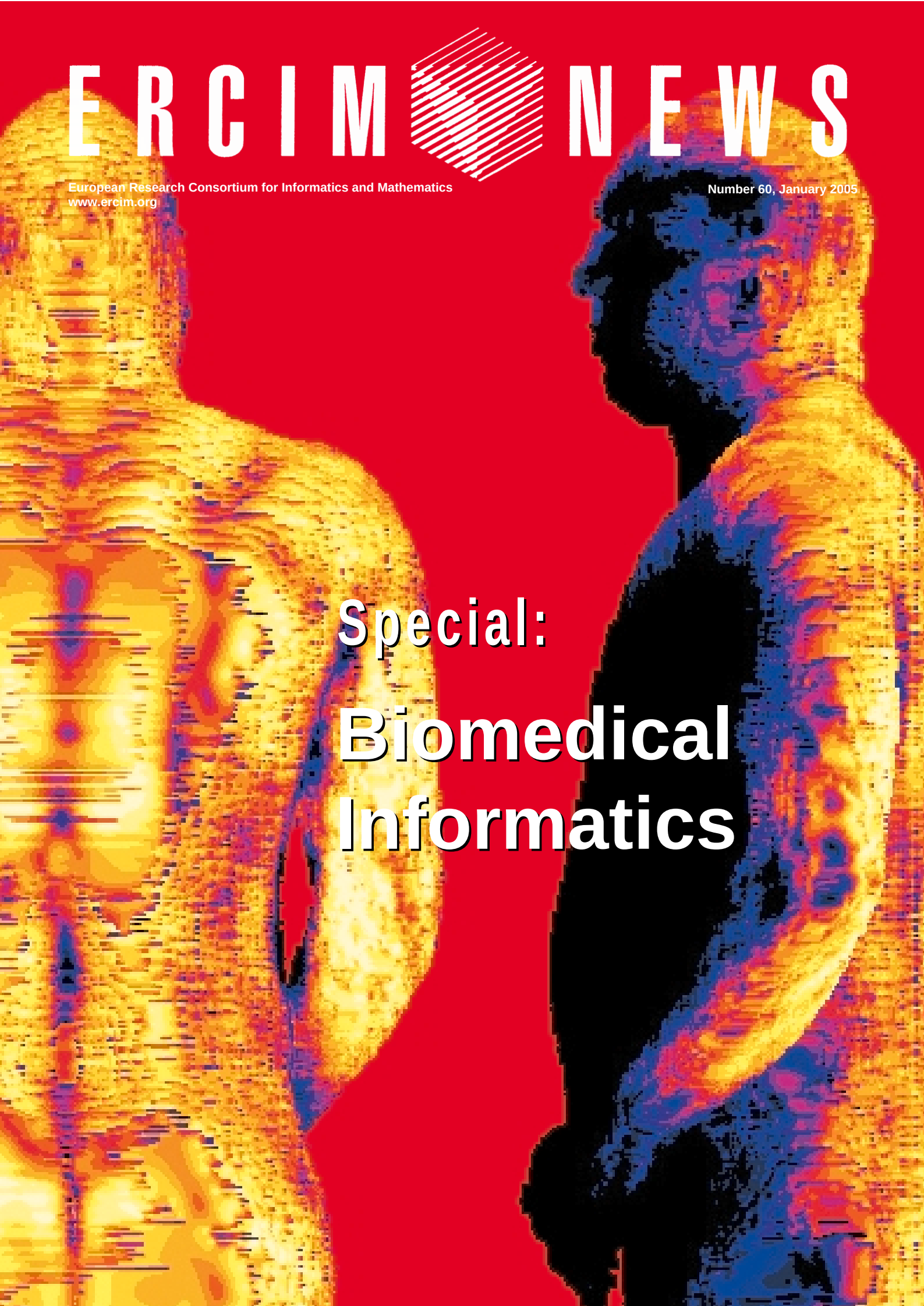
ERCIM NEWS

European Research Consortium for Informatics and Mathematics
www.ercim.org

Number 60, January 2005



Special:

Biomedical Informatics

CONTENTS

Next issue: April 2005
Special theme: Environmental Modelling

ERCIM invites applications for the position of

Manager

The Manager will be directly employed by the French National Institute for Research in Computer Science and Control (INRIA) and will be the representative of ERCIM vis-à-vis third parties. He/she will be responsible for:

- the implementation of ERCIM's general policy within the framework specified by the membership
- working in coordination with Executive Committee members on various activities
- participating in the Board of Directors and Executive Committee meetings
- assisting in selecting new members of the EEIG, as well as assimilating the new members into the consortium
- preparing the budget and the administrative and financial reporting to the consortium members
- the coordination of activities related to the management of European projects
- organising dissemination activities
- the management of approximately 25 employees from both the ERCIM Office and W3C Europe
- the consolidation and the launching of new scientific activities through working groups, EC projects, etc.
- participating in ERCIM task forces and in particular in the Strategic Task Group
- representing ERCIM to the European Commission, European or international scientific authorities, and W3C.

Qualifications

- PhD in computer science or mathematics or equivalent
- strong leadership skills, rigorous, assertive with a positive outlook and performance-oriented attitude; team-player and builder with excellent interpersonal skills; organisational and analytical skills; and high integrity.
- excellent verbal and written communication skills; must be fluent in English and French
- engaged in the European scientific environment and at ease liaising with the European Commission
- real capacity for administrative and financial management, with a good knowledge of management laws in private enterprise
- strong writing skills (scientific articles, employee and community communications, etc.)
- must be available for travel.

About ERCIM

The European Research Consortium for Informatics and Mathematics (ERCIM) is a European Economic Interest Grouping (EEIG) that comprises leading research institutes from eighteen European countries. ERCIM aims to foster collaborative work in the domain of information science technology (IST) and applied mathematics within the European research community and to increase co-operation with European industry. Activities managed by ERCIM include:

- sponsorship of events related to IST and applied mathematics (conferences, workshops and summer schools) in Europe
- organisation and supervision of strategic workshops bringing together representatives from the EU and the US National Science Foundation
- promotion of ERCIM working groups which are dedicated to specific scientific themes
- international fellowship programme for post-doctorates
- management of multi-national research projects funded in part by the European Commission.

The ERCIM administrative office is located in Sophia Antipolis (near Antibes) in France. The ERCIM Office serves as the front-end to the consortium where the budget is managed, financial and legal matters are addressed, projects are coordinated at a European level and communication channels are directed.
ERCIM is also the European host of the World Wide Web Consortium (W3C). W3C develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential.

Further Information
Daniele Eecke, INRIA, Tel: +33 1 3963 7808, E-mail: daniele.eecke@inria.fr
A detailed description will be available at: http://www.ercim.org/activity/jobs.htm

Biomedical Informatics: The Opportunity and Challenge for Multidisciplinary Research

Rosalie Zobel, Director of Directorate C: Miniaturisation, Embedded Systems, Societal Applications, Information Society Directorate-General, European Commission.

Biomedical Informatics (BMI) is a multidisciplinary field rising from the synergy of medical informatics, bioinformatics and neuroinformatics. The main mission of BMI is to provide a framework for developing, integrating and sharing biomedical knowledge related to human health from very different research disciplines such as genomics, proteomics, clinical research and epidemiology. The ultimate objectives of BMI are to support molecular medicine and personalised healthcare.

The advances in information and communication technologies (ICT), coupled with the increased knowledge about the human genome, have opened new perspectives for the study of complex diseases. There is a growing need to integrate and translate the knowledge about human genome into concrete benefits for all citizens such as more effective disease prevention mechanisms, individualised medicines and treatments and many other aspects of future citizen centred healthcare systems.

To carry out the work it is important to add value to the data that is stored in huge, publicly accessible research databases around the world generated by functional genomics and proteomics research by linking it with patient's clinical and genetic information that is stored in mostly smaller and secured clinical information databases and electronic health records. GRID technologies (see ERCIM News 59) constitute one of the promising tools in this direction. The use of GRID technologies and infrastructures in health sciences has been supported by the IST programme for several years now (HealthGrid).

BMI deals not only with the integration of health related data on different levels (molecular, cellular, tissue, organ, person and population) but also with computationally demanding tasks of data mining, modelling, simulation and visualisation. New in-silico modelling and simulation has the potential to accelerate new drug design and development, improve understanding of underlying biological processes, support predictive medicine as well as provide novel tools for training and surgery planning.

The Conference "Synergy between Research in Medical Informatics, Bioinformatics and Neuroinformatics", that took place in Brussels on December 14th, 2001 (http://14dec2001.ramit.be/index.html) marked the beginning of European Commission activities in BMI. A group of participants in this conference formulated the first R&D roadmap which was used as input for the formulation of activities for the 6th Framework Programme for research and technological development (FP6). While the first pilot BMI projects working on the aspects of genomic based medicine, such as INFOGENMED have been supported in FP5, more targeted calls have been issued in FP6 (http://www.cordis.lu/ist/directorate_c/ehealth/). You will find short descriptions of some of the current projects in this issue of the ERCIM News magazine.

Biomedical Informatics is one of the examples of new highly multidisciplinary ICT that lies at the crossroads with life sciences. The challenge for the future activities of the EU programme will be to create favourable environments for such multidisciplinary research and to accelerate the harvesting of societal and economic benefits.

Rosalie Zobel

Christof Teuscher Winner of the Cor Baayen Award 2004

Christof Teuscher from Switzerland was presented the 5,000 Euro Cor Baayen Award for his research on unconventional biologically-inspired machines during a ceremony at the ERCIM fall meetings in Malaga, Spain on 3 November 2004.

