1836–1920) developed a theory of music harmony, which stimulated one of his students Wilhelm Friedrich Ostwald (1853–1932) to study colour harmony after he had received the Nobel Prize in chemistry. Emil Kraepelin (1856–1926), the founder of modern psychiatry, is by far the most important psychologist who has worked in Estonia. His inaugural lecture at the University of Tartu was reprinted recently once again to demonstrate his extraordinary understanding which sounds very modern (Kraepelin 2005). His successor Wladimir von Tchisch (1855–1922), another student of Wilhelm Wundt, continued Kraepelin's work in experimental psychology. The lives of Wolfgang Köhler (1887–1967), the master mind of the Gestalt psychology, who was born in Reval (Tallinn), and Oswald Külpe (1862–1915), the founder of the Würzburg school, who graduated from the University of Dorpat, extend the link between the history of psychology and Estonia (Allik 2007). Karl Gustav Girgensohn (1875–1925), the founder of the Dorpat School of the psychology of religion, stretched the use of experimental methods to the study of religious experience (Allik 2007).

In 1919, after Estonia had obtained independence, Tartu University became a national university with the language of instruction being primarily Estonian. The first professor of psychology was Konstantin Ramul (1879–1975) who is first and foremost known as a historian of an early period of experimental psychology (Ramul 1960, 1963). In 1921, one of Ramul's assistants Juhan Tork (1889-1980) spent one semester in Leipzig where he attended, among the others, lectures of Wilhelm Wundt. Returning from Germany, he brought with him equipment that was needed to launch a laboratory of experimental psychology. Juhan Tork is an author of the most innovative psychological study that was carried out between the two wars. In 1940, he published doctoral dissertation (Tork 1940) on the intelligence of Estonian children based on the study of about 6,000 school children, containing many observations what anticipated the discoveries in psychological thinking.

After Soviet occupation in 1940 Estonian science was absorbed into one of the most isolated and inefficient science systems in the world which also suffered from the mania grandiosa – they cited their own publications disproportionally more compared to citations of their works by all other scholars. Psychology served mainly teaching purposes and its scientific ambitions were severely censored and restricted. Only a few years before regaining independence in 1991, Estonian science including psychology started to move back to its habitual place in the world science community.

PSYCHOLOGY DURING THE LAST TWENTY YEARS

The PsycINFO (American Psychological Association) is a database that provides systematic coverage of the psychological literature from the 1800s to the present. It covers comprehensively more than 1,500 titles of scholarly, peer-reviewed, and regularly published journals which form approximately 80% of the database. The remaining 20% are books, book chapters and other mainly non-periodic publications. In total it has over 3 million weekly updated records from which more than million journal articles, books, and book chapters contain 48 million references. Based on this database we can observe the development of psychology in Estonia and two other Baltic states during the first two decades of independence.

Fig. 1 demonstrates the growth curves of papers published annually by Estonian, Latvian and Lithuanian psychologists in journals and books indexed in the PsycINFO database from 1990 to 2010. While in 1990 there were only a few publications, in 2010 Estonian psychologists published approximately 100, Lithuanian – 50, and Latvian – 10 articles. More than a half of the Estonian papers are produced by researchers working at the Institute of Psychology at the University of Tartu. Since in 1990 the PsycINFO indexed 58,532 papers and 151,595 papers in 2010, even the relative share of Estonian papers increased approximately ten times from 0.007% to about 0.07%.

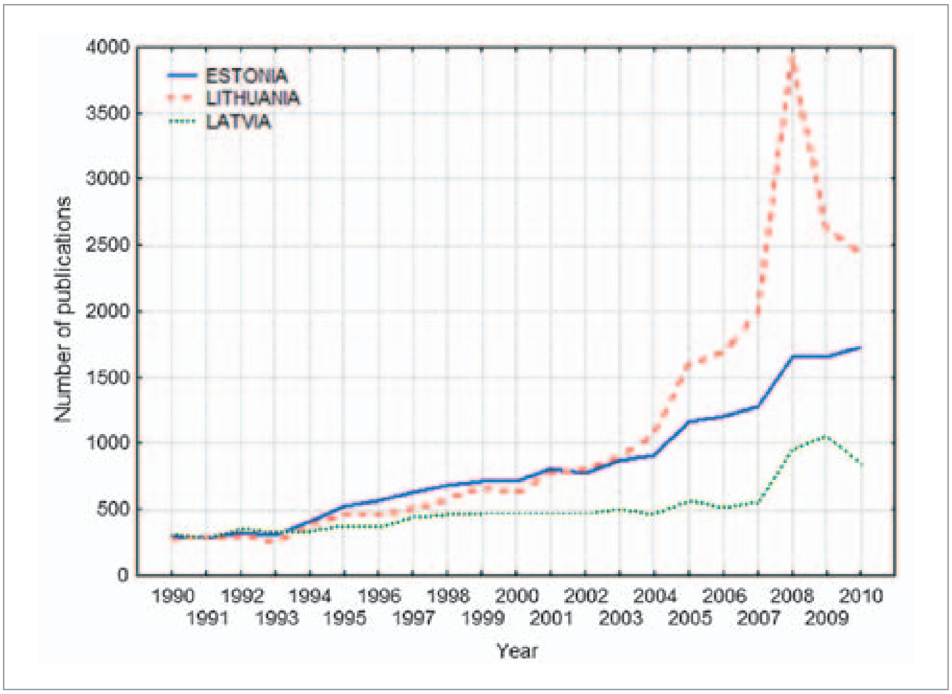


Figure 1.
The number of publications by Estonian, Lithuanian, and Latvian authors in the PsycINFO database from 1990 to 2010.

However, it is almost meaningless to compare productivity of Estonian and, say, American psychologists who produced about 600 and 600,000 papers respectively during the last 11 years from 2000 to 2010. One way how to compare psychology-friendliness of different countries is to compute how many papers their psychologists have produced per million inhabitants living in their country. Table 1 demonstrates the top 30 of the most psychology-friendly countries in the world during the last eleven year period 2000–2010. In addition the results for four countries – Latvia, Lithuania, Russian Federation, and People's Republic of China providing reference points for comparisons are also shown.

Although psychologists of the United States authored or co-authored 50.6% of all 1.25 million psychology papers indexed during that period, the most psychology-friendly countries are Canada, Australia, and Israel. Estonia occupies the 21st position being able to produce about 50 papers per one million inhabitants. From former Communist bloc countries only Slovenia was more productive. It is certainly remarkable that in relative terms Estonia is even ahead of such countries as Italy and France. Nevertheless, the relative productivity of our closest neighbour and partner Finland was about 3 and half times larger. About two decades ago, however, the gap of productivity between Estonian and Finnish psychologists was close to seven times. Two other Baltic countries, Latvia and Lithuania, are in the lower part of the list closer to the Russian Federation and China.

The number of publications in international peer-reviewed journals is not an automatic guarantee of their quality. Unfortunately, in the Essential Science Indicators (ESI; Thomson Reuters) psychology is the minor partner of the joint category psychiatry/psychology. Although a considerable growth in citations per paper published by Estonian psychiatrists and psychologists was observed during the last decade 2000–2010, their impact is still –30% below the world average (see "Estonian science estimated through bibliometric indicators" in this volume). There are all reasons to think, however that the performance of Estonian psychology is superior to that of psychiatry. There were also 3 individual psychology papers co-authored by Estonian psychologists which citation rate exceeded the top 1% threshold for articles in the category psychiatry/psychology and they all belonged to psychologists (Schmitt et al. 2007, 2008; Swami et al. 2010). Beside these 3 several other papers authored by psychologists also reached the top 1%, nevertheless in other fields of science such as clinical medicine and neuroscience (Näätänen et al. 2001, 2005, 2007).

Table 1.

The top 30 of the most psychology-friendly countries in the PsycINFO database during the period from 2000 to 2010.

	Country	Publications	Population	Publications per million
1	Canada	74 770	34 473 000	216.9
2	Australia	48 544	22 626 868	214.5
3	Israel	16 590	7 746 400	214.2
4	Netherlands	34 911	16 673 100	209.4
5	Iceland	657	318 452	206.3
6	New Zealand	9 072	4 407 300	205.8
7	United States of America	632 085	311 481 000	202.9
8	Sweden	17 624	9 428 054	186.9
9	Norway	9 022	4 949 300	182.3
10	United Kingdom	107 929	62 008 048	174.1
11	Finland	9 360	5 384 490	173.8
12	Switzerland	12 994	7 866 500	165.2
13	Ireland	7 000	4 470 700	156.6
14	Denmark	7 325	5 560 628	131.7
15	Hong Kong	8 545	7 061 200	121.0
16	Belgium	10 890	10 827 519	100.6
17	Germany	52 427	81 802 000	64.1
18	Austria	5 376	8 404 252	64.0
19	Slovenia	1 100	2 051 900	53.6
20	Spain	23 314	46 148 605	50.5
21	Estonia	667	1 340 122	49.8
22	Italy	26 753	60 626 442	44.1
23	France	28 290	65 821 885	43.0
24	Greece	4 561	11 306 183	40.3
25	Croatia	1 632	4 425 747	36.9
26	Luxembourg	183	502 100	36.4
27	Portugal	3 482	10 636 888	32.7
28	Hungary	3 043	10 014 324	30.4
29	Japan	26 281	127 930 000	20.5
30	Czech Republic	1 854	10 515 818	17.6
…				
…	Lithuania	366	3 225 300	11.3
…	Latvia	83	2 221 100	3.7
…	Russian Federation	3 568	142 905 200	2.5
…	People's Republic of China	18 200	1 339 724 852	1.4

Note: Population data were retrieved from http://en.wikipedia.org/wiki/World_population (June 5, 2011).

SOME OBSERVATIONS ABOUT GENERAL TRENDS

It is not a coincidence that many, possibly a majority of currently leading Estonian psychologists started their career from cognitive psychology. Of course, it is possible that our countryman Endel Tulving of University of Toronto, one of the leading authorities of the modern memory research, played an incentive role but the main motivation for many of us was to find a refuge from an ideological pressure. Although some eager political activists sincerely believed that the Soviet system was able to breed a distinctive *Homo Soveticus*, most of the guardians of political correctness still tolerated a presumption that senses and cognitive apparatus of a human being are basically identical irrespective of political or economic system. Most logically the first publications authored by Estonian psychologists which appeared in internationally monitored journals concerned eye movement (Allik et al. 1981) or visual masking (Bachmann, Allik 1976). Two internationally available monographs published on the initial wave of these studies were relatively successful. Peeter Tulviste published his "The cultural-historical development of verbal thinking" in 1991 (Tulviste 1991). Since its appearance this book has been cited in 47 articles published in the journals indexed by the Web of Science (WoS) and even in 166 sources accessed by the Harzing's Publish or Perish (PoP; August 5, 2011). Few years later Talis Bachmann published his "Psychophysiology of masking: The fine structure of conscious experience" (Bachmann 1994) which has been cited 84 times by sources indexed by WoS and 104 times according to PoP.

Like any other Post-Communist countries, the most dramatic changes happened in the clinical and counselling psychology. These changes obtained more in the teaching curricula and professional practices than in academic publications. Still, an increasing number of reports about successful adoption of various measurement instruments is a solid yet superficial sign of changes in Estonian psychology as a profession. Not only academics took a leading role in the translation of the most popular psychological measurement instruments such as NEO-PI or PANAS (Allik, Realo 1997; Pulver et al. 1995) but many of these adaptations were motivated by practical needs to secure required standards for diagnostic tools (Aluoja et al. 1999).

There are several trends that can be observed in the development of Estonian psychology many of them parallel to global developments that were happening in the world psychology. One perhaps inevitable shift was from 'hard' to 'soft' psychology. Lee Cronbach, who is mainly known for the coefficient of reliability named in his honour, pointed to the gap between two types of psychology – correlational and experimental – in his inaugural talks as the elected president of the American Psychological Association. As mentioned above, most senior Estonian psychologists started with the 'hard' experimental psychology practicing either laboratory or field experiments. With the collapse of the Soviet ideology the doors were opened to softer psychology, associated mainly with personality and social psychology, which principal tool is still correlation. Because entering the market of international ideas is a tricky business, especially for non-native English speakers from the former Russian colony, many researchers were happy to join a bandwagon of studies initiated by Western colleagues. Although it was not only for this reason that Estonia was present in virtually all major cross-cultural surveys studying either sexual behaviour (Buss et al. 1990), self-esteem (Schmitt, Allik 2005) or even gelatophobia – a fear of being laughed at (Proyer et al. 2009). However, it took some time until original concepts developed by Estonian psychologists (Põder 2006; Toomela 2003), their meta-analyses (Strenze 2007), or cross-cultural projects initiated by Estonian teams started to penetrate the most influential international journals (Mõttus et al. 2011).

Another remarkable movement goaded Estonian psychology from almost completely 'dry' psychology towards 'wet' one. It could be asserted that the modern research paradigm of Estonian psychology was born, partly at least, as a result of so called cognitive revolution. Unlike many other places cognitive psychology practiced in Estonia was very dry with more theoretical rather than practical links to its wet foundations, neurons and transmitters. The obsolete Soviet economy was anything but supportive to study of brain and central nervous system. With about two decades of independence, studies relating psychologically interesting phenomena to brain functions have become regular in Estonian psychology. For example, location of phonemes in brain was discovered (Näätänen et al. 1997); noradrenergic innervation from the locus coeruleus was linked to depression (Harro, Oreland 2001), short alleles of NOS1 ex1f-VNTR went along with impulsive behaviour (Reif et al. 2011), or non-REM sleep led to relative more positive slow brain potentials, compared to wakefulness (Stamm et al. 2011). One obvious advantage of Estonian psychology has been openness to those who have different basic trainings. From current prominent psychologists Eve Kikas came from physics and both Jaanus Harro and Aaro Toomela have medical training.

Another remarkable shift that characterises Estonian psychology was moving from 'cool' to 'hot' psychology. Like cognitive psychologists who were inspired by Endel Tulving, another *émigré* from the North America Jaak Panksepp played a pivotal role in shifting not only Estonian psychology towards emotions and affective phenomena. His "Affective Neuroscience" (1998) which become a classic (Panksepp 1998) which has been cited 1,100 times according to WoS and 1,795 times according to PoP (August 5, 2011) provided a role-model and inspiration. It started with a modest adaptation of the most popular measuring instruments of affect, PANAS, (Allik, Realo 1997) and evolved into a broad spectrum on studies involving such questions as how emotions affect judgements of life-satisfaction (Kuppens et al. 2008), how recognition of facial expressions deteriorates with age (Mill et al. 2009), and which brain regions are most likely linked to depression (Harro et al. 2011).

Although Estonian psychology has always been characterised by a relatively high-level of culture awareness, the link between psychology and culture has clearly tightened during the last few years. Perhaps living on the cross-road of different cultural influences has sharpened senses of Estonian psychologists towards cultural issues. It is certainly not a coincidence that the Founding Editor of a prestigious international journal with an emblematic title "Culture and Psychology", Jaan Valsiner of Clark University, is our good colleague and graduate of the University of Tartu. One traditionally prominent line of studies concerns how different cultural institutions and practices, such as school education, transform human mind (Hannust, Kikas 2010; Toomela 1996; Tulviste 1991). According to a recent report published in the "Science" magazine, psychology seems to be committing a grave error studying mostly WEIRD subjects – people from Western, Educated, Industrialised, Rich, and Democratic cultures – who are, according to some psychologists some of the most psychologically unusual people on the planet (Jones 2010). Indeed, approximately 96% of subjects used in psychological studies come either from the United States, other English-speaking and/or European countries. Only three percent are from Asia and less than 1% from Latin America, Africa or the Middle East. It seems that Estonian psychology has managed to avoid this error being involved in a large number of cross-cultural comparisons of such different topics as talking during the meal time (De Geer et al. 2002), social axioms (Bond et al. 2004), desire for sexual variety (Schmitt, 118 Members of the International Sexuality Description Project 2003), and cultural tightness or looseness (Gelfand et al. 2011). Thanks to these and numerous other studies we are relatively well informed about location of Estonia on different thematic maps which show how people feel, think and behave in different situations.

CONCLUDING REMARKS

There may obtain an impression, perhaps not totally wrong, that the current generation of Estonian psychologists continue the research tradition which foundation was established by eminent scholars of the 19th century who worked at the University of Tartu. It is certainly not usual that such a small country can maintain decent research activities and achieve prominent results in so many different fields of psychology. Future of Estonian psychology obviously depends not only on available intellectual resources which are in short supply but also on the supporting environment. Unfortunately, there is not much room for an optimism that this environment could change more favourably in near future. In such a small country as Estonia the number of active researchers is more or less directly dependent on the number of psychology students. Since psychology has been one the most popular subjects the state policy has been to shift the financial burden on the shoulders of those students who can afford paying their study fees. It is indeed ridiculous that the state of Estonia has been ready to finance only 15 bachelor and 12 master students of psychology at the University of Tartu where psychology courses have been taught without major interruptions since 1802. This is obviously not enough to keep busy even one active researcher who needs to share his time between research and teaching. These numbers are even less than that for theologians, particularly for a country known as one of the most secular in the world, to say nothing about lawyers, economists, and librarians. Another stumbling block is what I perceive as technocratic mentality of the local science policy makers. When it concerns establishment of scientific centres of excellence or any other large scale investments, priority is always given to physics or chemistry and rarely if ever to behavioural sciences or even neurosciences, though their scientific achievements are less impressive. Nevertheless, there is a chance that Estonian psychology can maintain its pace of development and maintain its competitiveness on the world market of new ideas.

REFERENCES

Allik, J. 2007. History of experimental psychology from an Estonian perspective. Psychological Research-Psychologische Forschung, 71, 6, 618-625.

Allik, J., Konstabel, K. 2005. G. F. Parrot and the theory of unconscious inferences. Journal of the History of the Behavioral Sciences, 41, 4, 317-330.

Allik, J., Rauk, M., Luuk, A. 1981. Control and sense of eye-movement behind closed eyelids. Perception, 10, 1, 39-51.

Allik, J., Realo, A. 1997. Emotional experience and its relation to the Five-Factor Model in Estonian. Journal of Personality, 65, 3, 625-647.

Aluoja, A., et al. 1999. Development and psychometric properties of the Emotional State Questionnaire, a self-report questionnaire for depression and anxiety. Nordic Journal of Psychiatry, 53, 6, 443-449.

Bachmann, T. 1994. Psychophysiology of Masking: The Fine Structure of Conscious Experience. Nova Science Publishers, New York.

Bachmann, T., Allik, J. 1976. Integration and interruption in the masking of form by form. Perception, 5, 79-97.

Bond, M. H., et al. 2004. Culture-level dimensions of social axioms and their correlates across 41 cultures. Journal of Cross-Cultural Psychology, 35, 5, 548-570.

Buss, D. M., et al. 1990. International preferences in selecting mates: a study of 37 cultures. Journal of Cross-Cultural Psychology, 21, 1, 5-47.

De Geer, B., et al. 2002. Socialization in communication: pragmatic socialization during dinnertime in Estonian, Finnish and Swedish families. Journal of Pragmatics, 34, 12, 1757-1786.

Gelfand, M. J., et al. 2011. Differences between tight and loose cultures: A 33-nation study. Science, 332, 1100-1104.

Hannust, T., Kikas, E. 2010. Young children's acquisition of knowledge about the Earth: A longitudinal study. Journal of Experimental Child Psychology, 107, 2, 164-180.

Harro, J., et al. 2011. Mapping patterns of depression-related brain regions with cytochrome oxidase histochemistry: Relevance of animal affective systems to human disorders, with a focus on resilience to adverse events. Neuroscience and Biobehavioral Reviews, 35, DOI: 10.1016/j.neubiorev.2011.02.1016.

Harro, J., Oreland, L. 2001. Depression as a spreading adjustment disorder of monoaminergic neurons: a case for primary implication of the locus coeruleus. Brain Research Reviews, 38, 1–2, 79-128.

Helmholtz, H. v. 1896. Über die academische Freiheit der Deutschen Univärsität. H. v. Helmholtz, H. v. (d). Vorträge und Reden. 2. 4 ed. Vieweg, Braunschweig, 191-212.

Jones, D. 2010. Psychology a WEIRD: View of human nature skews psychologists' studies. Science, 328, 5986, 1627-1627.

Kraepelin, E. 2005. The directions of psychiatric research. History of Psychiatry, 16, 3, 350-364.

Kuppens, P., Realo, A., Diener, E. 2008. The role of positive and negative emotions in life satisfaction judgment across nations. Journal of Personality and Social Psychology, 95, 1, 66-75.

Mill, A., et al. 2009. Age-related differences in emotion recognition ability: A cross-sectional study. Emotion, 9, 5, 619-630.

Mõttus, R., et al. 2011. Comparability of self-reported conscientiousness across 21 countries. European Journal of Personality, DOI: 10.1002/per.840.

Näätänen, R., et al. 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature, 385, 6615, 432-434.

Näätänen, R., et al. 2001. 'Primitive intelligence' in the auditory cortex. Trends in Neurosciences, 24, 5, 283-288.

Näätänen, R., et al. 2007. The mismatch negativity (MMN) in basic research of central auditory processing: A review. Clinical Neurophysiology, 118, 12, 2544-2590.

Näätänen, R., Jacobsen, T., Winkler, I. 2005. Memory-based or afferent processes in mismatch negativity (MMN): A review of the evidence. Psychophysiology, 42, 1, 25-32.

Panksepp, J. 1998. Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford University Press, New York.

Proyer, R. T., et al. 2009. Breaking ground in cross-cultural research on the fear of being laughed at (gelotophobia): A multi-national study involving 73 countries. Humor-International Journal of Humor Research, 22, 1-2, 253-279.

Pulver, A., et al. 1995. A Big Five personality-inventory in two non-Indo-European languages. European Journal of Personality, 9, 2, 109-124.

Põder, E. 2006. Crowding, feature integration, and two kinds of "attention". Journal of Vision, 6, 2, 163-169.

Ramul, K. 1960. The problem of measurement in the psychology of the eighteenth century. American Psychologist, 15, 256-265.

Ramul, K. 1963. Some early measurements and ratings in psychology. American Psychologist, 18, 653-659.

Reif, A., et al. 2011. A functional NOS1 promoter polymorphism interacts with adverse environment on functional and dysfunctional impulsivity. Psychopharmacology, 214, 1, 239-248.

Schmitt, D. P., et al. 2007. The geographic distribution of big five personality traits – Patterns and profiles of human self-description across 56 nations. Journal of Cross-Cultural Psychology, 38, 2, 173-212.

Schmitt, D. P., et al. 2008. Why can't a man be more like a woman? Sex differences in big five personality traits across 55 cultures. Journal of Personality and Social Psychology, 94, 1, 168-182.

Schmitt, D. P., 118 Members of the International Sexuality Description Project. 2003. Universal sex differences in the desire for sexual variety: Tests from

52 nations, 6 continents, and 13 islands. Journal of Personality and Social Psychology, 85, 1, 85-104.

Schmitt, D. P., Allik, J. 2005. Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. Journal of Personality and Social Psychology, 89, 623-642.

Stamm, M., Aru, J., Bachmann, T. 2011. Right-frontal slow negative potentials evoked by occipital TMS are reduced in NREM sleep. Neuroscience Letters, 493, 3, 116-121.

Strenze, T. 2007. Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. Intelligence, 35, 5, 401-426.

Swami, V., et al. 2010. The attractive female body weight and female body dissatisfaction in 26 countries across 10 world regions: Results of the International Body Project I. Personality and Social Psychology Bulletin, 36, 3, 309-325.

Toomela, A. 1996. How culture transforms mind: A process of internalization. Culture & Psychology, 2, 3, 285-305.

Toomela, A. 2003. Relationships between personality structure, structure of word meaning, and cognitive ability: a study of cultural mechanisms of personality. Journal of Personality and Social Psychology, 85, 723-735.

Tork, J. 1940. Eesti laste intelligents. [The intelligence of Estonian children]. Koolivara, Tartu.

Tulviste, P. 1991. The Cultural-Historical Development of Verbal Thinking. Nova Science Publishers, Commack, N.Y.

RESEARCH AND ECONOMY

U. Varblane, K. Ukrainski

INTRODUCTION

A substantial body of empirical work confirms the importance of innovation for long-run economic growth (Kuznets 1973; Hasan, Tucci 2010). The country benefits from a well developed science base and properly functioning linkages between science and technology system and business sector. It all suggests the need for the well functioning national innovation system (Freeman 2002). Therefore the current chapter is focusing on the analyses of the problems surrounding the building of a knowledge based national innovation system in Estonia as the catching-up economy. Specific attention is given to the problems of developing linkages between the academia and industry. In the first section of the chapter short overview is presented about the theoretical framework of industry academia cooperation as one of the cornerstones of well functioning national innovation system. Next section is devoted to the description and analysis of the designing of Estonian national innovation system. The following section is presenting the main trends in the funding of research and development (hereinafter R&D) activities in Estonia with major emphasis on the business sector. Next section provides data about the dynamics of researchers employed in the Estonian business sector. The following empirical section is dealing with the analysis of the university and industry linkages focusing on the joint publication, co-supervision and patenting. The last section is devoted to the role of universities and other research organisations as the source of knowledge for business sector innovations. The chapter concludes with discussion and recommendations.

THEORETICAL FRAMEWORK OF THE INDUSTRY AND ACADEMIA COOPERATION

It is widely accepted that innovation is among the major roads for firms and nations to remain competitive while following the path to sustainable development. The role of knowledge in generating competitive advantage has steadily increased and the ability to generate new knowledge requires functioning of the knowledge based system of innovation. Within the last couple of decades different models were proposed for the study of knowledge production process and innovation systems (Mode 2; national innovation systems, Triple Helix). Behind those models is the understanding about the second revolution in academic life and the changing role of universities (and more broadly research) in the national system of innovation. Authors of Mode 2 argued that post-modern development had led to the so-called de-differentiation of the relationship between science, technology and society (Gibbons et al. 1994; Nowotny et al. 2001). According to their arguments, this emergent mode is transdisciplinary, organisationally non-hierarchical, socially accountable, and reflexive. Knowledge is increasingly being produced in 'the context of application', that is, with societal needs having direct impact on the knowledge production from the early stages of investigative projects. This model of knowledge production supposes already nonlinear interactive type of innovation model.

Another approach was proposed by H.Etzkowitz and L.Leydesdroff as the Triple Helix model in mid 1990s (Etzkowitz, Leydesdorff 1998). In the Triple Helix model university, industry and government are performing in addition to their traditional functions also the roles of others. "Thus universities take on entrepreneurial tasks like marketing knowledge and creating companies, while firms develop academic dimension, sharing knowledge among each other and training employees at ever higher skill levels." (Leydesdorff, Etzkowitz 1998: 198). Based on their research results, Etzkowitz and Leydesdorff have suggested that there is not a single model of the Triple-Helix relations but in fact three different configurations (Etzkowitz, Leydesdorff 2000; Leydesdorff, Etzkowitz 1998). The first version – Triple Helix I or statist, was found in the former Soviet Union and East Europe – the nation state encompasses the university and industry and directs their mutual relationships. The second version of the model – Triple Helix II or 'laissez faire' model with industry, academia, government acting separately and interacting only modestly across strong borders. By this model industry is the driving force. The third version of the model – Triple Helix III – denotes a knowledge infrastructure made up of overlapping institutions that take the roles of each other and produce hybrid organisations.

Both above mentioned models are incorporated into national innovation system approach proposed by evolutionary economists, which centres around the idea of the need for systemic approach integrating institutions to create, store, and transfer the knowledge, skills and artefacts (OECD 1999; Lundvall 2007).

THE DEVELOPMENT OF THE ESTONIAN NATIONAL INNOVATION SYSTEM

After the regaining of its independence in 1992 Estonia faced typical problems of ex-Soviet economy in the process of building market based economy – liberalisation of markets, privatisation of the state owned firms and stabilisation of the financial market. Therefore during the first half of 1990's the problems of building national innovation system were not on agenda for Estonian policy makers. They moved ahead without having defined innovation policy and one could define Estonian innovation policy as 'no-policy policy' (Karo et al. 2010). But in the second half of 1990's awareness about the need for the systemic approach toward innovation spilled slowly from the neighbouring Finland over to Estonia. Finland was used as the role model for the systemic approach toward innovation as it has been among the first countries in the world declaring that they build their society using national system of innovation approach. In addition massive inflow of Finnish foreign direct investments into Estonian economy created strong firm level cooperation between Estonia and Finland. Historically close links between Finnish and Estonian universities also facilitated this process.

Late 1990s two important documents about the Estonian innovation policy were prepared: the Estonian State Innovation Programme (approved in 1998) and the National Development Plan for the years 2000–2002 (1999). Both documents were rather ambitious in the catching up economy context. Unfortunately none of those plans were actually implemented. The major reason was reluctance of allocation of resources to the fulfilling objectives defined in those plans (Tiits et al. 2003). The policy makers applied wishful thinking and forgot the differences in the starting positions of Estonia and Finland. In order to provide some insights how deep was the gap between those two countries, the following Table 1 is constructed with some key indicators about economic development, funding of R&D and higher education.

Nonetheless, Estonia defined its R&D strategy in a document, "Knowledge-based Estonia: Estonian R&D Strategy" (hereinafter KBE) for 2002-2006, passed by Estonian Parliament end of 2001. The document established even more ambitious goal – to increase the share of R&D expenditures to 1.5% of the GDP by 2006. It also defined three key areas of R&D activities in Estonia: user-friendly information technologies and development of the information society, biomedicine and material technologies. The proposed actual list of key technologies mirrored the list in Finland and many other highly developed economies and did not consider the real structures of Estonian industry with its dominance of less R&D intensive sectors, which are not likely to change overnight. (Koch et al. 2007).

Table 1.

Estonia and Finland compared by some indicators of economic development and funding of R&D and higher education in 1995.

Indicators	Finland	Estonia	Estonian level from the Finnish (in %)
GDP per capita in Purchasing Power Standards from EU-27 average (%)	108	36	33
Labour productivity in industry per employee (1000 EUR)	53.5	3.9	7.3
Gross R&D expenditures from GDP (%)	2.26	0.58	25.7
Business R&D expenditures from GDP (%)	1.43	0.08	7.7
R&D expenditures per full time researchers in higher education (EUR)	71656	9840	13.7
Patents (total EPO filings)	717	3	0.4
Total researchers (full time)	20857	2978*	14.3
Business enterprise researchers	10378	291*	2.8

Source: calculations based on Estonian Statistics, OECD 2011, EIUS2010.
Note: Earliest available data from 1998 used by Estonia *based on OECD 2011 data.

Positive still was that the new strategy document outlined the role of state to tackle with different market and other failures, acting as an investor, catalyst and regulator. But due to its political nature, the document was full of value statements and did not provide specific action plan. However without properly functioning public policy instruments, the document remains just a document (Kalvet 2001: 34-35). In short, by building this strategy Finland was mainly used as the model for copying. But in reality policy makers were able to copy only a tiny fraction of Finnish national innovation system – very linear approach, were expenditures on R&D were seen as the main aim and three key technologies. Unfortunately the whole system of learning aspect of Finnish NIS was not copied – attention was not given to the development of the system of absorption and diffusion of knowledge produced in the world. Estonia is so small a country that even in case of increasing its relative R&D expenditures to the level of world leading countries the domestic research potential is extremely limited. It could not alone solve the problems of upgrading the technological capability and productivity of Estonian main economic sectors. Consequently the success of Estonian economy depends heavily on the capability and willingness of firms to search, adapt, utilise knowledge produced outside Estonia (see also Varblane et al. 2007).