Well before getting his master's degree, Christof Teuscher participated very actively in some of his lab's research projects and published several papers. His promising graduation thesis on neural networks earned him several awards. During his PhD work at the Swiss Federal Institute of Technology, Christof Teuscher published 47 scientific papers. Based on the very promising and excellent work proposed in his PhD as well as on the recommendation of his PhD supervisor Prof. Daniel Mange from EPFL (as member of SARIT, the Swiss Association for Researchers in Information Technologies (http://www.sarit.ch) nominated Christof Teuscher as their candidate for the 2004 Cor Baayen Award .

ERCIM President Keith Jeffery (left) presents the Cor Baayen Award to Christof Teuscher.

In addition to his scientific research, Christof Teuscher was also exceptionally active in all kinds of other activities: for instance, he initiated and successfully organized several international workshops, such as the 5th International Workshop on Information Processing in Cells in Tissues, was publicity chair of the First International Workshop on Biologically Inspired Approaches for Advanced Information Technology. He also published a book on Turing's connectionist ideas and edited an impressive definitive collection of commemorative essays on Alan Turing's life and legacy.

His main research interests are mostly related to biologically inspired computation and machines, and cover a wide variety of interdisciplinary research topics like emergence, complexity, artificial intelligence, information theory, novel hardware and reconfigurable architectures, smart and ubiquitous computation and computational modelling of cognitive phenomena. The gradual process of developing and self-generating lifelike artefacts by which complex adaptive behaviours emerge from the interaction of simple processing elements was a central piece of his PhD work. In his thesis he also investigated inventive approaches which materialized within the 'Biowall' (http://lslwww.epfl.ch/biowall).

Christof Teuscher is currently a postdoctoral researcher at the University of California, San Diego (UCSD), Department of Cognitive Science, and with this award, ERCIM acknowledges the excellence of his scientific achievements and wishes him a very successful scientific career at the borders of technical and life sciences.

See also his article 'The Quest for Novel Computational Paradigms and Machines' on page 65.

Call for Candidates

Cor Baayen Award 2005

The Cor Baayen Award, awarded to a most promising young researcher in computer science and applied mathematics, was created in 1995 to honour the first ERCIM President, and is open to any young researcher having completed their PhD thesis in one of the 'ERCIM countries': Austria, Belgium, Czech Republic, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, Norway, Slovakia, Spain, Sweden, Switzerland, The Netherlands and the United Kingdom.

The award consists of a cheque for 5000 Euro together with an award certificate. The selected fellow will be invited to the ERCIM meetings in autumn 2005. A short article on the winner, together with the list of all candidates nominated, will be published in ERCIM News.

Rules for Nomination

Nominations from each country are made by the corresponding ERCIM Executive Committee member (also referred to as 'national contact'). Those who wish that a particular candidate be nominated should therefore contact the ERCIM Executive Committee member for their country (see http://www.ercim.org/contacts/execom) or the ERCIM office (office@ercim.org).

Nominees must have carried out their work in one of the 'ERCIM countries'. Nominees must have been awarded their PhD (or equivalent) not more than two years prior to the date of nomination. Each ERCIM institute is allowed to nominate up to two candidates from its country. A person can only be nominated once for the Cor Baayen Award. The selection of the Cor Baayen award is the responsibility of the ERCIM Executive Committee.

Submitting a Nomination

To submit a nomination, fill out the Cor Baayen Award nomination form (see http://www.ercim.org/cor-baayen.html) and provide a copy of the candidate's PhD thesis as well as copies of the candidate's best papers (max. 5), preferably provided as links to electronic documents.

Deadlines

- **Nominations are to be received by the national contacts by 15 April, 2005.**
- **National contacts are to send their two selected nominations to the coordinator by 30 April 2005.**

Further information can be obtained from your national contact or from the Cor Baayen Award coordinator Lubos Brim, Masaryk University Brno/CRCIM, (lubos.brim@ercim.org).

http://www.ercim.org/activity/cor-baayen.html

Keith Jeffery appointed new ERCIM President 'ERCIM Needs Closer Cooperation'

Keith Jeffery, director of IT for CCLRC took office as ERCIM president on 1st January. His main message: ERCIM members must work closer together, on the level of the working groups, but also in management. Keith believes that is the only way to play a role in the European knowledge society.

With the election of Keith, ERCIM is led by a geologist — not the most obvious choice for an ICT research consortium. His geology research, however, led Keith in an early stage to the subject of information systems. Since then he has worked on database systems, scientific computing and e-business. Keith has experience of all the major structures of ERCIM. He was a member of ERCIM's Executive Committee from 1994-1998 and a member of ERCIM's Board of Directors since 1998. He was also on the Director's strategy task force and he led the (now discontinued) Database Research Working Group from 1991.

Keith succeeds Stelios Orphanoudakis who had to resign for health reasons in October 2004. "Of course I am happy my collegues in the Board of Directors elected me as president," Keith says. "But I would have liked the circumstances to be different. I can speak for all directors when I wish Stelios a rapid recovery and return to ERCIM."

Organization

"I will carry on the main line of my predecessors," Keith responds when asked about his plans for ERCIM, "but with my own added flavour." During the November meeting in Malaga, just before his appointment, Keith already gave a small taste of that flavour. The Board of Directors agreed to his proposal to change ERCIM's management structure. The Board, supported by the Strategic Task Group, determines ERCIM's goals. This policy is then carried out by the Executive Committee with the help of four task groups, Projects, Finance, HR Management and Public Relations.

"My observation of ERCIM since 1991 was that the organization was somewhat fuzzy. Decision-making, priority-setting and resource management require clear lines of communication between organizational units with clear objectives and without significantly overlapping tasks. The new structure is clear and simple. It resembles the way a company is organized."

Lisbon Strategy

The new structure is necessary, Keith believes, for ERCIM to play a role in the realization of the Lisbon strategy (the European Council set out a ten-year strategy in March 2000 in Lisbon to make the EU the world's most dynamic and competitive economy). "Lisbon requires European cooperation on a scale and a depth hitherto unseen. ERCIM can and should be a leading player in its realization. However, currently we do not have this level of cooperation within our own organization."

Keith hopes the new organization paves the way for a common ERCIM strategy and more focused research. He confirms that this means more influence from ERCIM on its members' research activities. "However, no pressure or diktats are involved. I believe the members will realize the benefits of an ERCIM-wide approach. Of course not all institutes have to be involved in every ERCIM project. Depending on the individual institutional objectives members can decide to join."

Working Groups

Working groups will also be stimulated to expand their collaboration. "ERCIM has several successful working groups, but on the whole they are not doing as well as they could be. I would like to see each working group manage three projects, one of them together with another ERCIM working group. Furthermore, I am preparing a proposal to reserve funding to stimulate working groups to adopt more uniform procedures and develop PR material."

He realizes these ambitions depend on the willingness of the ERCIM members. "There has been a tendency for each institute to nurture its self-interests. It is my job to ensure everybody sees the advantages of belonging to ERCIM and participating to their full capacity. As a start it would be a good idea to try to gain an 'Integrated Project' from the EU, for instance on a combination of semantic web and GRIDs. That would definitely be an inspiring example."

Keith also wants to strengthen ties with ICT companies. "In my opinion ERCIM is oriented too much towards academic blue-sky research. I believe members can gain valuable experience and contacts by collaboration with commercial partners. This does not mean that commercial projects should dominate ERCIM research. My experience is every ICT problem has multiple aspects. ERCIM members can address predominantly basic research components, while companies deal with predominantly commercial components."

One of the issues that will undoubtedly be high on the agenda during Keith's presidency is the further expansion of ERCIM. According to Keith, the consortium will actively look for members in the new EU countries. "ERCIM was ahead of the EU in recruiting Central Europe. But since that time the Union has overtaken us. We are adressing this through the new Strategic Task Group."

Keith Jeffery was interviewed by Fedde van der Lijn, CWI, ERCIM local editor for The Netherlands.

ERCIM Working Group on Software Evolution

by Tom Mens

In November 2004, the ERCIM consortium approved a new Working Group on Software Evolution. The purpose of this ERCIM Working Group (WG) is to build and maintain a network of ERCIM researchers within the particular scientific field of software evolution. In this sense, the WG will serve as a natural successor to an existing Scientific Network called RELEASE that will continue to be financed by the European Science Foundation until autumn 2005.

The inaugural meeting of the WG was held on 2 October 2004 in the Universita Degla Studia in Rome. It was co-located with ICGT 2004, the International Conference on Graph Transformation. The goal of the meeting was twofold: from the point of view of research, it served as a workshop for researchers to present recent results from a broad range of activities that fit within the general topic of 'Software Evolution through Transformations' (for more information, see http://wwwcs.upb.de/cs/ag-engels/ag_engl/Segravis/Events/SETra04/); the second goal was to start up a new ERCIM Working Group on Software Evolution.

The meeting was very successful, with thirty participants from eleven European countries attending. Twelve participants were members of the ongoing European Science Foundation 'RELEASE' research network, and representatives were present from eight ERCIM institutes.

Following the success of the meeting, the ERCIM Board of Directors decided to approve the proposed Working Group as an official ERCIM WG. At the time of writing, members of over 25 different research groups in European Universities, fourteen of which are ERCIM partner institutes, have already expressed their interest in becoming members of the WG, and this number is growing. If you are interested in joining the WG, please consult our membership policy and follow the procedure explained on our website (see below).

Motivation

The scientific motivation for the WG is that numerous scientific studies of large-scale software systems have shown that the bulk of the total software-development cost is devoted to software maintenance. This is mainly because software systems need to evolve continually to cope with ever-changing software requirements. Unfortunately, existing techniques and tools for the support of software evolution have many limitations, including being (programming) language-dependent, not scalable, difficult to integrate with other tools and lacking in formal foundations.