The high targets set in KBE 2002-2006 remained unfulfilled. (Fig. 1). Expenditures on R&D were planned to reach to the 1.5% of GDP in 2006.

The real level achieved was 1.13% – by far below the targets. Under the strong time constraint a new strategy document KBE for 2007-2013 (often referred as KBE II) was quickly prepared. It proposed even more ambitious aims: “the target of 3% of GDP expenditures on R&D, as agreed in the Lisbon strategy, is planned to be achieved by 2014” (p. 15 KBE II). It foresees also to double Estonian GDP by 2014, increase productivity by 80% of EU average level etc.

Now specifically after the serious economic crises becomes clear that targets set in the KBE II strategy remain unfilled too. Following Fig. 1 is presenting the actual and planned levels of gross expenditures of R&D in Estonia 1995-2015. First line about the planned level of gross R&D expenditures (GERD) is derived from the KBE 2002–2006 and second from the KBE 2007–2013.

The precondition of the fulfilment of the targets of Knowledge Based Estonia II by research expenditures is strong growth of the research activities of private sector. Without rapid structural change in Estonian economy and growth of knowledge intensive sectors the KBE II target is unrealistically high.

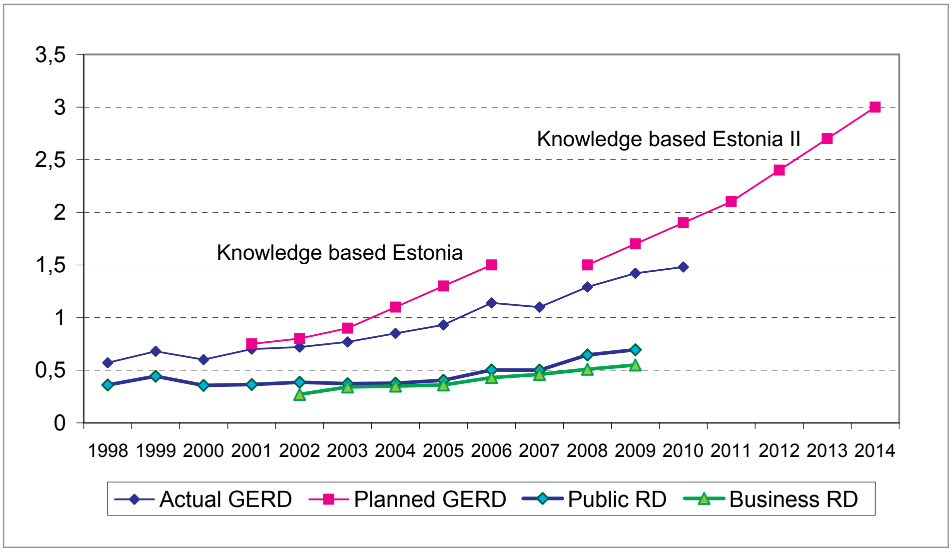


Figure 1.
The planned and actual expenditures on R&D in Estonia 1998-2014 (% from GDP).

Estonia started with the very low level of gross expenditures on R&D (GERD). It was only 0.57% in 1998, stagnated up to the mid 2000s, but then started to grow rather quickly reaching to the 1.48% in 2010 (Fig. 1 and 2). In the international comparison Estonian GERD is still far away from its Nordic neighbours, US and Asian new industrialised countries (Fig. 2), but converges toward the EU-27 average. Between 1995 and 2009 growth of Estonian GERD was among the fastest in the EU-27. Certainly it was enabled by the low starting position, but it also reflects the change in the priorities of the governmental policy making in favour of funding the R&D.

Previous data used in Fig. 1 provide relative share of R&D expenditures from GDP. Hence those indicators should be looked combined with the absolute figures, as the same percentage of expenditures from the GDP provides rather different real R&D expenditures per capita or per researcher terms. Therefore Table 2 was created, which provides GERD per capita at current prices and using purchasing power parity (eliminating effect of price differences in GDP calculations).

The data in Table 2 describe the internationally comparable expenditures on R&D in Estonia and some EU-27 and other OECD member countries. Table 2 reveals that in 2008 Estonia was able to invest per capita 283

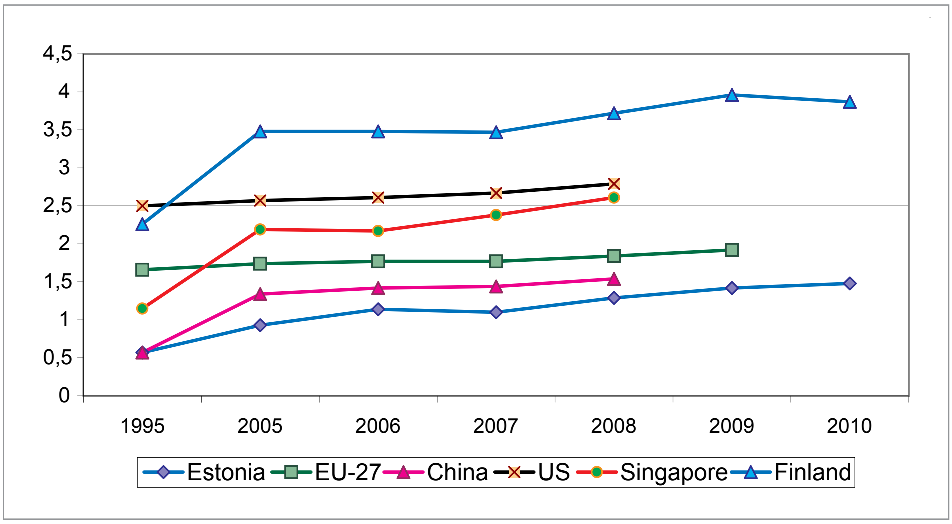


Figure 2.
Gross expenditures on R&D of Estonia in international comparison (% from GDP). Source: OECD 2011.

USD into R&D activities, which was 48% of EU-27 average. Top performers were Singapore with 1425, Finland 1406 and US 1306 USD. Particularly remarkable is the growth of Singapore from 294 USD in 1995 up to 1425 USD in 2008. Inside EU-27 remarkable growth is experienced also by Portugal, which increased R&D funding per capita basis 5.8 times during last 15 years and reached the level of 415 USD per capita.

The major reason behind the very low level of GERD lies in the small research intensity of the Estonian business sector. Governmental sector has steadily increased its contribution to the funding of GERD. The contribution of Estonian business sector to the R&D expenditures was 0.55 from GDP in 2009. This ratio constitutes one fifth of the Finnish or Korean levels, and two times below the EU-27 average (Fig. 3).

Particularly remarkable is the growth of the role of private sector in South Korea, which is already catching up with Japan in relative terms, but also Singapore and Finland. Estonian private sector contribution has increased with the same speed as in EU-27 average, therefore convergence has not happened within last decade. Unfortunately the private sector role in financing R&D in the whole EU-27 has stagnated and gap with Japan or US has widened. Therefore the Barcelona Summit target from March 2002 to increase not only the overall proportion of GDP devoted to R&D, but also to improve the private sector contribution to its financing remains unmet for the whole EU.

Table 2.

Gross expenditures on R&D per capita at current prices and PPPs in USD.

Countries	1995	2004	2005	2006	2007	2008	2009
Singapore	294	874	995	1081	1287	1425	n.a.
Finland	425	1031	1068	1151	1256	1406	1397
US	690	1023	1090	1163	1236	1306	n.a.
Korea	295	580	636	731	841	903	n.a.
EU-27	289	445	467	513	545	590	599
Slovenia	198	310	337	396	394	483	511
Czech Rep	122	241	288	346	377	379	389
Portugal	71	148	166	226	282	375	415
Estonia	n.a.	126	154	216	233	282	283
Poland	47	73	78	84	95	109	128

Source: OECD 2011.

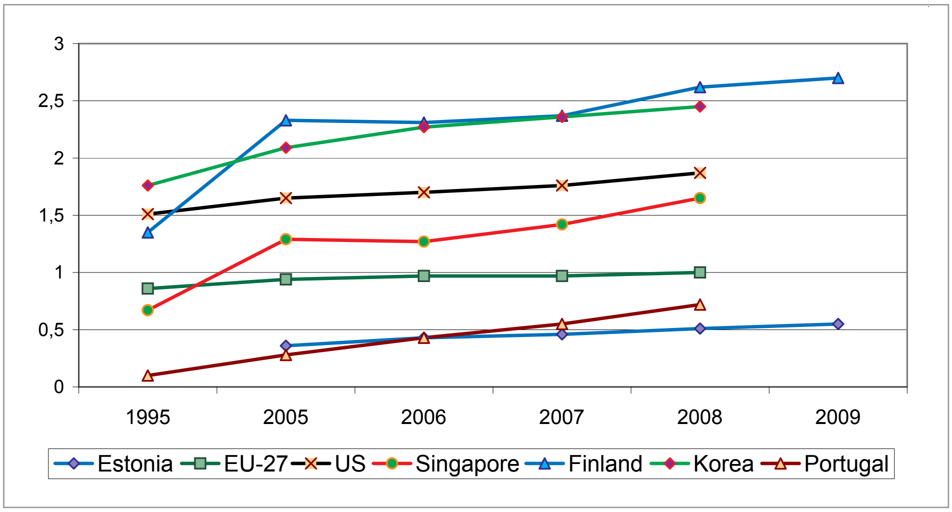


Figure 3.
Industry financed gross expenditures on R&D (% from GDP). Source: OECD 2011.

Table 3.

Business expenditures on R&D as percentage of value added of industry.

Countries	1995	2005	2006	2007	2008	2009
Finland	2.31	4.05	4.06	4.02	4.49	4.86
Korea	2.35	3.15	3.44	3.62	3.77	n.a.
US	2.77	2.88	2.97	3.09	3.25	n.a.
Singapore	0.87	1.71	1.69	1.9	2.25	n.a.
EU-27	1.58	1.71	1.75	1.76	1.81	1.87
China	0.28	1.09	1.19	1.23	1.32	n.a.
Portugal	0.18	0.52	0.8	1.04	1.31	1.35
Czech Rep	0.86	1.24	1.38	1.3	1.24	1.27
Estonia	n.a.	0.63	0.76	0.79	0.84	0.96
Poland	0.34	0.26	0.25	0.25	0.27	0.28

Source: OECD 2011.

The low ability (or willingness) of Estonian business sector to invest into R&D activities is revealed in Table 3, which shows the ratio of the business expenditures on R&D as percentage of value added of industry. From the whole value added created by Estonian industrial sector only 0.96% was spent on R&D activities in 2009. It is two time less than in EU-27 and more than 4 times below the Finnish level. Portugal is an interesting case of rapid growth of expenditures on R&D as percentage of value added of industry – from 0.18% in 1995 up to the 1.35% in 2009. Singapore has undergone an equally rapid growth, but from the higher starting level.

DYNAMICS OF RESEARCHERS EMPLOYED IN ESTONIAN BUSINESS SECTOR

The strength of Estonian business sector research activities depends on their own research capability, which is embodied in the employed researchers, but also on the intensity of co-operation with the Estonian and foreign research institutions. The following section is providing a short overview about the trends in employment of researchers in Estonia. In 1998 out of 4600 full-time researchers employed in Estonia only 440 worked in business sector (Fig. 4). The majority of researchers, 3077 worked in higher education sector. Between 1998 and 2009 the total number of researchers increased 18 percent, but in private sector the growth was 4.32 times.

Hence the relative share of private business employed researchers soared from 9.5% of all researchers in 1998 to 35.4% in 2009. By this ratio Estonia moved upward in the international comparison. However in 2008 the

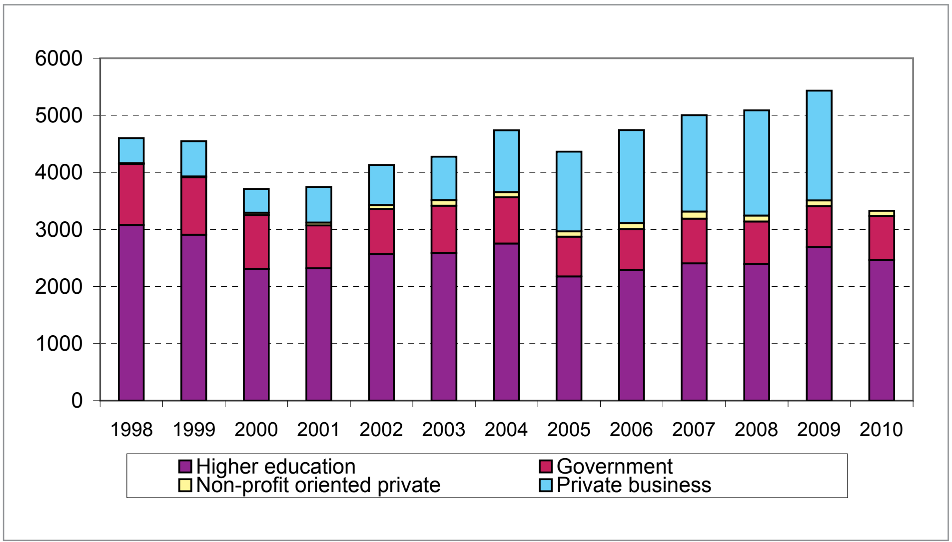


Figure 4.
Full-time researchers in Estonia employed by different sectors 1998–2010 (data about private sector researchers for 2010 not yet available). Source: Statistics Estonia 2011.

ratio of business researchers as percentage of national total in US was as high as 80, Japan 75, China 68.6, Finland 57.9 and EU-27 average 44.9. Another internationally acknowledged indicator of the intensity of research in business sector is the amount of full time R&D personnel per thousand employed in industrial sector. The following Fig. 5 indicates that in Estonia 4.1 researchers are employed per thousand employees. It is still far away from the leading countries Finland and Sweden employing 18.8 and 18.2 researchers per thousand employees. But it is also lower than EU-27 average 7.5. Estonia is on the same level as Hungary, Spain or Portugal. Therefore Estonian business sector has a long way to go in order to establish itself as an important knowledge producer with significant amount of full time researchers.

Clearly the limitation for the growth of number of researchers is the current structure of Estonian economy, which is biased toward sectors, where the research intensity is in general lower. But on the other hand the positive example is Finland, where forestry, wood, pulp and paper industry plays an important role. In general those sectors are treated as the low research intensity sectors, but in Finland they invest heavily into R&D activities.

In order to get a better understanding which sectors of Estonian economy are employing researchers Fig. 6 was constructed about the number of employees involved in R&D activities by economic sectors in Estonia between 1998 and 2009. The starting position in 1998 was characterised by 786 em-

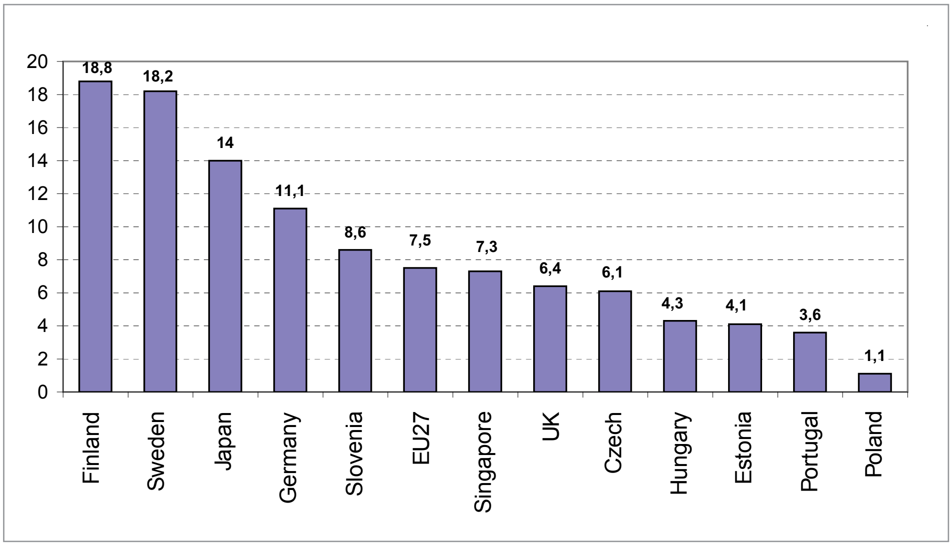


Figure 5.
Business R&D personnel (in full time equivalent) per thousand employees in industry. Source: OECD 2011.

ployees concerned with R&D activities in Estonian private sector. The whole industrial sector of Estonia employed only 202 researchers in 1998, which was approximately 20 times less than in the neighbouring Finland. It mirrored the low research intensity of Estonian economy in mid 1990s, when the major competitive advantage was low factors costs (mainly low labour costs) and extremely limited attention was accorded to the R&D activities. Positive changes started from the mid 2000 – between 2003 and 2009 the number of employees involved in R&D increased two times – from 1529 to 3122 (Fig. 6). The strongest input into the growth of R&D employees was given by the computer and software development sector, which increased their employment from the 237 employees in 2003 up to the 978 employees in 2008. R&D service providers were another sector which experienced rapid growth, particularly during the most recent years. From the 262 researchers employed by this sector in 2007 it reached, by the end of 2009, the 643 full time researchers' positions. It is partly associated with the emergence of many firms providing support in using EU structural funds in the fields of research, development and innovation. Manufacturing industry has not been very active in opening up new positions of researchers and during the economic crises they reduced full time researchers from 899 positions in 2007 down to 663 in 2009.

It is highly important that within the process of coming out from the economic crisis Estonian business sector will prioritise the creation of its own

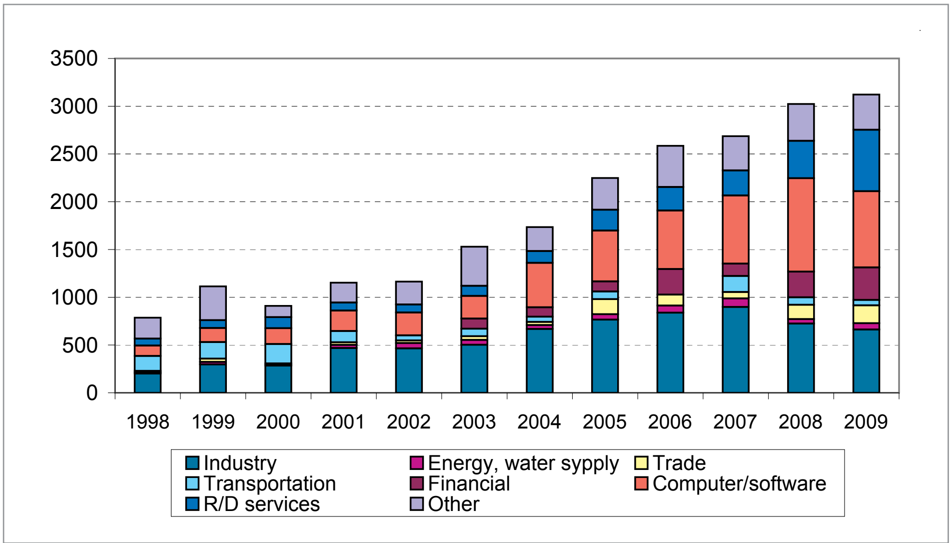


Figure 6.
Number of employees involved in R&D activities by economic sectors 1998–2009. Source: Statistics Estonia 2011.

research personnel. Special attention should be given to the employment of researchers with doctoral degree. In this respect, positive changes started since mid 2000s (Fig. 7). In 2004 the number of researchers employed in business sector exceeded 100, in 2008 already 200 and it is highly probable that in 2011 already 300. In the rapid growth process the frontrunner is R&D service providing sector, which employed as many as 161 doctoral degree holders by 2009, followed by ICT sector (46) and manufacturing industry (32).

ENTREPRENEURIAL ACTIVITIES OF UNIVERSITIES

In first paragraphs of this chapter, the changing role of universities in society has been described. By reason of the ideas set out above, the universities have extended the set of activities where they could specialise as research performers, but also as entrepreneurs. The universities have diversified the set of users of their knowledge and research to adapt their capabilities. Empirical assessment of university-industry partnership is not as widespread and standardised as in research performance; some indicators are presented below to describe Estonian situation.

Considering the general level of co-authored scientific publications, it is twice as low in Estonia (19) as against the average level of EU (36.2 publi-

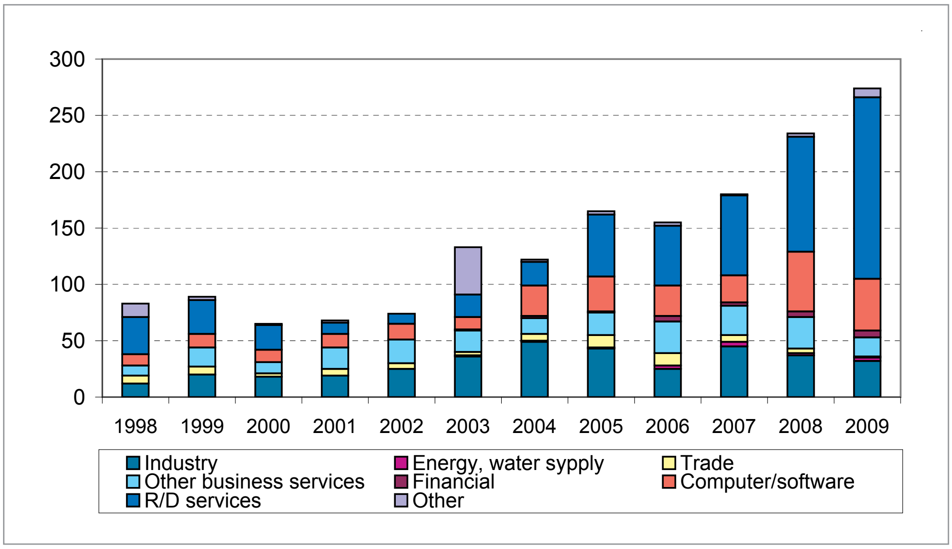


Figure 7.
Number of doctors working in Estonian economy by sectors 1998–2009. Source: Statistics Estonia 2011.

cations per million inhabitants). Fig. 8 indicates also, that the co-publication levels are very diverse across European countries, but generally growing more rapidly in new member states (it has tripled in Estonia within the observed six-year period). The reasons of low co-publication activities are related to different attitudes of industry and academia towards publishing the results of joint research, but also to misalignment or absence of the capabilities of industry partners (Varblane et al. 2008). The latter, but also lower capabilities of the university side (in terms of number of papers) are the reasons of the University of Tartu being in the lowest position in specific comparison with Finnish universities (see also Box 1).

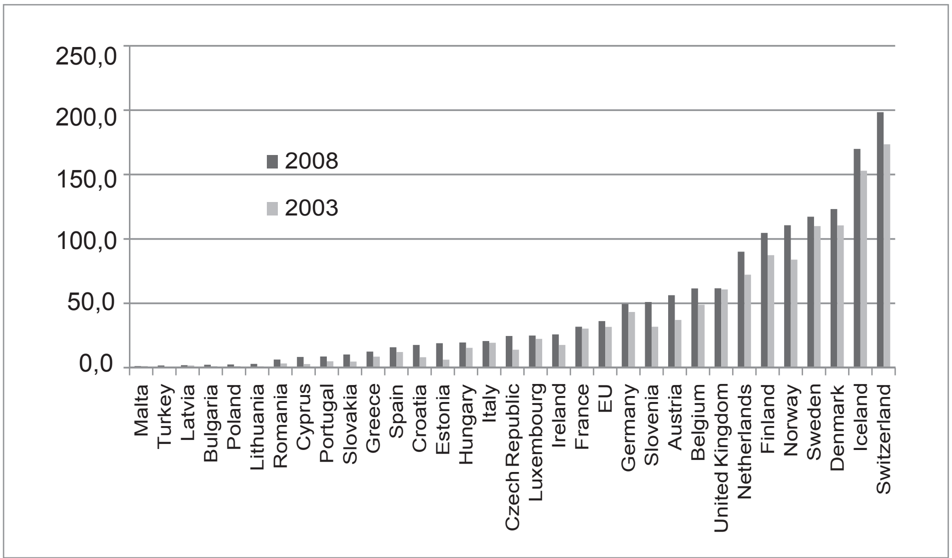


Figure 8.
Number of public-private co-authored publications (per million of inhabitants). Source: IUS 2010.

Considering the other form of aligning the activities in universities towards society is joint supervision of student research. Our analysis based on ETIS (Estonian Research Portal) reveals that in University of Tartu, the co-supervision with industry comprises about 3.4% and 4.9% of the master and doctoral theses respectively in recent years. But it is mainly limited with supervisors from spin-off firms or researchers working in both institutions. The share of co-supervised theses with employees of public institutions such as museums or ministries remains about 5–6%. The co-supervision with academics from other domestic, but also foreign universities has increased es-

Box 1. Co-publications in selected Estonian, Finnish and Swedish universities.

Relatively lower orientation of universities towards joint research is shown in university-industry co-publication rankings, which share in total publications is relatively low (Helsinki University of Technology 6.1%; University of Turku 3.9% and University of Helsinki 3.6% (Evaluation 2009)). This can be considered quite typical for selected universities if judged by the rankings of Leiden CWTS below.

Name	Country	UIC intensity (ranking range)	UIC volume (volume range)
Helsinki University of Technology	Finland	11-25	251-500
University of Turku	Finland	101-200	251-500
University of Helsinki	Finland	201-300	501-1000
University of Uppsala	Sweden	51-100	1001-2000
University of Tartu	Estonia	301-400	101-250

pecially in case of doctoral theses and has reached 7.3% from all supervised doctoral dissertations.

Patenting is besides research contracts another form for commercialising the research results of universities. Generally, the patent revenues are very low in Estonia (0.13% from GDP), but this is also true for most European countries (Fig. 9). The number of patent applications in Estonia is quite high (1.99 per billion GDP) when compared to other new member states, but also other small countries. Compared to the old member states, patenting is not as active, because of less favourable industrial structure, high patenting costs, but also complexity and legal uncertainty related to patenting (EMER 2011).

Despite of generally low patenting activity, universities have played a significant role in patenting. It is argued (based on the data of 2007) that about a half of Estonian patent applications comes from universities (Ojakivi 2008). The University of Tartu case presented in Box 2 reveals growing success in patenting. In Estonia, it seems that leading universities are not specialised in industry partnerships, but rather the opposite, are choosing to patent their own knowledge and technologies. One reason for that is also absence of suitable business partners with capabilities and competence in areas matching those of universities.

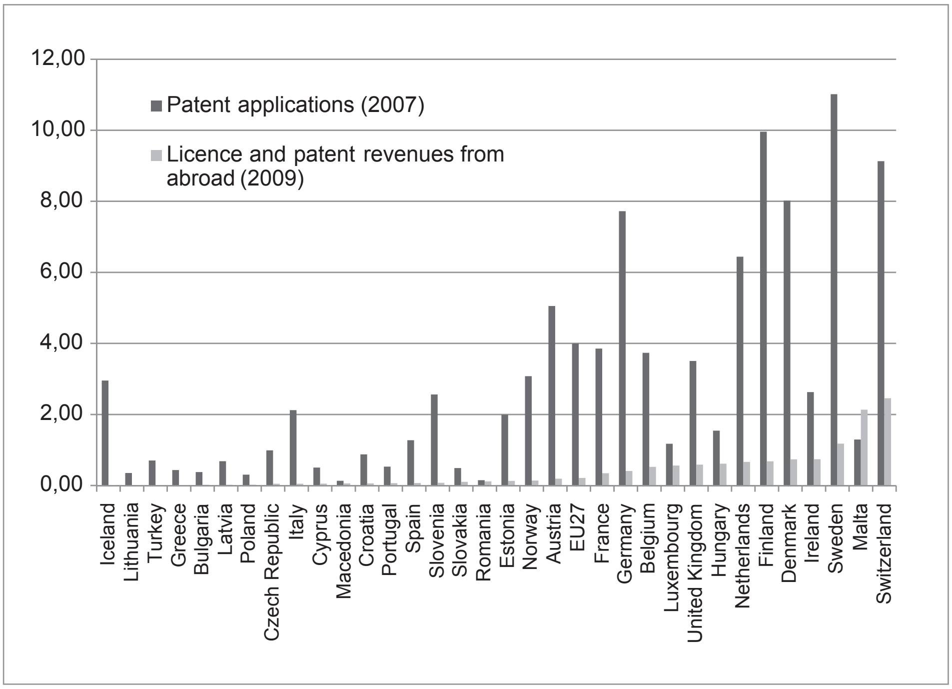


Figure 9.
Number of patent applications and patent revenues in European countries. Note: Patent applications are per billion GDP (in PPS€), revenues (% of GDP). Source: IUS 2010.

Estonia has targeted in its KBE's for 2000–2006 and 2007-2013 ICT, materials technologies and biotechnology as the key technologies achieving societal goals of knowledge-based economy. However, the relative importance of patenting in those fields has been quite low. Estonia is for example on the same level with Slovenia in ICT, but twice as low in biotech (IUS 2010 data). With more advanced countries, the differences are very large, for example Estonian patents in ICT comprise about 3% and in biotech about 8% of the respective numbers of Finland (IUS 2010 data).

By looking at the cases of individual universities, it seems that these do in fact have different specialisations by fields, whereby the diversification trend of patent portfolios by fields is evident. It can be seen from Box 3 that University of Helsinki is specialised in organic chemistry, University of Tartu in biochemistry and Helsinki University of Technology in biochemistry, electricity, but also in paper and pulp production (related to a significant sector of Finnish economy, but also global paper industry controlled by Finnish multinational corporations).

Box 2. Patenting in selected Estonian and Finnish universities.

It is evident, that all three selected universities are becoming more active in patenting, most importantly Helsinki University of Technology with 23 patents in the period of 1996–2008. University of Tartu is in the middle with 21 patents and University of Helsinki with 8 patents (Fig. 2.1).

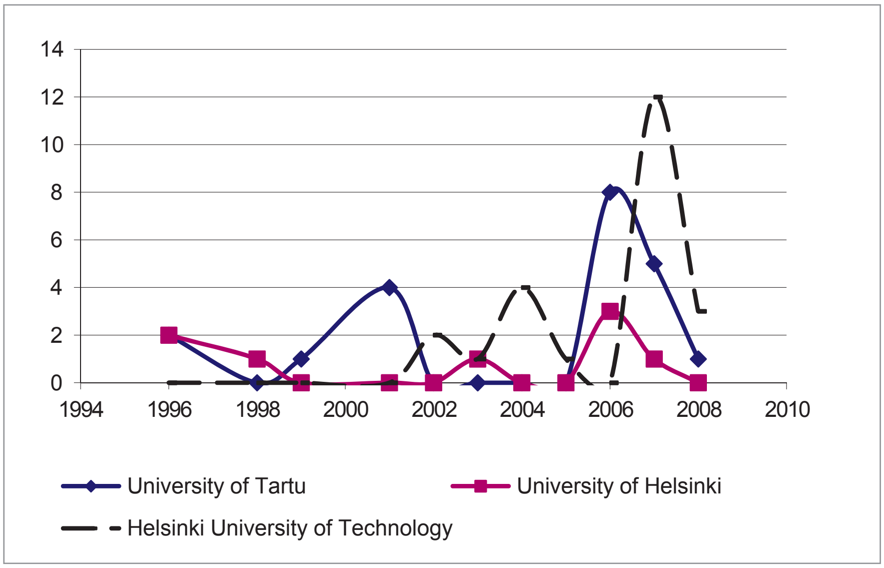


Figure 2.1.
Number of patents in University of Tartu, University of Helsinki and Helsinki University of Technology Source: authors' calculations based on EPO 2010.