The principal goal of the proposed WG is therefore to identify a set of formally founded techniques and associated tools to aid software developers with the common problems they encounter when evolving large and complex software systems.

The WG intends to address a diverse array of evolution problems. We will not only attack the technical aspects of software evolution (the how), but also try to understand and model the fundamental principles behind software evolution (the what and the why). Following is a tentative and inevitably incomplete list of topics to be addressed:

- specification or analysis of the evolution of software artefacts in a broad sense (including, but not limited to requirement specifications, architectures, designs, models, metamodels, programs, components, tests, documentation, bug reports, version control information, log files, release histories, language descriptions, APIs and protocols)
- re-engineering and reverse engineering
- software restructuring, refactoring and renovation
- model-driven engineering and model transformation
- co-evolution, consistency maintenance and inconsistency management
- impact analysis, effort estimation, cost prediction, evolution metrics
- traceability analysis and change propagation
- version control and configuration management
- run-time adaptation and dynamic reconfiguration
- family and product-line engineering
- methods, processes and tools for managing software evolution
- development of a formal theory of software and systems evolution.

Activities in 2005

As for all other ERCIM WGs, recurrent activities include proposing new European research projects, enhancing intra-European collaboration at both research and teaching levels, collaborating with other WGs, and organising scientific events (eg international workshops).

In 2005, the WG will organise the following activities:

- a symposium on Software Restructuring, 6 January 2005, Gent, Belgium (co-financed by FWO/FNRS, Belgium)
- the annual WG meeting, to be held in co-location with the international conference ICSM 2005 in Budapest, September 2005.

Other activities may be added to the calendar; an up-to-date list can be found on the WG website.

Link:
http://w3.umh.ac.be/evol/

Please contact:
Tom Mens, Institut d'Informatique,
Université de Mons-Hainaut/FNRS/FWO
Tel: +32 65 37 3453
E-mail: tom.mens@umh.ac.be

Beyond the Horizon — A New European Action coordinated by ERCIM

by Grigoris Antoniou

Beyond the Horizon (B-T-H) is new so-called 'coordinated action' supported by the Information Society Technologies/Future and Emerging Technologies programme (IST-FET) of the European Commission.

The purpose of this action is to provide input about IST-related emerging trends and strategic research areas that require support, through a well-organised, extensive and systematic consultation of the relevant research community throughout Europe, involving the main actors and experts in the related fields. This input will assist the IST/FET programme with establishing its funding priorities for its 'proactive scheme' under the forthcoming 7th Framework Programme.

The main aims of B-T-H are the following:

- identify advanced strategic areas and grand science and technology challenges related to information and communication technologies (ICT)
- discuss the scientific, commercial and social importance of these challenges
- draw basic research directions in ICT and related disciplines for addressing the above challenges
- design roadmaps for making advances in these areas with a timeframe of fifteen years
- identify new frontiers for ICT basic research, and boundaries between 'pure ICT' research and other disciplines
- identify the potential for cross-fertilization of research in disciplines involved in these areas
- establish communication and cooperation mechanisms within and beyond Europe in order to facilitate and support the formation and functioning of a related scientific community during the project lifecycle in a field characterised by rapid and continuous evolution.

Approach

The project will last 18 months and will be based on a number of thematic groups (expert panels) working in a parallel, yet coordinated fashion. Each thematic group will focus on a particular research area, and will deliver a staged roadmap for the particular area, depicting the research paths that are believed to be most promising for achieving substantial progress in the particular area. They will also analyse the potential scientific, industrial and societal impact of advances in its designated area.

The groups will be established at a dedicated opening workshop to be held during the first four months of the project. The intermediate findings of the groups will be discussed at a consolidation workshop to be held after the fist year. Subsequently, they will finalize their work, and the outcomes of the individual reports will be integrated into a final project report.

The operation of the thematic groups will be based on teleconferencing and on advanced on-line communication facilities rather than on physical meetings, to increase cost efficiency and to minimize time and place constraints that would be imposed on their members. An online community dedicated to each of the groups will be established, based on the deployment of an information portal and a collaborative work system. Technological support for Beyond-the-Horizon online communities will be provided through a suite of web-based tools that facilitate the sharing of information among a group of people, and collaboration towards the achievement of shared goals. The technological infrastructure will be designed to support authorization, virtual networking facilities, task allocation and monitoring, voting and survey tools, and navigation and query facilities in an easy-to-use environment.

Thematic Areas

The final structure of the project's work will be established at the Opening Workshop. However, a number of key areas have already been identified:

- pervasive computing, coordinated by Prof. Alois Ferscha, University of Linz, Austria
- nanoelectronics and nanotechnology, coordinated by Sir Harold Kroto and Prof. Kosmas Prassides, University of Sussex, UK
- security and trust, coordinated by Prof. Michel Riguidel, ENS Telecom Paris, France
- biomedical informatics, coordinated by Prof. Fernando Martin-Sanchez, Institute of Health Carlos III, Spain
- intelligent and cognitive systems, coordinated by Prof. Rolf Pfeifer, ETH, Zurich
- engineering and quality assurance of software-intensive systems, coordinated by Prof. Martin Wirsing, TU Munich, Germany.

Management

B-T-H is managed by Bruno le Dantec, ERCIM office. The scientific coordinator is Grigoris Antoniou, FORTH. The scientific steering committee consists of Stefan Jähnichen, Fraunhofer Gesellschaft, Keith Jeffery, CCLRC, Jean-Eric Pin, CNRS and Arne Sølvberg, NTNU. B-T-H commenced on 1 January 2005 and will last until 30 June 2006. The next important step is the opening workshop that will take place in April 2005.

Interested members of ERCIM institutes are asked to participate actively.

Link:
http://www.beyond-the-horizon.net.

Please contact:
Grigoris Antoniou, ICS-FORTH
Tel: +30 2810 391624
E-mail: antoniou@ics.forth.gr

Bruno Le Dantec, ERCIM office
Tel: +33 4 9238 5010
E-mail: bruno.le_dantec@ercim.org

Strategic IST-FET/NSF Workshop 'Engineering Software-Intensive Systems'

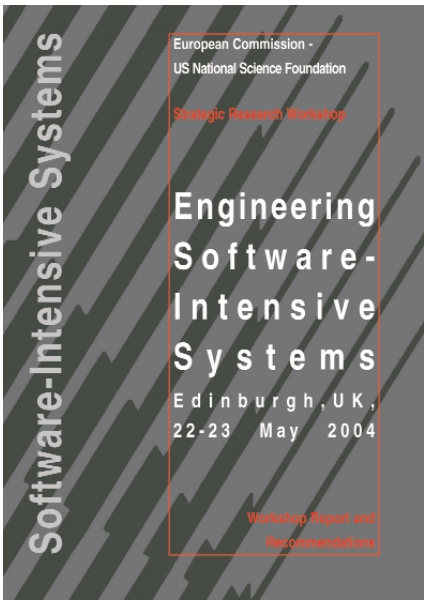
ERCIM has just published a report on the workshop 'Engineering Software-Intensive Systems' as part of the strategic workshop series under the auspices of the European Union (Information Society Technologies — Future and Emerging Technologies action) and the US National Science Foundation, 'Computer and Information Sciences and Engineering' division.

The strategic research workshop on 'Engineering Software-Intensive Systems' was organised in Edinburgh, Scotland, 23-24 May 2004, with the objective to present and discuss future R&D directions, challenges, and visions in the emerging area of software-intensive systems. About 20 leading experts from Europe, the United States and Australia participated in the workshop and identified research issues and challenges.

Software has become a key feature of a rapidly growing range of products and services from all sectors of economic activity. Software-intensive systems include large-scale heterogeneous systems, embedded systems for automotive applications, telecommunications, wireless ad hoc systems, business applications with an emphasis on web services etc. Our daily lives depend on complex software-intensive systems, from banking to communications to transportation to medicine. In the near future, software-intensive systems will exhibit adaptive and anticipatory behaviour; they will process knowledge and not only data, and change their structure dynamically.

Software-intensive systems will act as global computers in highly dynamic environments and will be based on and integrated with service-oriented and pervasive computing. However, actual practice shows that the techniques for engineering software-intensive systems suffer from many severe deficiencies in quality and methodological shortcomings:

- pragmatic modeling languages and techniques have no clean scientific foundations which inhibits the construction of powerful analysis and development tools



- formal approaches are not well-integrated with pragmatic methods and do not scale up to complex software-intensive systems
- aspects such as change, adaptation, heterogeneity, quality of service, security, trust, and highly dynamic and unpredictable environments, are important for software-intensive systems, but are not well supported by actual engineering methods.

Given the above, the grand challenge is to develop practically useful and theoretically well-founded principles, methods and tools for engineering high-quality software-intensive systems. Mastering the complexity of software-intensive systems requires a combined effort for foundational research and new engineering techniques that are based on mathematically well-founded theories and approaches. The new methods should support the whole system life cycle including requirements, design, implementation, maintenance, reconfiguration and adaptation. Research is required for:

- developing innovative engineering support for software-intensive systems to ensure required levels of quality and trust
- putting change and adaptation at all levels of system development
- developing a science of software-intensive systems
- bridging the gap between pragmatic development techniques and foundational validation and verification methods.

More information on this workshop, its main findings and the list of participants can be found at: http://www.ercim.org/EU-NSF/sis.pdf

The workshop is one of ten workshops organised or co-organised by ERCIM in Europe and the United States since 2001 to identify key research challenges and opportunities in information technologies. Other workshops covered the following topics:

- Bionics — Bio-Inspired Information Technologies
- Future Information Processing Technologies
- Semantic Web
- R&D Strategy for a Dependable Information Society
- Middleware for Mobile Systems
- Digital Human Ontologies
- Interdependencies
- The Disappearing Computer
- Unconventional Programming Paradigms.