It is interesting to note that all universities have different co-patenting strategies that concern joint patenting with other universities or companies. Helsinki University of Technology didn't fail joint patent application with any institutional co-applicant in the period of observation. University of Helsinki on the contrary has long co-patenting traditions with a couple of strategic academic partners (62.5% from all patents in 1996–2008), the role of industry partners is considerably smaller (12.5%). University of Tartu's main patenting partners are business firms, which amounted to 9.5% from all patents in 1996–2008, but emerged only in last two years of the observation.

Box 3. Patenting by fields in selected Estonian and Finnish Universities.

It is evident from Table 3.1 that the strategies are different among the universities; University of Helsinki has mainly concentrated on organic chemistry connected with medical or veterinary science. In recent period, when the number of patents has been large neither, one patent has been applied in the field of basic electric elements. The other two universities are having more diversified pattern of patenting, although having also the largest part in chemistry (University of Tartu 50% and Helsinki University of Technology 43.7% in latter period). Within chemistry, the largest part of patents in both universities (predominant for University of Tartu, 46.7% from all patents in 1996-2008) has been applied for in biochemistry, microbiology, enzymology and genetic engineering. The raising fields for University of Tartu are medical science, different kinds of tools. Single patents in recent years have been applied for in shipbuilding, machinery for liquids, basic electric elements and agriculture, forestry and fishing. In case of Helsinki University of Technology, the field of metallurgy has emerged, but the share in paper and cellulose production has diminished, and the patents in mechanical engineering and physics have not been applied for recently.

Table 3.1

Distribution of patent applications by fields, %.

Field by European classification	University of Tartu		University of Helsinki		Helsinki University of Technology	
	1996-2004	2005-2008	1996-2004	2005-2008	1996-2004	2005-2008
Human necessities	28.6	14.3				12.5
Performing operations, transportation		21.4				6.3
Chemistry, metallurgy	57.1	50.0	100.0	75.0	28.6	43.7
Textiles, paper					28.6	6.3
Mechanical engineering		7.2			14.3	
Physics	14.3				14.3	
Electricity		7.1		25.0	14.2	31.2
Total	100.0	100.0	100.0	100.0	100.0	100.0

Source: Authors' calculations using EPO (2010) data. Note: Here the patent applications have been distributed according to the first field in patent application. As often these include several fields, considering all fields involved might have different, but also elaborated results.

UNIVERSITIES AS SOURCES OF KNOWLEDGE FOR INNOVATION

Increasing competition, shorter product development cycles and rapid development in technologies have generally promoted the idea that universities, in addition to their basic research function, have an additional function to fulfil – to provide applicable knowledge in collaboration with firms that can lead to commercial innovations. The firms might locate their R&D activities abroad as the major source of knowledge as part of the globalisation process – increasing interaction with universities is also found in research-intensive industries where the close proximity and interface of different disciplines is relevant and finally, the respective efforts of governments enhance this kind of interaction (Faulkner et al. 1995).

Estonian firms have been rather active in renewing their products and processes compared to other European countries (Fig. 10), such firms comprised 56.4% of all firms. Considering the novel (new to the market) innovations, the share has been lower (19.1%), but still Estonian position in Europe looks fine. It is natural that most of the innovations are incremental. It is found that Estonian firms are more engaged in procurement and imple-

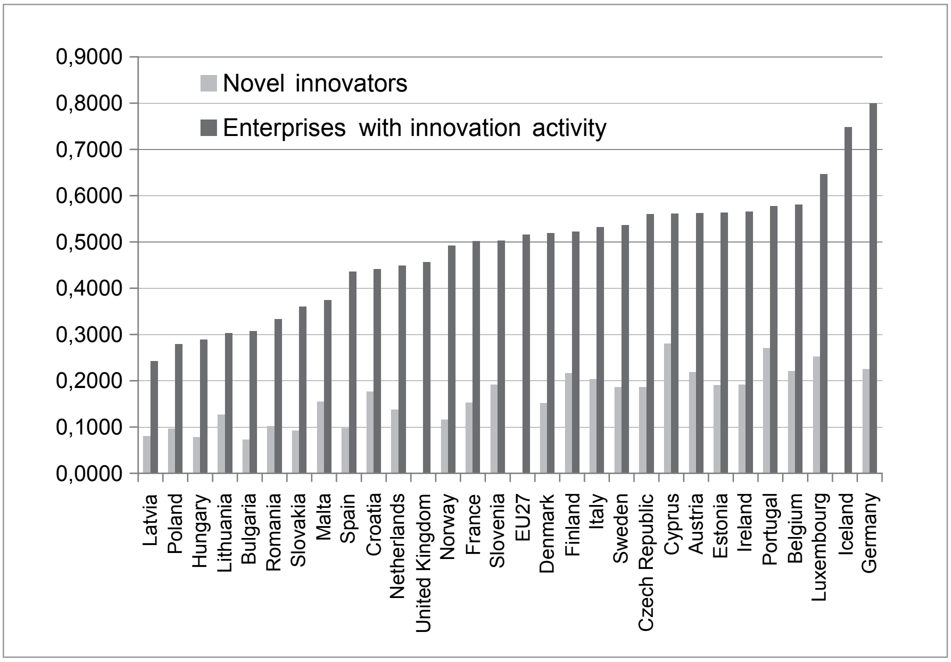


Figure 10.
Firms with innovation activities (as % from all firms). Source: EUROSTAT 2011.

mentation of new machinery and equipment, but relatively less interested in marketing and organisational innovations (Innovation in 2011).

Strategic research in firms had always some involvement of public sector research, but in some firms (especially in SMEs), the strategic research couldn't be done without some public funding, in others, public funding encouraged research in some peripheral areas (Faulkner et al. 1995). Mansfield (1998) finds by studying different industries over ten years period in US that more than 15% of new products and 11% of new processes would have been delayed or could not have been developed at all without academic research involved, and 8% of products and 7% of processes were developed with substantial assistance from recent academic research. In Estonia, it is evident that the contribution of universities and research institutes in innovation processes may not be as important. The share of firms that have not consulted universities and research institutes for their innovations has decreased in time, but the importance in innovation processes is rather low and growing slowly (Table 4). The growth is also associated with the improvement of capabilities and awareness of firms, but certainly with the public innovation support measures targeted towards industry-academia collaboration.

Table 4.

The role of universities and R&D Institutes in innovation processes of firms in Estonia (as % of innovative firms).

Importance for innovation activities	Universities			R&D Institutes		
	1998-2000	2004-2006	2006-2008	1998-2000	2004-2006	2006-2008
High	1.3	2.8	3.0	0.9	0.8	0.7
Medium	8.1	6.6	8.9	3.4	3.8	5.1
Low	10.6	9.7	14.2	7.4	6.9	9.3
Not used	80.0	80.9	73.9	88.3	88.5	84.9

Source: Innovation in 2011.

Faulkner et al. (1995) have found that the firms receive from the linkages with the universities mainly two types of knowledge: new knowledge in specialist fields of engineering and science, practical help for problem solving, instrumentalities. However, some sectoral research (Ukrainski (2008) for wood industries) has indicated that relatively higher shares of knowledge with low importance for innovation may also be connected to the non-technological knowledge or knowledge outside the firms' core activities (e.g. marketing, waste disposal, etc.) received from the universities. Although the issue is not studied in depth, it points to the importance of broader knowledge base of the universities that lies outside technological core-competences of firms.

Group of factors influencing the collaboration of firms with public sector research institutions is related to the industry or sectors the firms are active in. This comprises the character of new product development in respective industry – whether it is knowledge-led or more interactive involving user-supplier feedback (the linkage to universities is stronger in the first case). Size of the firms together with R&D intensity increases the intensity of interaction, whereby the size effects tend to overlay sectoral differences (Faulkner et al. 1995). Small firms typically face considerable resource constraints to both, formal and informal linkages, and communication problems (Corsten 1987). However, they may need even more the local universities as sources of knowledge, because in case the existing knowledge base is not sufficient (especially in fields new to the firm), the linkage is more relevant.

Considering the role of universities as highly important sources for innovation, it can be seen that only a small share of firms in Europe benefit from that knowledge (in case of universities 2–10% and in case of research institutes 1–7%). Different establishment of research systems is also revealed on the Figure via varying proportions of universities *vis-à-vis* research institutes.

Faulkner et al. (1995) discuss that some firms are simply more 'extrovert' than others and the reasons are hard to tell. They found the recruitment

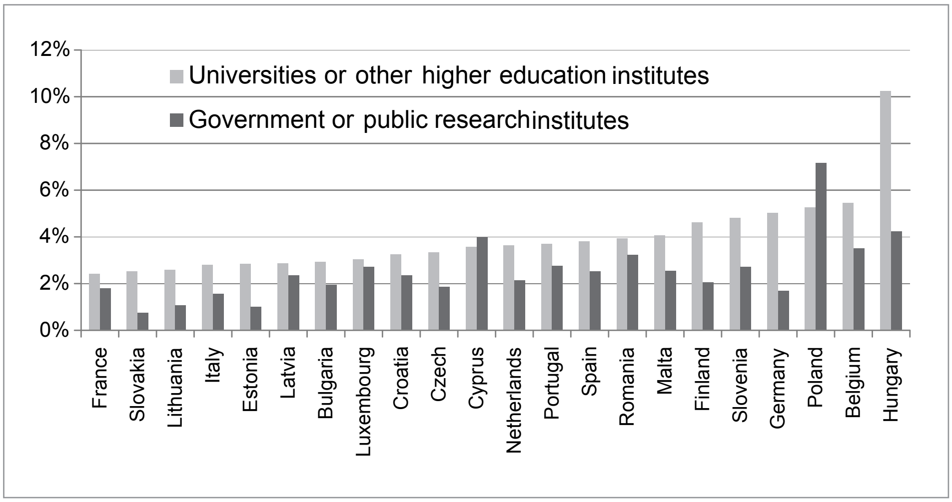


Figure 11.
Share of innovative firms recognising listed sources of knowledge as highly important for their innovation activities. Source: EUROSTAT 2011

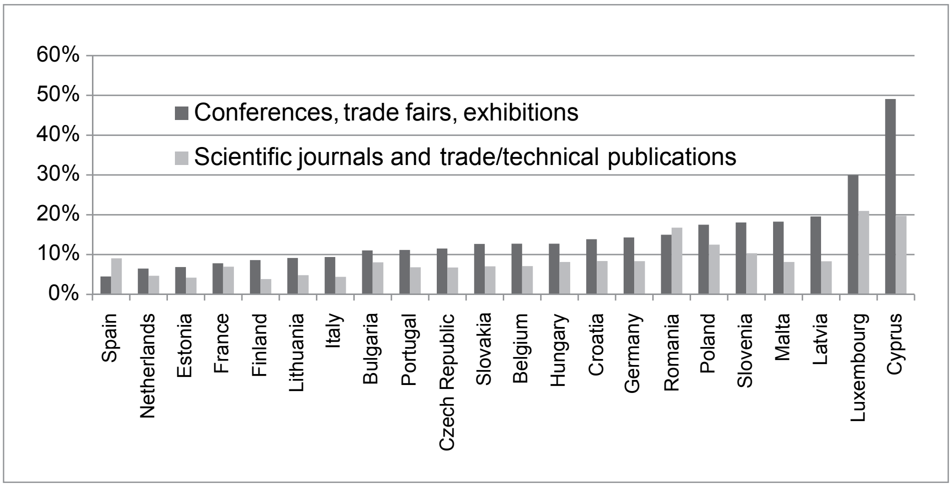


Figure 12.
Share of innovative firms recognising listed sources of knowledge as highly important for their innovation activities. Source: EUROSTAT 2011.

policy is relevant because of the networks new employees can bring to the firm (e.g. employees with doctoral or post-doctoral experience usually have networks within academia, but also know literature sources). The other factors they suggest are proprietary concerns of the firm, but also the openness of senior RD&D staff, which from one side reflects, but from the other side also influences the organisational culture. If mostly firms in the traditional sectors are not considered to have a close knowledge exchange with public sector R&D base, Meeus et al. (2004) and Schartinger et al. (2002) find the contrary.

Codified scientific knowledge is available in scientific and technical journals, but additionally, some expertise may also be obtained from conferences and exhibitions. These are relatively less costly sources, but require perhaps more in-house expertise on behalf of firms for using these sources. It seems that Estonian firms use such sources for innovation quite passively. It is argued that in process-related technologies (where the role of industrial users is more relevant) the use of public sector research is less relevant (compared with the product-related technologies) (Faulkner et al. 1995). Therefore also the relative importance of tacit knowledge seems more relevant for Estonian firms and the extent to which they use codified knowledge is also limited.

The intensity of formal and informal contacts can change over the development stage of industry or technology. Faulkner et al. (1995) find that in emerging and more dynamic fields the intensity of interactions is higher, and as the technology becomes more established the intensity declines. This can also be revealed in Estonian case (Fig. 13), where it can be seen that more dynamic industry groups – advanced knowledge providers (R&D, computers and business services), science-based manufacturers (electronics, chemicals) and specialised supplier industries (machinery, medical and optical equipment) have relatively higher interaction with universities (Fig. 13). The more traditional industry groups – scale intensive industries (food, metal) and supplier dominated industries (textile, wood, clothing etc.) rely more on their business partners and less on university knowledge. It is also true for most service industry groups: physical infrastructure (construction, electricity, water and gas supply) and supplier-dominated services (hotels, restaurants, retail, and repair services). In case of network infrastructure (post, telecommunication, financial intermediation and insurance), the internal R&D capabilities are mainly build up inside specialised departments of firms and therefore the more general university knowledge is not required for their innovation. It can be argued that based on R&D activities of those sectors the technological innovations in public sector (E-Government, electronic tax declaration, etc.) have become possible.

The availability of expertise in public sector research is a crucial factor in industry-public sector research linkage. The examples of Faulkner et

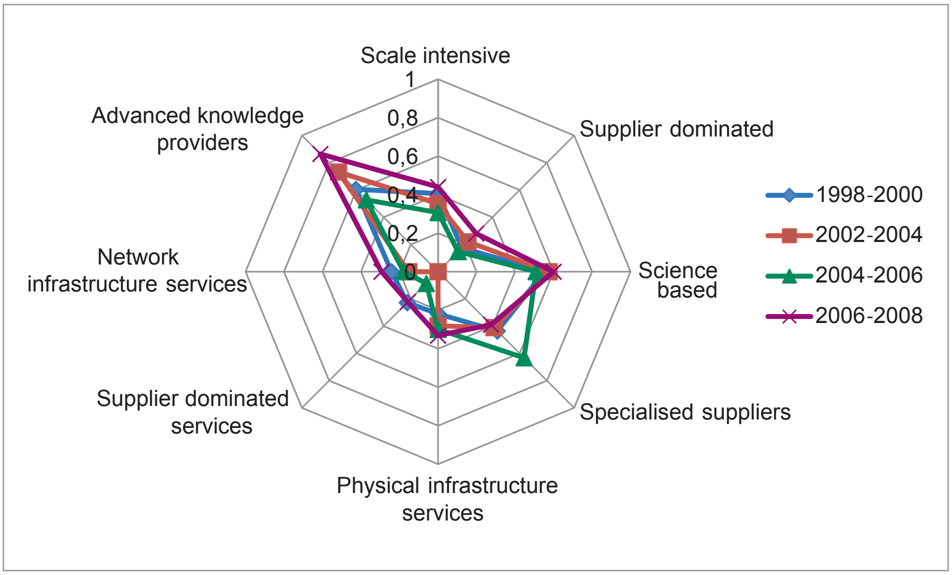


Figure 13.
Firms, who recognised universities as highly important source of knowledge (% of innovative firms). Source: Innovation in 2011.

al. (1995) show that the low level of using public sector research institutes as knowledge source does not imply the lack of research interest by the industry in respective research frontiers, but rather the paucity of available public knowledge and no modern equipment vital for development. Therefore there is less scope for the contribution of public sector research to the innovation of respective business field. Moreover, the availability of specific expertise is relevant for firms despite of general research activities (Faulkner et al. 1995). Public policy context is relevant especially for small firms overcoming the resource barriers for cooperation with public sector research institutions.