Links:
EU-NSF strategic workshops:
http://www.ercim.org/EU-NSF

Workshop report
'Engineering Software-Intensive Systems':
http://www.ercim.org/EU-NSF/sis.pdf

Please contact:
Remi Ronchaud, ERCIM office,
Tel: +33 4 9238 5010
E-mail: remi.ronchaud@ercim.org

ERCIM assesses the Future of ICT

'Strategy for ICT in Europe' is the title of the report produced by ERCIM on the suggestion of European Commission officials. The report asseses the future of research and development in information and communication technologies (ICT) in Europe and is intended to serve as a basis for consultations on future framework programmes.

Some technologies that rely heavily on connectivity have already been widely adopted: the rapid development of mobile phones and their various features may be the most omnipresent example. But in the next 10 years, ERCIM foresees 'smart dust' that senses where you are and sends information about it to the network of your choice, virtual reality that allows you to hold meetings with co-workers around the globe, GRIDs with self-repairing, self-managing properties that allow the seamless networking of datastores and computers across Europe, and automated programs that check code for flaws and security problems as the code is being written.

This almost overwhelming landscape of opportunities raises the obvious question: how can European investment, both by government and industry, best be targeted so that Europe finds its niche in the information world? Existing abilities and limitations will be critical in deciding where best to put research money, the report states.

We have to consider both our ability to build the systems that are needed to create the good society and products that have the potential of financing the desired society, the ERCIM report says.

Universities are clearly key players in crafting this new landscape – and as such must provide their students with the best possible knowledge and tools. Graduates are critical in reducing the 'time gap' between the creation of new knowledge through research and the infiltration of new knowledge into industry, public administration and society at large. The report also sees the importance of forming research consortia in partnership with industry, where well-financed basic research helps fuel overall business growth.

The report looks at three aspects of future and developing ICT:

- user- and system-centric components, including connectedness, for architected application systems
- electronic, storage, computing and communication components to support these user and system centric components
- system development methods to construct components for user- and system-centric components, and applications for these components.

More information:
The full report is available in pdf from the ERCIM website:
http://www.ercim.org/publication/policy/ERCIM_IT_Strategy_2004.pdf

or in html at:
http://www.ntnu.no/ikt/ercim_summary/

PhD Fellowships in GRID Research

The CoreGRID Network of Excellence is starting its Fellowship Programme (FP) to allow postgraduate students to join their research groups. The fellowships are open to PhD holders from all over the world. The duration will be 14 to 18 months in two different CoreGRID partner institutes sharing a joint research agenda.

Open Positions

- **Scheduling algorithms for higher-order components on the GRID**
  Duration: 18 months
  Hosts: University of Muenster, Germany
  Technical University of Delft, The Netherlands
- **Task Flow-based GRID application programming model and runtime environments**
  Duration: 18 months
  Hosts: Vrije University Amsterdam, The Netherlands
  Politecnica of Catalunya, Spain
- **Object oriented environment for HPC applications on the GRID**
  Duration: 14 months
  Hosts: HES-SO (EIA-FR), Switzerland
  University of Pisa, Italy
- **Dynamic software components composition in GRID environments**
  Duration: 18 months duration
  Hosts: University of Westminster, UK
  INRIA, France.

Conditions

- candidates must be fluent in English
- an institute is not allowed to host a fellow of the same nationality or a fellow who was previously affiliated to this institute.

The Fellows will participate in regular in-house discussions and seminars and contribute to CoreGRID in order to integrate scientific activities and dissemination across Europe. Particular care will be paid to ensure that all candidates get equal opportunities and all partners are well aware of the necessity to promote gender issues during the recruitment procedure. Throughout and on completion of the fellowship programme, the network will provide support and guidance to the Fellow in his/her effort to obtain a position in related research activities.

Deadline for Applications

11 February 2005, 17:00 local time (Central European Time)

Application Form and Detailed Information

http://www.coregrid.net/jobs/

W3C10 Program Looked Back, Looked Forward

The World Wide Web Consortium (W3C) marked its tenth anniversary with a day-long symposium on 1 December in Boston, Massachusetts, USA. W3C10 brought together Web and Internet technical leaders from around the globe to both remember the W3C's origins and look to the future of the Web and W3C's role in it.

Ethernet inventor and Internet pioneer Bob Metcalfe was the event's master of ceremonies. The rich program included equal parts reflection and projection. Sessions covered the early days of the Web and W3C's emergence, through the commercial and social impacts of the Web on the world we now experience:

- the *How It All Started* session speakers were Jean-François Abramatic (ILOG), Tim Berners-Lee (W3C), Alan Kotok (W3C), Dave Raggett (W3C), David Singer (IBM), Al Vezza (Foretec) and Steve Zilles (W3C Advisory Board)
- the *Impact on Science and Industry* session, moderated by Michel Cosnard (INRIA and ERCIM), had the speakers Denis Lacroix (Amadeus e-Travel) and Teri Richman (National Association of Convenience Stores)
- the *Impact on Society and Culture* session was run by Daniel Weitzner (W3C) and Lee Rainie (Pew Internet & American Life Project).

Other sessions looked at the impact of the Web and of Web standards, with an eye towards new frontiers for Web technical development, and tensions that may require resolution:

Photo: Ralph R. Swick, W3C

Cutting the anniversary cake. From left to right: Tim Berners-Lee, Jim Bell, Bob Metcalfe, Michel Cosnard, Nobuo Saito and Steve Bratt.

- the *Web of Meaning* session speakers were Tim O'Reilly (O'Reilly Media) and Bill Ruh (Cisco Systems); the moderator was Eric Miller (W3C)
- the *Web on Everything* session was moderated by Rohit Khare (CommerceNet Labs) and featured Takeshi Natsuno (NTT DoCoMo), Balaji Prasad (EDS) and Philipp Hoschka (W3C) as speakers
- the *Web for Everyone* was moderated by David Berlind (ZDNet) with speakers Bill Gillis (Center to Bridge the Digital Divide) and George Kerscher (DAISY Consortium).

Finally, the day ended with reflections and projections, as well as a Q&A session conducted by both Bob Metcalfe and Tim Berners-Lee.

Links:
W3C10 speakers: http://www.w3.org/2004/09/W3C10-Speakers.html
W3C10 program: http://www.w3.org/2004/09/W3C10-Program.html
W3C10 slides: http://www.w3.org/2004/Talks/w3c10-Overview/

Future Web Work at W3C

Current work is expanding the reach of the Web to anyone (regardless of culture, abilities, etc.), anything (on devices ranging from powerful computers with high-definition displays to smaller palm devices, and from general purposes consumer products to specialized tools), anywhere (from high to low bandwidth environments), and via any mode of interaction (touch, pen, mouse, voice, assistive technologies, computer to computer). These new technologies will lead to new discoveries, expand commercial opportunities, increase benefits, and create and solve new challenges for humankind. As of 1 December 2004, these technologies have been identified as follows:

- Web Applications
- Semantic Web Services
- Multimodal Web
- Mobile Web
- Accessible Web
- Advancing International Reach
- Policy Aware Web
- Next Generation Privacy for Enterprises.

Link:
http://www.w3.org/2004/11/15-presskit-single.html

W3C Workshop on Semantic Web for Life Sciences

The W3C held a workshop on 'Semantic Web for Life Sciences' on 27-28 October 2004 at the Radisson Hotel in Cambridge, MA. The goal was to discuss emerging and future applications of Semantic Web for Life Sciences and to get feedback on what additional specification or coordination efforts might be necessary to support this area. Specifically, how can Semantic Web technologies help to manage the inherent complexity of modern life sciences research, enable disease understanding, and accelerate the development of new therapies for disease?

The workshop was oversubscribed at 115 attendees coming from all over the world and from all sectors of life sciences. The workshop program included seven panel discussions on specific topics related to the future of Semantic Web for Life Sciences, and a closing discussion about next steps. The position papers submitted by the workshop participants also provide further details on these issues and a list of attendees.

Tim Berners-Lee, W3C Director, led off the workshop with an introduction to the W3C Semantic Web technologies: of a web

of machine-processable information, evolutionary and decentralized. Ken Buetow, Director of the U.S. National Cancer Institute's Center for Bioinformatics, began the second day of the workshop with his efforts to build the Cancer Bioinformatics Grid (caBIG) and of the applications of semantics and standards in a decentralized research environment.

Panel discussions included the pharmaceutical industry perspective on Semantic Web (SW), the interactions between SW and scientific publishing, the difference between domain-specific ontologies and broad, 'bridging' vocabularies, Web services, the challenge of unique identifiers in the life sciences, chemistry and SW. Identified issues are aggregation and inferencing on complex scientific knowledge and data.

The workshop ended with a panel discussion of next steps for the W3C. There was broad consensus that the existing data integration approaches — intra-enterprise, cross-community, etc. — are in need of help. There was also broad consensus that Semantic Web represented a potential set of solutions that could ease some of the difficulties posed by the life sciences data and knowledge domain.

However, there was also consensus that work remains to be done, in particular on 'cross-domain' vocabularies describing such areas as data provenance, context, cross-reference, navigation, versioning and so forth. The goal of work on such vocabularies would be to stimulate cross-community data integration.

Finally, there was a discussion of an 'implementation' interest group — for working on issues of complex data integration, aggregation, query and visualization. This idea met with broad support.