DISCUSSION AND CONCLUSIONS

Success of the Estonian research & development & innovation policy is currently too much measured using the ratio of R&D expenditures to GDP. Different evaluation teams have raised a question – does this 3% objective in so near future make sense in an Estonian context. KBE II strategy reveals too simplistic understanding of policy-makers about the mechanistic relationship between increased R&D spending and higher per capita GDP. The proposed rapid increase of R&D expenditures up to the proposed level of 3% in 2013 without significant reforms in the structure of R&D spending could be misallocation of resources. Experience of Finland, Ireland and Korea revealed that increased R&D spending and GDP *per capita* goes together with the growing share of private sector R&D. It means that private sector should be involved into this process, but it requires that innovation policy should not be limited to promoting only R&D. It is important to keep in mind that this is input target – expenditures, which say very little about how efficiently those expenditures will be made. Unless accompanied by sensible policy measures, there is also the risk of spending money unwisely, which is clearly not to the needs of the Estonian innovation system (see e.g. Koch et al. 2007).

Next to the high level of R&D expenditures Estonian innovation system should address the issue how firms obtain skills to search, understand, use and adapt the knowledge. Estonian firms are in general small and therefore mechanism about the awareness of available innovations and access to the relevant channels of communication should be organised. This system should help to avoid the situation that firms are not able to identify which technologies they need and may therefore use inappropriate technologies. But in the Estonian context the technology transfer is really a problem of learning. Hence the future of Estonian society depends heavily on success of implementation of learning (technical, managerial etc.) in firms. Learning is not an automatic process, but certain motivation to enter the learning cycle must exist. Competitive pressure from the world market should be a driver for learning.

But on the other hand governmental policy should support creation of various learning networks (regional, sectoral clusters) with the aim to increase the capacities of the firms. It reflects the current trends in the innovation policy where the earlier models of science push or market pull interpretations of the innovation model have been further developed and innovation as the social process model is used instead. The importance of social networks in the innovation process is acknowledged. Estonian economy and its innovation system are competitive only when it could utilise the existing and improve the knowledge base in the society. Therefore linkages between university and business sector play crucial role. Current chapter revealed the positive changes in the university-industry relationships in Estonia. The levels of joint publishing, co-patenting, consultancy, etc. is still below the levels of best performing countries, but it progresses rather quickly. Estonian government expenditures on research and development activities improved within the last decade, but private sector role is not yet sufficient. Therefore the above recommended policies oriented toward facilitating learning processes in firms are relevant.

REFERENCES

Cohen, W. M., Levinthal, D. A. 1990. Absorptive capacity: a new perspective of learning and innovation. Administrative Science Quarterly, 35, 128-152.

Corsten, H. 1987. Problems with cooperation between universities and enterprises: a comparative study on size of enterprise. Technovation, 6, 295-301.

EMER 2011 = Estonian Ministry of Education and Research. Research and development in Estonia. Overview and Statistics. Tartu.

EPO 2010 = Worldwide collection of published patent applications from 80+ countries. http://ep.espacenet.com/quickSearch?locale=en_EP

Etzkowitz, H. 2003. Innovation in innovation: the Triple Helix of University–Industry–Government relations. Social Science Information, 3, 42, 293-337.

Etzkowitz, H., Leydesdorff, L. 1998. 'Introduction' to 'The endless transition: a "Triple Helix" of University–Industry–Government relations. Minerva, 36, 203-208.

Etzkowitz, H., Leydesdorff, L. 2000. The dynamics of innovation: from national systems and "Mode 2" to a Triple Helix of University-Industry-Government relations. Research Policy, 29,109-123.

Evaluation of the Finnish National Innovation System – Full Report. 2009. Helsinki University Print.

EUROSTAT 2011 = EUROSTAT Science, Technology and Innovation Database.

Faulkner, W., Senker, J., Velho, L. 1994. Knowledge Frontiers: Industrial Innovation and Public Sector Research in Biotechnology, Engineering Ceramics and Parallel Computing. Clarendon Press, Oxford.

Freeman, C. 2002. Continental, national and sub-national innovation systems – complementarity and economic growth. Research Policy, 31,191-211.

Gibbons, M., et al. 1994. The New Production of Knowledge: the Dynamics of Science and Research in Contemporary Societies. Sage, London.

Hasan, I., Tucci, C. 2010. The innovation-economic growth nexus: global evidence. Research Policy, 39, 1264-1276.

Innovation in Estonian Enterprises 2006-2008. Innovation Studies X/X, forthcoming.

IUS (Innovation Union Scoreboard) 2010 = Database.

Kalvet, T. 2001. ICT, Innovations and Innovation policy: The Case of Estonia Working Paper. Archimedes Foundation and PRAXIS Center for Policy Studies.

Karo, E., Kattel, R. 2010. The copying paradox: why converging policies but diverging capacities in Eastern European innovation systems? International Journal of Institutions and Economies, 2, 2, 167-206.

Knowledge-based Estonia 2002. Estonian R&D Strategy 2002-2006. Research and Development Council.

Knowledge-based Estonia 2007. Estonian R&D and Innovation Strategy 2007-2013. Estonian Ministry of Education and Research, Tartu.

Koch, P., Pukl, B., Wolters, A. 2007. OMC Policy Mix Review Report Country Report Estonia. Tallinn.

Kuznets, S. 1973. Modern economic growth: findings and reflections. The American Economic Review, 63, 3, 247-258

Lundvall, B.-Å. 2007. National Innovation System: Analytical Focusing Device and Policy Learning Tool. Swedish Institute for Growth Policy Studies, Working Paper 4.

Meeus, M. T. H., Oerlemans, L. A. G., Hage, J. 2004. Industry-public knowledge infrastructure interaction: intra- and inter-organizational explanations of interactive learning. Industry and Innovation, 11, 4, 327-352.

Nowotny, H., Scott, P., Gibbons, M. 2001. Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty. Polity Press, Cambridge.

OECD 1999 = National Innovation Systems. OECD, Paris.

OECD 2011 = Main Science and Technology Indicators database.

Ojakivi, M. 2008. Eesti ülikoolide leiutised taotlevad rekordarvu patente. Äripäev, 21.01.2008. http://prahijaht.epl.ee/artikkel/415426

Schartinger, D., Rammer, C., Fischer, M. M., Fröhlich, J. 2002. Knowledge interactions between universities and industry in Austria: sectoral patterns and determinants. Research Policy, 31, 303-328.

Tiits, M., Kattel, R., Kaarli, R. 2003. Competitiveness and Future Outlooks of the Estonian Economy. R&D and Innovation Policy Review. Research and Development Council, Tallinn.

Ukrainski, K. 2008. Sources of Knowledge Used in Innovation: an Example of Estonian Wood Industries. Tartu University Press, Tartu.

Varblane, U., et al. 2007. Can the national innovation systems of the new EU member states be improved? Post-Communist Economies, 2007, 19, 4, 399-416.

Varblane, U., Mets, T., Ukrainski, K. 2008. The role of the University-Industry-Government linkages in the innovation processes of the small catching-up economy. Industry and Higher Education, 22, 6, 373-386.

ESTONIAN SCIENCE ESTIMATED THROUGH BIBLIOMETRIC INDICATORS

J. Allik

INTRODUCTION

The idea of bibliometric analysis of individuals, institutions, or countries has transformed from a noble but not a very practical idea (de Solla Price 1965) into a regular tool for evaluation of the scientific productivity and quality (Moed 2005). Although a panel of knowledgeable experts is perhaps the best way how to evaluate the present situation and prospects of development of scientific research in a given country (Kristapsons at al. 2003), it is much cheaper and faster to rely on bibliometric indicators available from such databases as Web of Science, Essential Science Indicators, or Scopus. For example, a recent report showed that the productivity of Russian scientists has remained approximately on the same level during the last 20 years (Adams, King 2010). Not only Russia's relative growth in scientific productivity lags behind against two of its counterparts in the BRIC group of nations, India and Brazil, the structure of strengths and weaknesses of the Russia's research are characteristic not to the current, but to the sciences of the 20th century (Adams, King 2010). In this regard Estonia, like its two neighbours, Latvia and Lithuania, is an interesting historical case since it is not only fascinating but also instructive to see what has happened to Estonian science after almost two decades of development separate from the obsolete Byzantine system.

A BIBLIOMETRIC PORTRAIT OF THE ESTONIAN SCIENCE

The analysis is based on the Internet version of the Thomson Reuters Web of Science (WoS) databases covering mainly the period from 1990 to 2010. The WoS includes about 11,000 of the most influential research journals in the world. It contains five citation indices: Science Citation Index (1900-present), Social Sciences Citation Index (1956-present), Arts & Humanities Citation Index (1975-present), Conference Proceedings Citation Index – Science (1990-present), and Conference Proceedings Citation Index – Social Science & Humanities (1990-present). The WoS also provides different analytical tools for the analysis of countries/territories and research institutions.

The Essential Science Indicators (ESI) is a resource that enables analytical tools for ranking scientists, institutions, countries, and journals. It is based on journal article publication counts and citation data from WoS and analyses a ten-year and plus n-months rolling period. ESI covers 10 million articles in 22 specified fields of research (excluding humanities), and is updated every two months. ESI provides both total citation counts and cites per paper scores for different countries or territories. ESI is limited to the journal articles indexed in the WoS only. No books, book chapters, or articles published in journals not indexed by WoS are taken into account in ESI, either in terms of publication or citation counts. The latest release of ESI covered the period from January 1, 2000 to December 31, 2010.

Productivity from 1990 to 2010

In 1990, just before the collapse of the Soviet Union, Estonia, Latvia and Lithuania had very similar starting positions. Scientists in each of these three Baltic countries published approximately 300 papers per year in journals indexed by the WoS. Twenty one years later, in 2010, Estonian, Latvians and Lithuanians published 1730, 841, and 2435 articles respectively from about 1.9 million total publications in that year. Fig. 1 demonstrates the growth of publication in the three Baltic countries from 1990 to 2010. The relative contributions of Estonia, Latvia, and Lithuania were 0.092%, 0.045%, and 0.130% from the total number articles published that year.

Thus, Estonia and particularly Lithuania have increased their publication activity considerably during the last twenty years while Latvia's increase has been rather modest. In both absolute and relative terms the intensity of scientific publications in the Baltic countries is still low. Finnish researchers, for example, published 12,784 articles in 2010 which is about 0.68% of the total publications covered by WoS. The contribution of Estonian science to the WoS is 7.4 times smaller than that of Finland. Since Finnish population is about 4.1 times larger than in Estonia, this means that even after taking into account the size of the countries the Estonian contribution to the World publications is 1.8 times smaller than in Finland. It is important to notice, however, that this gap was much wider ten years ago.

In 2006/2007 the current list of about 10,000 journals in the WoS was expanded by about 700 regional journals. Among these 700 regional journals 27 are published in Lithuania and 7 in Estonia (Zavadskas et al. 2011). This expansion probably accounts for a large peak in 2008 for Lithuanian papers.

Figure counts all publications (also abstracts) included in the WoS based not on their nominal publication date but on the actual time they were entered into the database. All publications were included since it has been argued that conference presentations are more important vehicles of scientific communication in some areas than others.

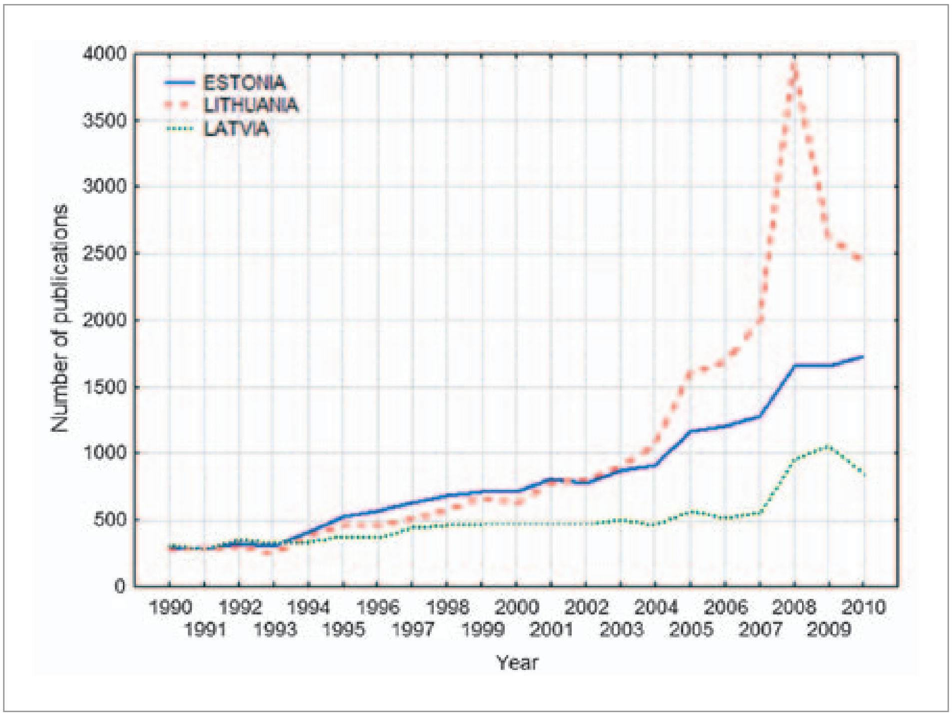


Figure 1.
The total number of publications in the Web of Science database authored by Estonian, Latvian, and Lithuanian scientists.

The impact of Estonian science (2000-2010)

The number of papers published in the journals indexed in the WoS characterises productivity of science in a given country. But science is more about quality of ideas: how much ideas what have been made public have inspired other researchers. One of the most basic indicators of the quality of scientific ideas is the citation rate. Table 1 provides the ranking of countries/territories on the basis of their impact, the number of citations per article. Only these countries/territories are listed which were able to exceed the 50% threshold of the essential science established for countries/territories and to publish 2,000 or more articles during the last 11-year period, from January 1, 2000 to December 31, 2010. As can be seen by the impact Estonia (33) is ahead of all former Communist bloc countries except Hungary (28). It is also ahead of several older EU members like Portugal (35) and Greece (39). Nevertheless, the impact of Estonian articles is still 10.8% below the world average.

Table 1.

Countries or territories producing more than 2,000 papers in the period from January 1, 2000 to December 31, 2010.