Links:
Agenda: http://www.w3.org/2004/07/swls-agenda.html
Summary report: http://www.w3.org/2004/10/swls-workshop-report.html
Position papers: http://www.w3.org/2004/10/swls/pospapers.html

W3C Adds FAQ and Tutorial

- **HTML and XHTML FAQ**
  http://www.w3.org/MarkUp/2004/xhtml-faq
  More FAQs available at:
  http://www.w3.org/Consortium/W3C-FAQs

- **XML Events for HTML Authors**
  http://www.w3.org/MarkUp/2004/xmlevents-for-html-authors
  This document is a quick introduction to XML Events for HTML authors. XML Events is a method of catching events in markup languages that offers several advantages over the HTML onclick style of event handling.
  More Tutorials available at:
  http://www.w3.org/2002/03/tutorials

Consensus among Mobile and Web Technology Leaders at W3C Mobile Web Initiative Workshop

The World Wide Web Consortium (W3C) hosted a Mobile Web Initiative Workshop on 18-19 November in Barcelona, Spain, colocated with a OMA (Open Mobile Alliance) meeting. The goal of this Mobile Web initiative is to make Web access from a mobile device as simple, easy and convenient as Web access from a desktop device.

Potential of Mobile Devices on the Web not yet Realized

Being able to access the wealth of information available on the Web from a mobile device is valuable in many day-to-day situations, eg when checking timetables, looking for product information, checking e-mail, transferring money or accessing a corporate Internet while traveling. Mobile Web access is considered to be a key enabler for mobile Internet services. However, even though many of today's mobile phones include Web browsers, accessing the Web from a mobile device has not become as popular as expected. Users often find that their favorite Web sites are not accessible or not as easy to use on their mobile phone as on their desktop device. Content providers have difficulties building Web sites that work well on all types and configurations of mobile phones offering Web access.

The Major Players Were There!

Over 40 position papers have been submitted by mobile and Web technology leaders from around the globe. These were operators (Vodafone, NTT DoCoMo, Orange/France Telecom, T-Mobile, etc.), browser vendors (ACCESS, Openwave, Opera, Obigo/Teleca, etc.), content providers (BBC, MSN, Yahoo!, etc.), authoring tool vendors (Adobe, etc.), mobile software vendors (HP, Oracle, PalmSource, Sun, etc.) and handset manufacturers (Nokia, RIM, SonyEricsson, etc.).

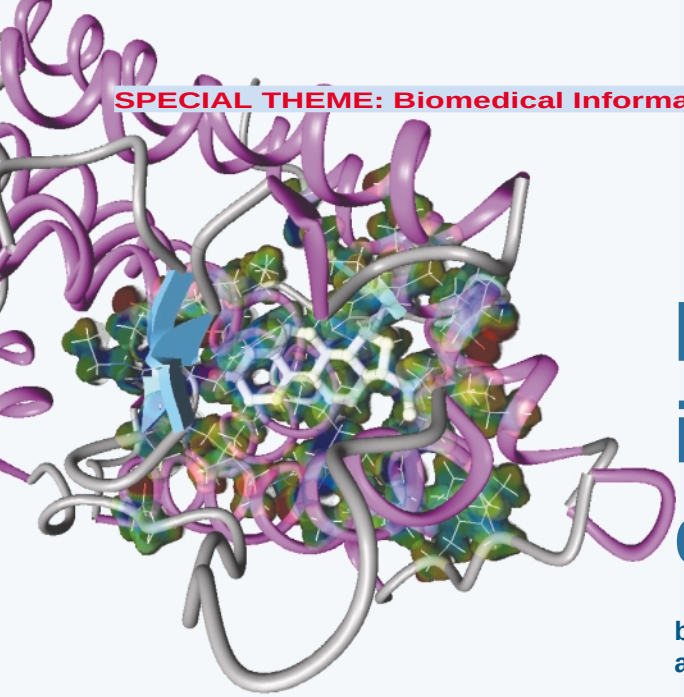
The workshop focussed on discussing the current challenges of 'mobile Web' access, and how to address them. Specifically, the participants discussed on developing of 'best practices' documents, providing support infrastructures for mobile developers, organizing training programs for Web content providers and creating validation and conformance testing services for Web-access from mobile devices.

Links:
Program: http://www.w3.org/2004/10/mwiws-program.html
Minutes: http://www.w3.org/2004/11/mwiws-minutes.html

Latest W3C Recommendations

- **15 December 2004: Architecture of the World Wide Web, Volume One;** Ian Jacobs and Norman Walsh
- **20 December 2004: XML Inclusions (XInclude) Version 1.0;** Jonathan Marsh and David Orchard

An exhaustive list of all W3C Technical Reports:
http://www.w3.org/TR/



Biomedical Informatics in Support of Individualized Medicine

by Stelios Orphanoudakis, Dimitris Kafetzopoulos and Manolis Tsiknakis

Historically, there have been few interactions between the research communities of medical informatics, medical imaging and bioinformatics. However, recent landmark achievements in genomics and the increased importance of genetics in health care are already changing the clinical landscape and are necessitating a highly interdisciplinary approach.

Recent and current developments include the following:

- A large number of genomes are now fully sequenced and public. The size of genomic databases has increased exponentially, now containing tens of higher organisms, hundreds of model and economically important species, thousands of microbial pathogens and almost all the important viral genomes. Comparative genomics allows the identification of conserved structural and regulatory elements within the genomes. Light has also been shed on the vast portion of the non-coding regions and 'gene deserts'.
- A large variety of proteins have been deduced from the various genome projects, and within them have been identified conserved or variant regions, functional and structural elements, features and domains. The continuum of life forms has become clearer and the differences between species measurable. Novel biocatalysts and the parameters relating structure to function have been identified from the diversity of living organisms, and the network of molecular interactions and complex biological processes has become available for modelling and 'in silico' experimentation.
- Gene expression profiles now allow clear identification, monitoring and classification of various organisms (eg pathogen strains), tissues and tumours, and states of health and disease. Profiling can highlight specific macromolecules and metabolic pathways (eg surface antigens) that could allow targeting of drugs or therapies.
- High-throughput screening of hundreds of targets is generating new functional coordinates within the chemical space. The primary tools for drug discovery are now the classification of chemical compounds and targets into functional groups, identification of relations between distant targets and drug effects, and knowledge visualization for chemical structures and properties. Advanced protein engineering via computer-aided design has proven a sophisticated aid in the development of new biocatalysts, therapeutics and diagnostic tools.
- Advanced methods (eg high-throughput crystallography, nuclear magnetic resonance (NMR)) have accelerated the resolution of new protein structures, and the modelling of macromolecules has been improved by new groups of 3D protein structures. Bioinformatics companies have developed information-integrating environments that allow computer-aided drug design and virtual screening for compounds.
- A variety of portable and distributed biosensors allowing simultaneous monitoring of several metabolites and biological signals have become widely available. In addition, molecular imaging techniques and other functional imaging methods such as positron emission tomography (PET) and functional magnetic resonance imaging (MRI) are assuming new and important roles in molecular-genetic imaging of cell metabolic states. This is used for the in vivo monitoring of protein interactions and gene expression.
- Functional genomics and genetic studies are elucidating the function of unknown genes, mostly by the use of holistic post-genomic approaches. The genetic determinants of multigenic diseases are being analysed and evaluated, and pharmacogenetics is identifying the genetic basis of drug efficiency and adverse effects. Pharmacogenomic information from clinical trials is generating the basis of the future 'targeted precise pharmacotherapy'; that is, the right drugs in the right doses to the right patient.
- Correlations between genotypes, gene regulatory networks and biochemical pathways now allow intervention and metabolic re-adjustments in combating complex diseases such as obesity, hypertension, hypercholesteraemia etc.

These developments and the increased importance of genetics in health care are already changing clinical care. Electronic genetic consulting is becoming common, and sequencing and genotyping are being established as laboratory routines in many health-care systems. Enterprises are also beginning to enrich the services they offer, providing, for example, analyses of health-related genomic information (ie subscription sequencing).

ARTICLES IN THIS SECTION

More importantly however, there are expectations that the new knowledge coming out of life science projects will change the world as much as or more than the Internet has, transforming the pharmaceutical and health-care industries and profoundly improving the practice of medicine. Since most individuals maintain unique genotype information, it is envisaged that they, or authorised health professionals, will in the future consult this information for their dietary choices, lifestyle and job placement decisions, prenatal diagnosis of suspected disorders and evaluation of possible disease symptoms and risks. Taken individually, classical epidemiological and clinical research and genomic research are no longer capable of advancing this so-called genomic medicine.

Genomic medicine integrates molecular medicine with individualized medicine; the former aims to explain life and disease in terms of the presence and regulation of molecular entities, and the latter applies genotypic knowledge to identify predisposition to disease and develop therapies adapted to the genotype of a patient. Needless to say, individual genotypic information, essential as it is to such approaches, must yet be the subject of extremely stringent security.

The exploitation of data from bioinformatics, medical informatics, medical imaging and clinical epidemiology requires a new and synergetic approach that enables a bi-directional dialogue between these scientific disciplines, and integration in terms of data, methods, technology, tools and applications. Biomedical Informatics (BMI) is an emerging discipline that aims to put these worlds together so that the discovery and creation of novel diagnostic and therapeutic methods is fostered. This will eventually herald a new era of what has become known as 'individualized medicine', whereby the drugs people are prescribed will depend on their personal genetic makeup, especially where there are significant costs or risk implications.

The mission of BMI is to provide the technical and scientific infrastructure and knowledge to allow evidence-based, individualized healthcare using all relevant sources of information. These sources include the 'classical' information as currently maintained in the health record, as well as new genomic, proteomic and other molecular-level information. BMI has the potential to improve the health and quality of life of the individual, as well as to reduce the overall costs of health-care systems, by enabling a shift from late-stage diagnosis to early detection or even prediction of disease.