No	Country/Territory	Papers	Citations	Citations Per Paper	Increase % 2007-2010
1	Switzerland	179,750	3,052,454	16.98	14.3
2	USA	3,069,500	49,593,619	16.16	14.0
3	Denmark	97,261	1,563,229	16.07	19.2
4	Netherlands	249,334	3,923,238	15.73	17.5
5	Scotland	109,965	1,724,071	15.68	19.4
6	Iceland	5,113	79,810	15.61	20.4
7	England	703,545	10,595,659	15.06	18.8
8	Sweden	180,269	2,702,126	14.99	17.9
9	Finland	89,436	1,244,315	13.91	15.7
10	Belgium	135,733	1,869,498	13.77	20.6
11	Canada	447,453	5,981,892	13.37	15.5
12	Germany	789,416	10,550,122	13.36	19.4
13	Austria	94,471	1,230,417	13.02	20.9
14	Israel	112,266	1,443,458	12.86	19.1
15	Norway	70,179	896,351	12.77	18.9
16	France	561,403	7,061,061	12.58	18.1
17	Wales	37,155	463,980	12.49	20.1
18	Australia	296,439	3,586,005	12.10	18.6
19	Italy	424,827	5,067,540	11.93	18.7
20	Costa Rica	3,258	38,754	11.90	35.4
21	North Ireland	18,131	214,789	11.85	21.5
22	Ireland	43,668	503,196	11.52	15.9
23	New Zealand	57,088	624,815	10.94	19.7
24	Spain	328,890	3,470,696	10.55	21.5
25	Peru	3,960	41,540	10.49	30.5
26	Japan	791,896	8,305,412	10.49	18.5
27	Kenya	7,468	77,584	10.39	21.4
28	Hungary	50,303	501,975	9.98	27.5
29	Uganda	3,328	32,799	9.86	10.3
30	Ecuador	2,083	19,897	9.55	– *

31	Singapore	62,947	589,433	9.36	45.8
32	Tanzania	3,738	34,904	9.34	31.0
33	Estonia	8,664	80,354	9.27	17.8
34	Uruguay	4,604	42,102	9.14	15.4
35	Portugal	59,274	540,370	9.12	28.5
36	Chile	32,639	291,069	8.92	18.8
37	Philippines	5,651	49,453	8.75	18.4
38	Luxembourg	2,369	20,391	8.61	–
39	Greece	81,295	699,215	8.60	32.1
40	South Africa	52,844	437,896	8.29	25.2
41	Senegal	2,198	18,210	8.28	10.7
42	Indonesia	6,346	51,614	8.13	26.6
43	Argentina	56,922	462,432	8.12	21.7
44	Zimbabwe	2,300	18,584	8.08	25.7
45	Czech Republic	65,804	517,145	7.86	25.4
46	Armenia	4,222	32,704	7.75	31.1
47	Nepal	2,052	15,534	7.57	–
48	Thailand	30,312	227,638	7.51	31.3
49	Yugoslavia	6,328	46,502	7.35	61.2
50	Ghana	2,670	19,500	7.30	30.6
51	Vietnam	6,655	48,209	7.24	26.1
52	South Korea	266,682	1,892,966	7.10	29.8
53	Mexico	73,720	521,828	7.08	22.1
54	Taiwan	169,555	1,195,634	7.05	27.0
55	Latvia	3,771	25,659	6.80	14.9
56	Slovenia	23,113	157,059	6.80	20.8
57	Sri Lanka	3,124	21,164	6.77	27.3
58	Poland	150,392	1,011,909	6.73	18.5
59	Cyprus	3,534	23,738	6.72	14.1
60	Venezuela	11,927	79,988	6.71	19.8
61	Colombia	12,365	80,798	6.53	6.5
62	Lebanon	5,159	33,470	6.49	29.3
63	Cameroon	3,595	23,289	6.48	15.3
64	Bangladesh	6,336	40,787	6.44	33.6
65	Brazil	200,199	1,285,332	6.42	17.2

66	Bulgaria	18,971	120,976	6.38	23.6
67	Slovakia	23,963	152,422	6.36	21.8
68	Cuba	7,319	44,084	6.02	27.0
69	Peoples R China	762,098	4,578,689	6.01	42.4
70	Ethiopia	3,469	20,781	5.99	24.0
71	Rep of Georgia	3,150	18,452	5.86	29.1
72	India	279,116	1,610,511	5.77	32.6
73	Moldova	2,024	11,594	5.73	53.2
74	Croatia	20,875	114,081	5.46	18.4
75	Kuwait	5,972	31,289	5.24	32.0
76	U Arab Emirates	6,052	31,237	5.16	36.5
77	Turkey	145,028	741,550	5.11	34.8
78	Egypt	34,802	176,413	5.07	32.0
79	Lithuania	11,371	57,545	5.06	2.2
80	Morocco	11,324	56,720	5.01	28.1
81	Oman	3,019	14,805	4.90	38.4
82	Russia	275,367	1,311,661	4.76	19.6
83	Saudi Arabia	17,895	80,024	4.47	17.6
84	Romania	34,574	152,168	4.40	7.6
85	Jordan	7,492	32,835	4.38	39.9
86	Malaysia	22,138	96,962	4.38	10.6
87	Belarus	10,507	44,961	4.28	24.4
88	Pakistan	17,673	75,191	4.25	18.1
89	Iran	66,136	272,282	4.12	26.4
90	Ukraine	44,044	180,035	4.09	25.1
91	Tunisia	14,027	57,336	4.09	29.4
92	Algeria	9,051	36,142	3.99	27.5
93	Kazakhstan	2,391	9,333	3.90	34.5
94	Uzbekistan	3,356	12,923	3.85	46.9
95	Nigeria	13,192	50,288	3.81	22.1
96	Azerbaijan	2,598	6,519	2.51	30.1
97	Serbia	11,251	25,001	2.22	–

Notes: * = Only countries published 2,000 or more papers during the 11-year period; the first 22 countries/territories are above the world average (9.5). The ESI also lists what it calls Yugoslavia as one of the fastest growing sciences. This was omitted in this analysis since it is unclear what is meant by this nonexistent political reality.

Beside the total number of papers, citations and citations per paper (impact), the last column show the percentage of the increase of the impact during the period 2000–2010 relative to the previously published ranking for the period of 1997–2007 (Allik 2008). Compared to the previously published list there are several newcomers who were absent or were not able to pass the 2,000 papers threshold in the list for 1997–2007. The average increase of the impact across all listed countries or territories was 24.1% during the last four years. The fastest growing sciences were in Moldova and Singapore where the number of citations per paper increased by 53.2% and 45.8% respectively. Although Estonian science generally preserved her world ranking several countries (e.g., Hungary, Singapore, and Tanzania) were growing with a faster pace and passed Estonia in the world ranking. It would be important to analyse how Hungary has achieved a more rapid increase (+27.5%) in the impact of their scientific papers. Among the three Baltic countries the fastest growth of the citations per paper was in Estonia (17.8%) which slightly exceeds the pace of Latvia (14.9%) but was considerably faster than the increase of citations per paper in Lithuania (2.2%) which was in fact the slowest among the listed 97 countries or territories. Although Lithuania was able to increase considerably scientific output (11,371 papers in the period 2000-2010) compared to her neighbours Estonia (8,664) and Latvia (3,771), its stagnated number of citations (5.06 per paper) has moved Lithuanian science from the 64th position in 1997–2007 to the 79th position in 2000–2010. This ranking is only three position ahead of Russia (4.79) occupying the 82nd position in the world ranking.

Fig. 2 demonstrates trends in the average citation rates of Estonian papers in 21 fields of science during the last five years (2006-2010) for the respective 11-year running periods. The Multidisciplinary category was excluded due to its smallness and volatile character. Ten out of 21 fields which demonstrated the increase of the impact in 2000-2010 compared to the citations per paper in 1996-2006 are shown. Molecular Biology and Genetics and Clinical Medicine demonstrated the largest (more than 20%) growth during the last five years. Two of the most successful fields in Estonian science – Plant and Animal Science and Environment/Ecology – were also growing even though they have already exceeded the world average level of the citations per paper. Some growth can be also observed in four other fields (Geosciences, Psychiatry/Psychology, Computer Science, and Immunology) which average level is still considerably below the world average citations rate.

Fig. 3 demonstrates the remaining 11 fields in which the growth was absent or even negative. Almost 50% decline of impact was observed in the Materials Science. This decline was expected since in 1996-2006 the impact

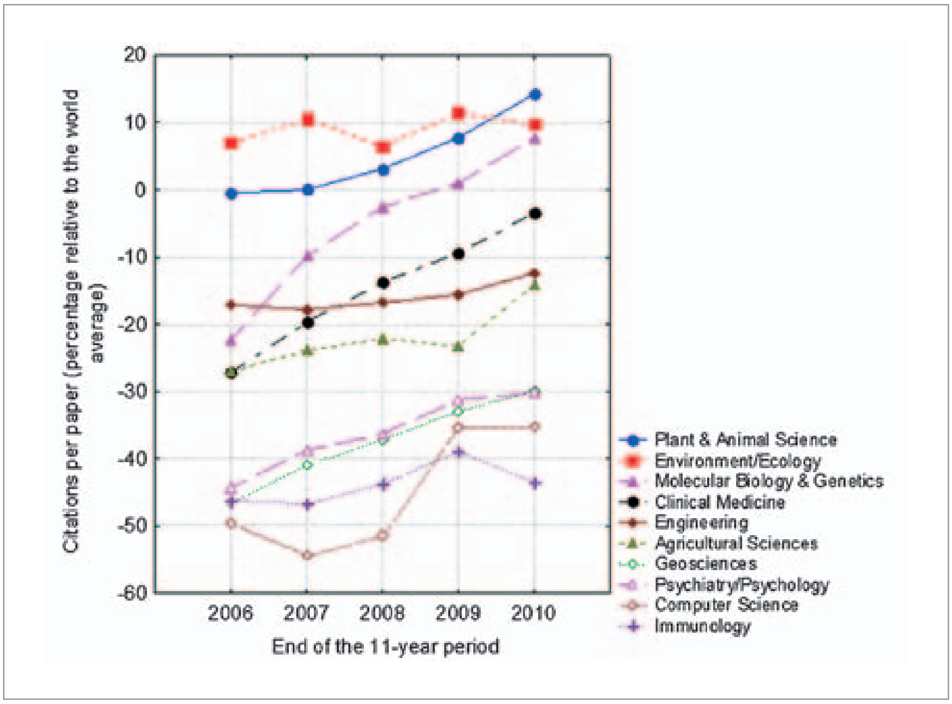


Figure 2.
Trends in the citations per paper of Estonian articles relative to the world average in fields where the growth was observed.

of the Estonian Materials Sciences was about 60% higher than in the field in general. Even after this deep decline the Materials Science in Estonia remains about 10% above the world average citations rate. Beside Materials Sciences a considerable decline was observed in three other fields: Pharmacology and Toxicology, Social Sciences general, and Space Sciences. In the rest of the fields (Chemistry, Biology and Biochemistry, Physics, Microbiology, Mathematics, Neuroscience and Behaviour, Economic and Business) the impact of an average paper has not changed much during the last few years.

It is important to notice, however, that all these trends happen on a very different level of performance. There were only 4 fields (Plant and Animal Science, Materials Science, Environment/Ecology, and Molecular Biology and Genetics) in Estonia which exceed the average citation level per paper. Two fields – Chemistry and Clinical Medicine – are very close to the average performance while 10 fields are more than -30% percent below the world average performance. For a comparison, the most successful field in Russian science is still Physics which nevertheless performs -26.1% below the world average. On average, a scientific paper authored by at least one Russian author

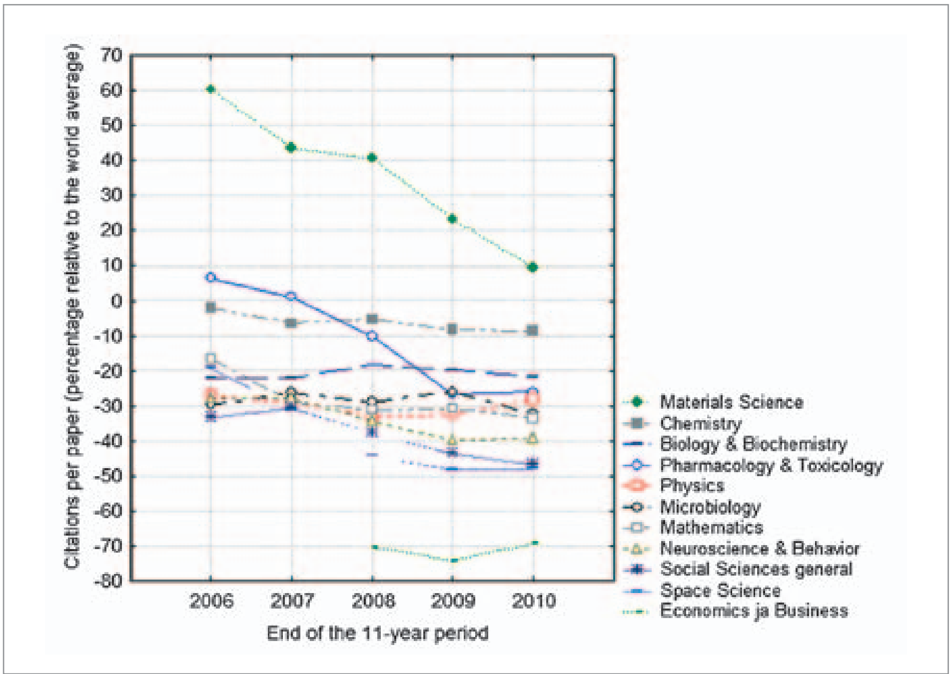


Figure 3.
Trends in the citations per paper of Estonian articles relative to the world average in fields where the growth was not observed.

is cited –56% less than a scientific paper in the Web of Science database on average.

There were 89 Estonian papers exceeding the top 1% threshold of citations. Latvia had 23 and Lithuania had 87 papers that were among highly cited papers in the respective fields. For comparison, Hungary had 467, Finland 1147, and Iceland 125 highly cited papers. The most cited Estonian paper collected 1,103 citations during six year period after its publication (Dormandy et al. 2005). The rest of the top 10 papers were cited 243 or more times each (Bayatian et al. 2007; Bjorksten et al. 2001; Forette et al. 2002; Giudice et al. 2004; Hibbett et al., 2007; Mahdavi et al., 2002; Mann et al., 2005; Remm et al. 2001; Wright et al. 2004).

There were 15 scientists having full or part time affiliation with one of Estonian universities or research organisation who exceeded the top 1% threshold of cumulative citations in their respective fields. Table 2 presents bibliometric characteristics of these 15 scientists. Beside them there were at least 4 researchers of Estonian origin who also exceed the top 1% citation threshold: Rando Allikmets in Molecular Biology and Genetics, Jaak Panksepp and Mart

Table 2.

Estonian scientists exceeding the top 1% cumulative citation threshold in one of the 22 fields of science.

Scientist	Field	Number of highly cited papers	Papers	Cita-tions	Cita-tions per paper	1% citation threshold of the field	% by which the threshold is exceeded	Total papers	Total cita-tions	Total citations per paper
Aarik, Jaan	Materials Science	0	33	579	17.55	519	11.6	74	1 223	16.53
Allik, Jüri	Psychiatry/Psychology	3	43	688	16.00	584	17.8	55	738	13.42
Dumas, Marlon	Computer Science	1	38	303	7.97	171	77.2	118	1 254	10.63
Karelson, Mati	Chemistry	0	61	1 091	17.89	1077	1.3	83	1 220	14.70
Kivisild, Toomas	Molecular Biology & Genetics	2	40	2 779	69.47	1368	103.1	61	3 217	52.74
Kollist, Hannes	Plant & Animal Science	5	19	533	28.05	440	21.1	19	533	28.05
Kukli, Kaupo	Materials Science	1	49	1 025	20.92	519	97.5	95	1 653	17.40
Kõljalg, Urmas	Plant & Animal Science	4	17	812	47.76	440	84.5	21	962	45.81
Langel, Ülo	Pharmacology & Toxicology	2	20	977	48.85	398	145.5	126	3 007	23.87
	Biology & Biochemistry		56	1 125	20.09	828	35.9			
Liira, Jaan	Environment/Ecology	2	24	598	24.92	453	32.0	48	806	16.79
Niinemets, Ülo	Plant & Animal Science	8	59	1 280	21.69	440	190.9	95	2 625	27.63
	Environment/Ecology		25	1 165	46.6	453	157.2			
Näätänen, Risto	Neuroscience & Behaviour	3	111	2 923	26.33	1072	172.7	163	4 344	26.65
	Psychiatry/Psychology		21	647	30.81	584	10.8			
Zobel, Martin	Environment/Ecology	7	30	727	24.23	453	60.5	63	1 133	17.98
Tedersoo, Leho	Plant & Animal Science	4	12	461	38.42	440	4.8	16	516	32.25
Villems, Richard	Molecular Biology & Genetics	2	37	2 077	56.14	1368	51.8	59	2 712	45.97

Saarma Neuroscience & Behaviour, and Endel Tulving in Psychiatry/Psychology. However, their bibliometric data contributed to their current residence countries, not Estonia. Three Estonian scientists (Ülo Langel, Ülo Niinemets and Risto Näätänen) excelled in two different fields. Ülo Niinemets exceeds the average impact in the Plant & Animal Science by more than 190%. The latest entrée to the top 1% list was Leho Tedersoo who needed only 12 papers to collect 461 citations which exceed the top 1% citation threshold (440) in the Plant & Animal Science. As expected, the highest concentration highly-cited scientists were in Plant & Animal Science (4) and in Environment/Ecology (3). Molecular Biology & Genetics, Materials Science, Psychiatry/Psychology are all represented by two researchers who exceeded the top 1% citation threshold. There were 12 fields (including Multidisciplinary) of science where researchers working in Estonia were not able to achieve the top 1% level of cumulative citations to their papers.