To achieve this, a new breed of techniques, systems and software tools are required for two main reasons: to convert the enormous amount of data collected by geneticists and molecular biologists into information that physicians and other health-care providers can use for the delivery of care and the converse, and to codify and anonymize clinical phenotypic data for analysis by researchers.

Significant progress is also necessary in a number of domains, including:

- ontology-based integration of heterogeneous biological and clinical databases and the creation of a life-long active electronic health record for every citizen
- methods and tools for knowledge discovery, representation and visualization
- advanced computational methods in support of drug discovery, rational drug design, clinical trials and pharmacogenomics
- molecular and metabolic imaging methodologies in medicine
- simulations and modelling of molecular interactions, metabolic pathways, cells, tissues, and organs
- Grid-based approaches for demanding molecular-biomedical computational applications
- novel security-related methods and technology.

The goals of the current issue of ERCIM news, the first issue dedicated to biomedical informatics due to the relatively new nature of the field, are to show selected approaches and results from the research community of ERCIM, and to present the goals and objectives of some recently funded national and EU projects in this area.

Of the many articles submitted to this special theme, the 32 selected were most relevant to the domain of biomedical informatics. While some of the articles are still rooted in individual fields (eg medical informatics and bioinformatics), we believe they are relevant for the future multidisciplinary domain of BMI. Work reported in these articles can be divided into the following main areas:

European R&D and national projects: articles in this category present the main objectives of selected European and national projects.

Integration and analysis of biomedical data: a central issue in BMI is the integration of genetic information with the medical information contained in electronic health records or population databases in order to develop advanced prognostic or therapeutic tools for health professionals.

A number of articles address this area, with several focusing on Grid-based approaches to these demanding molecular-biomedical applications.

Data mining and visualization of biomedical data: the integration and exploitation of data from the disciplines of bioinformatics, medical informatics, medical imaging and clinical epidemiology require new methods and tools; several articles look at state-of-the-art technology in this area.

Simulation and modelling of biomedical processes: several articles deal with computationally demanding tasks of modelling and simulation, including modelling a living cell and developing 'in silico' virtual experiments.

Biomedical image analysis: several articles describe research in biomedical imaging, with an emphasis on molecular-genetic imaging. Some also relate to traditional medical image analysis, but provide a treasure chest of tools and methodologies that can easily be applied in the biological domain.

A couple of articles are also included which are related to chemoinformatics, and to the linking and integration of risk and environmental data. They are both relevant to future 'holistic' approaches to biomedical data and information management.

Finally, the newly formed ERCIM Working Group on Biomedical Informatics is introduced. The article describes the group's objectives, activities to date, and short-term plans in its attempt to create a highly interdisciplinary and distributed BMI research community from ERCIM organisations and other relevant stakeholders in Europe.

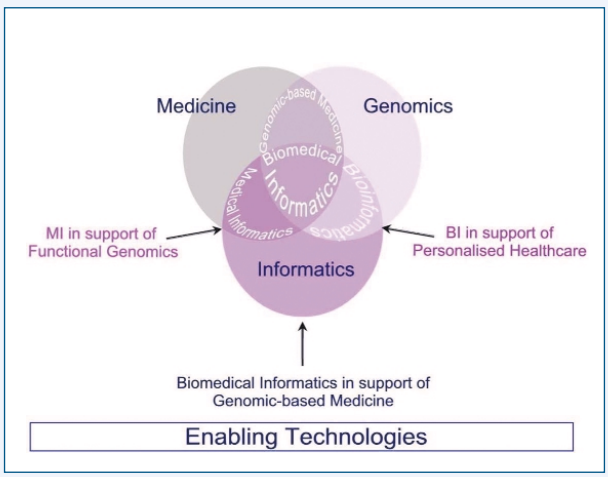
Please contact:
Manolis Tsiknakis or Dimitris Kafetzopoulos, FORTH
E-mail: tsiknaki@ics.forth.gr, kafetzo@imbb.forth.gr

ERCIM Working Group on Biomedical Informatics

by Manolis Tsiknakis and Dimitris Kafetzopoulos

On an initiative by scientists from FORTH, ERCIM's Board of Directors has established a Working Group on Biomedical Informatics.

Following the release of the Human Genome Project data, bioMedical informatics (BMI) has emerged as a scientific field worldwide, bringing together the disciplines of medical informatics, medical imaging, bioinformatics and possibly neuroinformatics in order to support individualized and proactive medicine in the post-genomic era. The mission of BMI is to provide the technical and scientific infrastructure and knowledge to allow evidence-based, individualized healthcare using all relevant sources of information.



Interdisciplinarity of BioMedical Informatics. Source: White Paper "Synergy between Medical Informatics and Bioinformatics: Facilitating Genomic Medicine for Future Healthcare" which resulted from the EC-IST 2001-35024 BIOINFOMED Study, available at http://bioinfomed.isciii.es/Bioinfomed/The White Paper/results/White Paper.pdf

BioMedical informatics aims not only to bring together these traditionally distant scientific disciplines but also to synthesize and exploit the whole spectrum of heath-related information, from molecules (ie molecular interactions, molecular imaging), to cells (neuron signalling, proteomic and transcriptomic profiles), to tissues, organs and organisms (cancer classification, pathogens), to individuals (integrated EHR, genotypes) and populations (epidemiology and public health records). BioMedical informatics touches on all basic and applied fields in biomedical science and is closely tied to modern information technology, notably in the areas of computing and communication.

Europe has demonstrated considerable interest, great momentum and substantial results in relevant fields over the years, by participating in most of the genome research projects, establishing outstanding facilities for bioinformatics (eg SwissProt and EBI) and by leading research and development in medical informatics through health telematics, medical imaging, health information networks and eHealth programs. Further, several groups in ERCIM institutions have oriented their interest in this multidisciplinary and emerging field of BioMedical Informatics, have played a crucial role in defining and guiding the R&D agenda in this area (see J. Biomed. Inform. (2004), 37:30-42) and have already achieved a high level of recognition. This momentum needs to be sustained and advanced.

The ERCIM BioMedical Informatics Working Group (WG) intends to promote interaction and collaboration between ERCIM R&D groups relevant to this area, and to facilitate cross-fertilization and synergies between distant scientific disciplines. The overall aim is to consolidate and advance this new field of research, enabling a better level of individualized health care in the post-genomic era.

Particular emphasis has been placed on understanding the link between genes, disease and the environment, and on the development of predictive models for diseases linked to genetic and environmental risk factors. This will allow appropriate preventive measures to be taken and progress to be monitored continuously. Researchers in the following fields could be interested and included in the activities of the WG:

- integration and analysis of genetic and medical information for health applications
- biomedical ontologies
- gene expression analysis (computational and experimental)
- genetic imaging
- modelling of genetic disorders and diseases
- Grid- based approaches to molecular-biomedical applications
- data mining and visualization of biomedical data
- computational methods and tools to support individualized medicine.

The Working Group also aims to promote awareness of promising basic and applied research results and their potential industrial adoption, in order to promote relevant standards and foster mobility of ERCIM researchers.

A number of ERCIM institutes and/or individual researchers from eleven European countries have actively participated in the initial meetings and discussions aiming at defining the scope and objectives of the WG or have expressed their interest in participating and contributing to future activities of the WG. The Working Group is open and we welcome new members. Potential collaborations and common activities have also been discussed with the recently formed Working Group of the

International Medical Informatics Association (IMIA).

The immediate plan of action for the WG includes:

- liaising with the EU, providing contributions towards the elaboration of the EU research agenda and priorities in the domain
- establishing R&D collaborations between the participating organizations (special efforts will be made in coordinating the submission of research proposals within the European Framework Programs)
- taking action in order to assist the mobility of (in particular young) researchers within the ERCIM BMI community
- organising a scientific workshop in the spring of 2005.

The WG has also been invited to submit a proposal for the organization of a satellite workshop during the VLDB international conference to be held in September 2005. The purpose of this would be to emphasize to the VLDB research community the significant challenges raised by biomedical informatics related to ontologies, semantic integration of heterogeneous and distributed resources, visualization of large and complex data sets, and so on.

The Web site for the Working Group will be available soon, with information and details of all activities, performed or planned.

Please contact:
Manolis Tsiknakis, FORTH, Greece
E-mail: tsiknaki@ics.forth.gr
Dimitris Kafetzopoulos, FORTH, Greece
E-mail: kafetzo@imbb.forth.gr

INFOBIOMED: A Joint European Effort to Support the Establishment of Biomedical Informatics

by Carlos Díaz

The European Commission funds a Network of Excellence to foster the development of Biomedical Informatics as a key integrative discipline for future healthcare.

Bioinformatics (BI) and Medical Informatics (MI) are disciplines that up to now have followed separate development with few contacts and synergies between them. The publication of the human genome has evidenced the need and the possibilities for a strong synergy between these two disciplines. The integration and exploitation of the data and information generated at all levels in both fields requires a new approach that enables a two-way dialogue between them that comprises data, methods, technologies, tools and applications. Biomedical Informatics (BMI) is the emerging area that aims to put these two worlds together. The mission of BMI is to provide the technical and scientific infrastructure and knowledge to allow evidence-based, individualised healthcare using all relevant sources of information. These sources include the "classical" information as currently maintained in the health record, as well as new genomic, proteomic and other molecular-level information. Aiming at a change from late stage diagnosis towards early detection or even prediction of disease, BMI bears the potential to foster discovery and creation of novel diagnostic and therapeutic methods, in order to improve the health and quality of life of the individual, as well as the efficiency of expenditure in healthcare systems.

With the objective of supporting the development of BMI in Europe, the INFOBIOMED Network of Excellence gathers 16 European organizations from 10 different countries (Belgium, Denmark, Germany, Great Britain, Greece, Italy, The Netherlands, Portugal, Spain, and Sweden). Funded by the European Commission for an initial period of 3 years, the Network brings together research groups with different backgrounds, creating a multidisciplinary team that provides an excellent framework to trigger the collaborative approach needed in order to enforce the establishment of BMI as a crucial scientific discipline for future healthcare. The network is composed by: Fundació IMIM, Institut Municipal d'Investigació Mèdica, Instituto de Salud Carlos III, Karolinska Institute, Edinburgh University, Custodix, Universidad Politécnica de Madrid, Universidade de Aveiro, Foundation for Research and Technology-Hellas, Danish Centre for Health Telematics, Informa S.r.l, Heinrich-Heine-Universität Düsseldorf, Erasmus MC-University Medical Center Rotterdam, Hvidovre Hospital, Academisch Centrum Tandheelkunde Amsterdam, and AstraZeneca Research and Development. The project is coordinated by the Research Unit on Biomedical Informatics (GRIB) of the Institut Municipal d'Investigació Mèdica (IMIM) in Barcelona, Spain.

The joint programme of activities that INFOBIOMED aims to implement in this initial phase has been designed to first, study all the significant aspects that are already relevant to Medical Informatics and Bioinformatics and that have the potential to provide a space for synergy between them. These aspects are included in two separate blocks of activities, one for data interoperability and management and the other for methods, technologies and tools. Each block is divided in several activities that reflect the main different areas that can require specific effort towards synergy: data characteristics and ontologies, data integration approaches, ethics and confidentiality, data mining

and information retrieval, image visualisation and analysis, and decision support systems. The first steps are directed to make a complete analysis of the state of the art and to identify the existing bottlenecks. Subsequently, the Network expects to tackle selected developments addressed to solve the detected gaps, so that obstacles to the realisation of BMI as a key discipline are reduced.

The INFOBIOMED project applies a vertical as well as a horizontal approach to Biomedical Informatics, so that a global perspective is possible. In this framework, all the knowledge gathered and created in the above-mentioned activities will be then tested into some 'vertical' pilot applications that aim to cover the whole range of information levels from molecule to population, from a practical perspective that works as testbed for the integrative approach pursued. These pilots will facilitate the analysis of the impact of BMI in key specific fields with the aim of investigating and knowing the requirements that these fields impose to BMI. The pilot applications also intend to create a bi-directional dialogue between BMI and other health-related disciplines, in order to prevent the isolation of scientists of different disciplines and to foster the creation of a solid, durable scientific community. The four areas covered by pilot applications are:

- pharmainformatics, which aims at investigating the impact of BMI at the different stages of the drug discovery process, from target identification to lead optimisation
- genomics and microbiology, focussed on the study of host and pathogen genetic polymorphisms, protein interactions and transcriptional/translational control and how these impact on microbial virulence and host immune responses to infection
- genomics and chronic inflammation, aimed at investigating the factors involved in susceptibility to adult periodontitis, as a model for complex diseases
- genomics and colon cancer, targeted at studying and improving the organization of screening in families with a high risk of developing colon cancer.

Link:
http://www.infobiomed.org

Please contact:
Carlos Díaz - INFOBIOMED Project Manager, IMIM, Spain
Tel: +34 93 2240302
E-mail: cmdiaz@imim.es

The BIOPATTERN Network of Excellence

by Emmanuel Ifeachor, Jo Thompson-Byrne and Michalis Zervakis

The 'Computational Intelligence for Biopattern Analysis in Support of eHealthcare' (BIOPATTERN) project is a Network of Excellence funded under FP6, Information Society Technologies Programme of the European Union.

The Grand Vision of the project is to develop a pan-European, coherent and intelligent analysis of a citizen's bioprofile; to make the analysis of this bioprofile remotely accessible to patients and clinicians; to exploit the synergy of information from different sources of medical and bioinformatics and the bioprofile to combat major diseases such as cancer and brain diseases.

The two key words in the project are *biopattern* and *bioprofile*. By biopattern, we mean basic information (pattern) that provides clues about the underlying clinical evidence necessary for diagnosis and treatment. Typically, it is derived from specific data types (eg genomic microarray, EEG and MRI) in a single medical investigation. A bioprofile is a personal dynamic 'fingerprint' that fuses together a person's current and past biohistory, biopatterns and prognosis. It combines not just data, but also analysis and predictions of future or likely susceptibility to diseases.

The focus of the project is to see how far we can go to realising the vision of a citizen's bioprofile; to identify barriers to the vision, to examine ways in which the bioprofile could be exploited for personalised healthcare, and to exploit synergies across boundaries, eg between bioinformatics and medical informatics, and between clinical areas. The project aims to identify how bioprofile could be exploited for individualised healthcare such as disease prevention, diagnosis and treatment.

The Grand Vision of the project opens up the possibilities of the development and support for future omnipresent and personalised active health support systems. Ultimately, a bioprofile would be constituted from data, information and analysis of a person's medical condition, involving sub-cellular genomic and proteomic information, intermediate scale phenomena such as EEG, MEG and ECG, and macroscopic scale patient records, demographic data, and body and brain imaging. This bioprofile could be dynamic and distributed across multicentres.

Roadmap

There are many barriers to the Grand Vision, including technological as well as ethical, security, and operability. BIOPATTERN integrates the research efforts of 31 institutions across Europe to tackle and reduce fragmentation in the new field of biopattern and bioprofile analysis which will underpin eHealthcare in the post genome era. It brings together leading researchers in medical informatics and bioinformatics from academia, the healthcare sector and industry in a new way, harnessing expertise and information to put Europe at the forefront of eHealth. The Network is set up to address, integrate and co-ordinate research efforts into the following generic themes:

- data acquisition
- analysis
- evaluation and bench marking

- e-delivery
- special interests areas - brain diseases, cancer and bioinformatics
- dissemination and exploitation
- management and co-ordination.

By organising the joint programme of activities into generic themes, this facilitates integration and ensures that our combined research efforts are channelled into the development and applications of techniques that have hitherto been undertaken in isolation. The specialist interest areas enable us to identify needs and priorities, to apply and evaluate new techniques, to ensure proper co-evolution of techniques and applications.

The idea of the Grand Vision is to move away from local solutions to local problems and towards European-wide solutions to European problems. Thus, the joint programme of activities include making information from distributed databases available in a secure way over the Internet, and providing on-line algorithms, libraries and processing facilities, eg for intelligent remote diagnosis and consultation. This will require the development of new and robust computational intelligence algorithms for biopattern analysis to support such facilities, including reference models of patients in the form of intelligent systems. Taken together, these will represent a unique on-line resource which can be used for remote diagnosis, decision support, trials and for medico-chemometrical research purposes. For example, for cancer diagnosis we would envisage a hospital (anywhere in Europe) would supply the necessary bio-profile, an appropriate set of algorithms are then used to analyse the bio-profile, on suitable machines which may be distributed, possibly utilising large distributed database(s) of examples, and the outcome of the analysis is then returned to the user.

The BIOPATTERN project is designed to reinforce European strengths in areas where it has established industrial and technology leadership (particularly true in the area of advanced biomedical data analysis), to overcome weaknesses in areas which are critical for European competitiveness and for addressing societal challenges (through the Vision of a citizen's bioprofile), to exploit new opportunities (through cross boundary networking) and respond to emerging needs, and to ensure the co-evolution of technology and applications.

Link:
http://www.biopattern.org

Please contact:
Jo Thompson-Byrne, Manager,
University of Plymouth, UK
E-mail: jbyrne@plymouth.ac.uk

Emmanuel Ifeachor, Project Coordinator,
University of Plymouth, UK
Tel: +44 1752 232574
E-mail: e.ifeachor@plymouth.ac.uk,

Michalis Zervakis,
Technical University of Crete, Greece,
Tel: +30 28210 37206
E-mail: michalis@danai.systems.tuc.gr

The INFOGENMED Project: A Biomedical Informatics Approach to Integrate Heterogeneous Biological and Clinical Information

by Ankica Babic, Victor Maojo, Fernando Martín-Sanchez, Miguel Santos and Antonio Sousa

Since the end of the 1990s, a growing number of researchers in Medical Informatics, Bioinformatics, Medicine and Biology are realizing of the advantages of linking genomic and medical data, knowledge, and methods to address new issues related to genomic medicine.

A conference carried out in Brussels, in 2001, supported by the European Commission, was the starting point for a series of European efforts to launch new initiatives in Biomedical Informatics. Since that event, many different approaches have been proposed with the aim of linking genomic and clinical data to support biomedical research and practice. One of these efforts has been the INFOGENMED project, carried out at the University of Aveiro, STAB VIDA (both Portugal), Linköping University (Sweden), and the Institute of Health Carlos III and Universidad Politecnica de Madrid (Spain). From 2002 to 2004, this project has been carried out and completed.