GENERAL EVALUATION

There is no doubt that bibliometric indicators reflect to some extent decisions made by politicians and science administrators many years earlier.

For example, almost ten years ago it was already rather transparent that science in Latvia was starting to stagnate (Allik 2003).

During the last 10 years Latvian scientists published 3 times less high quality papers in international peer-reviewed journals than Lithuanians although only twenty years ago all three Baltic countries had approximately the same level of productivity. It is surprising that Latvia was able to preserve and even increase the quality (citation per paper) while the total number of papers demonstrated an increase only during the last three years. Although there was a considerable increase in productivity during the last three years, it is still an open question whether the Latvian science has enough resources for maintaining quality of the published papers on the current level during the next few years.

In turn, Lithuania decided to emphasise quantity rather than quality of their scientific publications. Dozens of local journals were established which publish not only in English but Lithuanian papers as well. These local journals have obviously less stringent criteria for the acceptance of scientific contributions especially when they come from the Lithuanian authors. When the scientific business of Thomson Reuters announced on May 28, 2008 that 700 new regional journals have been added to the Web of Science, Lithuania succeeded to include 27 of their local journals (e.g. "Elektronika ir Elektrotechnika", "Medicina-Lithuania", "Journal of Vibroengineering", "Materials Science – Medziagotyra", "Inzinerine Ekonomika – Engineering Economics", "Logos-Vilnius", "Lithuanian Journal of Physics", "Veterinarija ir Zootechnika", "Mechanika" (Kaunas University of Technology), "Problemos", "Informatica", "Zemdirbyste – Agriculture", "Transport", "Chemija" and many others). As a consequence, the impact of Lithuanian scientific papers has dramatically dropped compared to all other countries in the world. One likely candidate for the decrease of the impact is a diminished motivation of Lithuanian researchers to publish their papers in the most prestigious and competitive international journals since it is much easier to publish in one of these local journals which impact is relatively low. As a consequence, the competitiveness of the Lithuanian science has dramatically decreased. The local science administrators are probably not even aware that in the result of their well intended measures Lithuanian science is rapidly dropping to the same league with Russia from which they hoped to distance about twenty years ago (Adams, King 2010).

Looking retrospectively it seems that political decisions concerning Estonian science have been relatively competent, especially in comparison with Latvia and Lithuania. The relative citations per Estonian scientific paper have steadily increased along with the total scientific productivity. Estonian science seems to be relatively well balanced since the threshold of the essential science was passed in all 22 scientific fields. Only 20 years ago Estonian research was a part of the Soviet science which core strengths were in the fields that were perceived supporting military strength of the Soviet Union such as physical sciences and engineering. Unlike Russia, Estonia was able to restructure her science to meet requirement of the 21st century which is dominated by life sciences and the environment. Emulating and in direct co-operation with Scandinavian countries the highest excellence of Estonian sciences was achieved in Plant & Animal Science, Environment/Ecology and Molecular Biology & Genetics.

For many years it was complained that the quality of Clinical Medicine in Estonia does not meet expectations (Kaasik, Uibo 2005). Data presented in this paper testify, however, that Clinical Medicine is one of the fastest improving fields which most likely will be ahead of the world average already in the next year. In some cases it is possible to guess that changes in the citations rate are caused by accidents such as a high-profile researcher changed his or her field of interests. In other cases observed trends reflect more substantial processes requiring intervention from the science administrators and strategists. For example, Space Science was one of the most visible fields on the Estonian science terrain with frequent and heavily cited papers in the most prestigious journals (Einasto et al. 1997; Zeldovich et al. 1982). Another evidence for the eminence was the Marcel Grossmann award for 2009 that was given to Jaan Einasto for his pioneering contributions in the discovery of dark matter and the cosmic web. Bibliometric data reveal, however, that Estonian researchers have published 176 papers in journals classified to the category of the Space Science during the last 11 years which have attracted only 1,318 citations.

Another disturbing development is the falling of the Estonian Social Sciences impact far below (almost -40%) the world average level. In combination with the lowest impact of the Economics & Business (approximately –70% below the world average) it should be viewed as a warning that the recent policies towards these areas have not been sufficiently effective. Equally alarming is the decline of the impact of Estonian papers in Neuroscience and Behaviour category. Although Risto Näätänen, one of the world most renowned neuroscientist, was elected professor of cognitive neuroscience at the University of Tartu few years ago, most of his previous papers still contribute to scientific statistics of Finland where these studies were executed. It may be not a coincidence that two other outstanding Estonian researchers Jaak Panksepp and Mart Saarma have passed top 1% level of citations also in Neuroscience and Behaviour category. Although more than a decade has passed from what was called the Decade of the Brain, it is rather obvious that no modern country pretending to have a knowledge-based economy can afford a weakness in the field that studies nervous system and human behaviour.

REFERENCES

Adams, J., King, C. 2010. Russia: Research and collaboration in the new geography of science. Evidence, Leeds.

Allik, J. 2003. The quality of science in Estonia, Latvia, and Lithuania after the first decade of independence. Trames, 7, 40-52.

Allik, J. 2008. Quality of Estonian science estimated through bibliometric indicators (1997-2007). Proceedings of the Estonian Academy of Sciences, 57, 4) 255-264.

Bayatian, G. L., et al. 2007. CMS physics technical design report, volume II: Physics performance. Journal of Physics G-Nuclear and Particle Physics, 34, 6, 995-1579.

Bjorksten, B., et al. 2001. Allergy development and the intestinal microflora during the first year of life. Journal of Allergy and Clinical Immunology, 108, 4, 516-520.

de Solla Price, D. 1965. Networks of scientific papers. Science, 149, 510-515.

Dormandy, J. A., et al. 2005. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events): a randomised controlled trial. Lancet, 366, 9493, 1279-1289.

Einasto, J., et al. 1997. A 120-Mpc periodicity in the three-dimensional distribution of galaxy superclusters. Nature, 385, 6612, 139-141.

Forette, F., et al. 2002. The prevention of dementia with antihypertensive treatment. Archives of Internal Medicine, 162, 18, 2046-2052.

Giudice, G. F., et al. 2004. Towards a complete theory of thermal leptogenesis in the SM and MSSM. Nuclear Physics B, 685, 89-149.

Hibbett, D. S., et al. 2007. A higher-level phylogenetic classification of the Fungi. Mycological Research, 111, 509-547.

Kaasik, A.-E., Uibo, R. 2005. Teadusmõte Eestis: Arstiteadus [Scientific thought in Estonia: Medicine]. Eesti Teaduste Akadeemia, Tallinn.

Kristapsons, J., Martinson, H., Dagyte, I. 2003. Baltic R&D systems in transitions: Experiences and future prospects. Zinatne, Riga.

Mahdavi, J., et al. 2002. Helicobacter pylori SabA adhesin in persistent infection and chronic inflammation. Science, 297, 5581, 573-578.

Mann, J. J., et al. 2005. Suicide prevention strategies – A systematic review. Jama-Journal of the American Medical Association, 294, 16, 2064-2074.

Moed, H. F. 2005. Citation analysis in research evaluation. Springer, Dordrecht.

Remm, M., Storm, C. E. V., Sonnhammer, E. L. L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of Molecular Biology, 314, 5, 1041-1052.

Zavadskas, E. K., Kirvaitis, R., Dagiene, E. 2011. Scientific publications released in the Baltic States. Scientometrics, 88, 1, 179-190.

Zeldovich, Y. B., Einasto, J., Shandarin, S. F. 1982. Giant voids in the universe. Nature, 300, 5891, 407-413.

Wright, I. J., et al. 2004. The worldwide leaf economics spectrum. Nature, 428, 6985, 821-827.

INSTITUTIONAL AFFILIATION OF AUTHORS

Aarik, Jaan
Institute of Physics, University of Tartu

Abarenkov, Kessy
Institute of Ecology and Earth Sciences, University of Tartu

Alavere, Helene
Estonian Genome Center, University of Tartu

Alles, Harri
Institute of Physics, University of Tartu

Allik, Annely
Estonian Genome Center, University of Tartu

Allik, Jüri
Institute of Psychology, University of Tartu
Estonian Academy of Sciences

Antonov, Maksim
Department of Materials Engineering, Tallinn University of Technology

Blinova, Irina
National Institute of Chemical Physics and Biophysics

Eerme, Martin
Department of Machinery, Tallinn University of Technology

Ellervee, Peeter
Department of Computer Engineering, Tallinn University of Technology

Engelbrecht, Jüri
Estonian Academy of Sciences
Institute of Cybernetics, Tallinn University of Technology

Esko, Tõnu
Estonian Genome Center, University of Tartu

Fischer, Krista
Estonian Genome Center, University of Tartu

Freiberg, Arvi
Institute of Physics, University of Tartu
Estonian Academy of Sciences

Haller, Toomas
Estonian Genome Center, University of Tartu

Hass, Merli
Estonian Genome Center, University of Tartu

Heinsalu, Els
National Institute of Chemical Physics and Biophysics

Hektor, Andi
National Institute of Chemical Physics and Biophysics

Hizhnyakov, Vladimir
Institute of Physics, University of Tartu
Estonian Academy of Sciences

Hollstein, Thomas
Department of Computer Engineering, Tallinn University of Technology

Hussainova, Irina
Department of Materials Engineering, Tallinn University of Technology

Ivask, Angela
National Institute of Chemical Physics and Biophysics

Jänes, Alar
Institute of Physical Chemistry, University of Tartu

Jervan, Gert
Department of Computer Engineering, Tallinn University of Technology

Jutman, Artur
Department of Computer Engineering, Tallinn University of Technology

Kaasik, Aidula-Taie
Estonian Genome Center, University of Tartu

Kahru, Anne
National Institute of Chemical Physics and Biophysics

Kaljurand, Mihkel
Department of Chemistry, Tallinn University of Technology

Kalm, Volli
Institute of Ecology and Earth Sciences, University of Tartu

Kannike, Kristjan
National Institute of Chemical Physics and Biophysics

Karelson, Mati
Department of Chemistry, Tallinn University of Technology
Institute of Chemistry, University of Tartu
Estonian Academy of Sciences

Karjust, Kristo
Department of Machinery, Tallinn University of Technology

Kasemets, Kaja
National Institute of Chemical Physics and Biophysics

Kers, Jaan
Department of Materials Engineering, Tallinn University of Technology

Kikas, Jaak
Institute of Physics, University of Tartu

Kirsimäe, Kalle
Institute of Ecology and Earth Sciences, University of Tartu

Kõljalg, Urmas
Institute of Ecology and Earth Sciences, University of Tartu
Natural History Museum, University of Tartu

Koppel, Ilmar
Institute of Chemistry, University of Tartu
Estonian Academy of Sciences

Krjutskov, Kaarel
Estonian Genome Center, University of Tartu

Kruus, Margus
Department of Computer Engineering, Tallinn University of Technology

Kübarsepp, Jakob
Department of Materials Engineering, Tallinn University of Technology

Kukli, Kaupo
Institute of Physics, University of Tartu

Kull, Kalevi
Institute of Philosophy and Semiotics, University of Tartu

Kulu, Priit
Department of Materials Engineering, Tallinn University of Technology

Küttner, Rein
Professor emeritus, Tallinn University of Technology
Estonian Academy of Sciences

Lang, Valter
Institute of History and Archaeology, University of Tartu
Estonian Academy of Sciences

Leego, Erkki
Hansson, Leego & Partner Ltd.
Institute of Computer Science, University of Tartu

Leete, Art
Institute of Culture Studies and Arts, University of Tartu

Leito, Ivo
Institute of Chemistry, University of Tartu

Leitsalu, Liis
Estonian Genome Center, University of Tartu

Lilienthal, Kadri
Estonian Genome Center, University of Tartu

Lõhmus, Asko
Institute of Ecology and Earth Sciences, University of Tartu

Lopp, Margus
Department of Chemistry, Tallinn University of Technology

Lushchik, Aleksander
Institute of Physics, University of Tartu

Lust, Enn
Institute of Chemistry, University of Tartu
Estonian Academy of Sciences

Majak, Jüri
Department of Machinery, Tallinn University of Technology

Meidla, Tõnu
Institute of Ecology and Earth Sciences, University of Tartu

Mellikov, Enn
Department of Material Science, Tallinn University of Technology
Estonian Academy of Sciences

Merilai, Arne
Institute of Culture Studies and Arts, University of Tartu

Metspalu, Andres
Estonian Genome Center, University of Tartu
Estonian Academy of Sciences

Milani, Lili
Estonian Genome Center, University of Tartu

Mõtus, Leo
Estonian Academy of Sciences
Department of Computer Control, Tallinn University of Technology

Müntel, Mait
National Institute of Chemical Physics and Biophysics

Nelis, Mari
Estonian Genome Center, University of Tartu

Õim, Haldur
Institute of Estonian and General Linguistics, University of Tartu
Estonian Academy of Sciences

Öpik, Andres
Department of Materials Science, Tallinn University of Technology

Pajusalu, Karl
Institute of Estonian and General Linguistics, University of Tartu

Pärtel, Meelis
Institute of Ecology and Earth Sciences, University of Tartu

Pehk, Tõnis
National Institute of Chemical Physics and Biophysics

Perola, Markus
National Institute for Health and Welfare, University of Helsinki

Pohlak, Meelis
Department of Machinery, Tallinn University of Technology

Raidal, Martti
National Institute of Chemical Physics and Biophysics

Raik, Jaan
Department of Machinery, Tallinn University of Technology

Saar, Enn
Tartu Observatory
Estonian Academy of Sciences

Samel, Nigulas
Department of Chemistry, Tallinn University of Technology

Sammelselg, Väino
Institute of Physics, University of Tartu

Selart, Anti
Institute of History and Archaeology, University of Tartu

Soomere, Tarmo
Institute of Cybernetics at Tallinn University of Technology
Estonian Academy of Sciences

Tammesoo, Mari-Liis
Estonian Genome Center, University of Tartu

Ubar, Raimund
Department of Computer Engineering, Tallinn University of Technology
Estonian Academy of Sciences

Ukrainski, Kadri
Institute of Economics, University of Tartu

Vaikmäe, Rein
Institute of Geology at Tallinn University of Technology

Valk, Ülo
Institute of Culture Studies and Arts, University of Tartu

Varblane, Urmas
Institute of Business Administration, University of Tartu
Estonian Academy of Sciences

Vasar, Eero
Department of Physiology, University of Tartu
Estonian Academy of Sciences

Veinthal, Renno
Department of Materials Engineering, Tallinn University of Technology

Zobel, Martin
Institute of Ecology and Earth Sciences, University of Tartu
Estonian Academy of Sciences

THE EDITORS EXTEND THEIR THANKS

S. Jakobson, T. Kuusmann, A. Pihlak, Ü. Rebo