The INFOGENMED project has been designed to provide a unified access to multiple, heterogeneous biological and medical databases over Internet. The database integration module, named OntoFusion, is an ontology-based system designed for biomedical database integration. It is based on two processes: mapping and unification. Mapping is a process to link a database schema with a virtual schema. Virtual schemas are created using an existing ontology, such as UMLS or Gene Ontology or building a new domain ontology. In its current version, databases are mapped to virtual schemas at a conceptual level. Unification integrates ontologies and databases. Then, virtual schemas are unified, providing integrated access to the actual physical data. To our know-

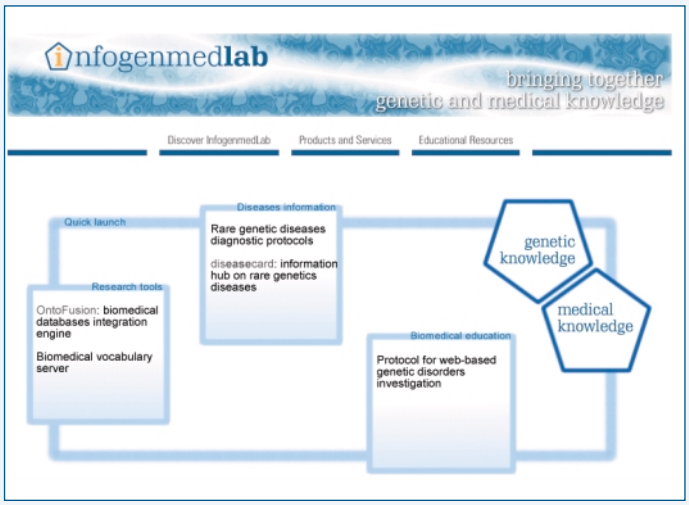


Figure1: Main screen of the INFOGENMEDLAB portal.

ledge, OntoFusion is the first database integration system that uses a high-level ontology description language to represent the virtual schemas. Our system incorporates tools to edit ontologies and to build the virtual schemas as well as a graphical ontology navigator.

OntoFusion is also capable to redesign database schemas. Using its mapping tool, physical database schemas can be modified. For instance, using OntoFusion, two different physical schemas can be mapped to a common virtual schema with two concepts. Unification is then automatic.

The system can integrate both private and public databases. There are currently over 500 biological databases (DBs) publicly available. These databases are the result of many biological research projects that have produced an enormous amount of data about genes, proteins and genetic diseases. Often, different public DBs include related types of data. In other cases, different organizations store their own information —eg, gene polymorphisms and mutations— but no integration of this disparate information is carried out. Many of these databases do not offer a direct connection and inquiries are made by means of Web forms. We have used the system to integrate a large number of public biomedical databases, such as OMIM, PubMed, Enzyme, Prosite and Prosite documentation, PDB, SNP, InterPro and others.

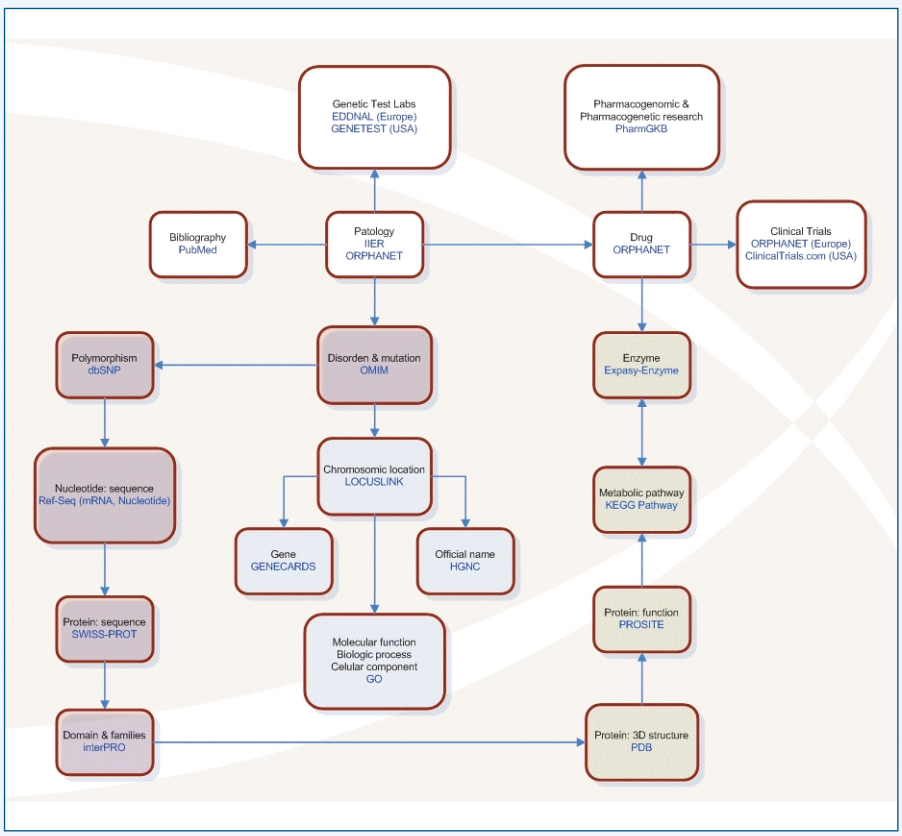


Figure 2: Map of the different public databases that can be accessed through the DiseaseCard module of INFOGENMED.

Another important result of the project has been the development of an assistant to help health practitioners to seamlessly navigate through local and remote Internet resources related to genetic diseases, from phenotype to genotype. A navigation protocol (a workflow for accessing public databases available on the web) was created by skilled users, familiar in retrieving information associated to rare diseases, both medical and genomic data. Based upon this protocol it was developed a web-based portal (DiseaseCard) that optimizes the execution of the information gathering tasks specified on the protocol.

One of the main future challenges of the system is to design the specific applications that can be useful to both biological and medical researchers and practitioners. Whereas, for instance, genomic researchers are common users of public Web-based databases, clinicians will need access to these kind of new information resources in order to fulfill the expectations of genomic medicine.

Links:
http://www.infogenmed.net
http://ribosome.ieeta.pt/infogenmedlab/

Please contact:
Antonio Sousa, University of Aveiro, Portugal
Tel: +35 1 234 370 500
E-Mail: asp@ieeta.pt

Victor Maojo,
Polytechnical University of Madrid, Spain
Tel: +34 91 336 7447
E-mail: vmaojo@fi.upm.es

Fernando Martin-Sanchez,
Institute of Health Carlos III, Madrid, Spain
Tel: +34 91 822 3219
E-mail: fmartin@isciii.es

SemanticMining – A Network of Excellence in the Field of Biomedical Informatics

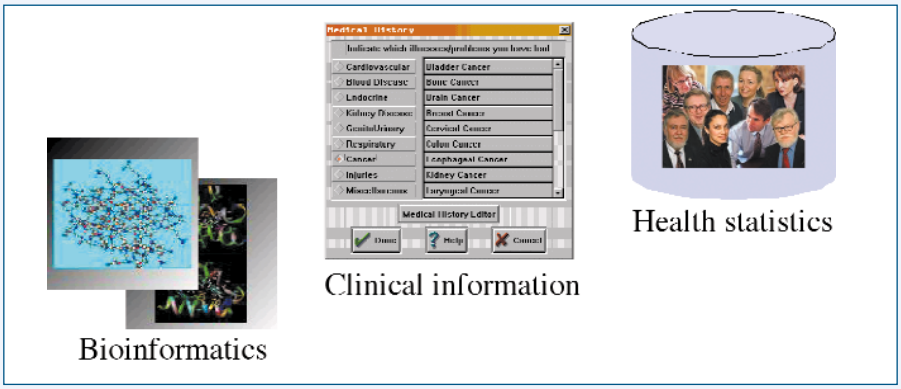
by Hans Åhlfeldt

The objective of the Network of Excellence entitled Semantic Interoperability and Data Mining in Biomedicine [SemanticMning] funded by the European Sixth Framework Programme, is to establish Europe as the international scientific leader in medical and biomedical informatics.

The long-term goal of the network will be the development of generic methods and tools supporting the critical tasks of the field; data mining, knowledge discovery, knowledge representation, abstraction and indexing of information, semantic-based information retrieval in a complex and high-dimensional information space, and knowledge-based adaptive systems for provision of decision support for dissemination of evidence based medicine.

The general objective of a Network of Excellence (NoE) is to bridge gaps in the European research infrastructure and to facilitate cross-fertilisation between scientific disciplines. Traditionally academic departments in the domain have their roots either in computer science, system engineering (including a variety of engineering disciplines) or in a medical or clinical context. The proposed network is composed of partners from these scientific areas, all bringing their experience and in-depths knowledge together into a common framework. An important aspect of this is the merging of medical or clinical informatics and bioinformatics including the new fields of genomics and proteomics.

Another bridging activity addressed by this NoE is knowledge transfer and co-operation between academia and organisations and SMEs in the health and welfare sector, including standardisation bodies and the different public and private institutions involved in health care delivery and management. The national institutes and organisations responsible for policy making and quality management with a regulatory and normative function will have an important role to play in the network. We believe that co-operation between these organisations and those involved in research departments needs to be strengthen, both in the early phase of research programme identification and in the later phases of implementation and large-scale evaluation of results and impact. The bridging activities between different levels of the health care system are exemplified in the figure.



The network addresses research issues on three levels; on the pre-clinical level of bioinformatics (functional genomics, proteomics etc.), on the clinical level of primary and secondary health care (hospital information systems, electronic health records etc.), and on the level of health statistics (population-based statistics, epidemiological surveying etc.). The identified areas of research share the basic problem of semantic interoperability, which means that semantics is preserved in communication between users and information systems.

The research activities in Semantic Mining is focused around seven areas:

- principles in ontology engineering
- evaluation of SNOMED CT
- impact of ontologies on health statistics
- concept systems in laboratory medicine
- the construction of a multi-lingual medical dictionary
- text mining and information retrieval in bioinformatics
- the concept-based electronic health record.

Researchers in the network play an influential role in the process of harmonisation and further development of terminology systems. Examples of areas of interaction are the Gene Ontology, the Foundational Model of Anatomy, and SNOMED CT. Part of the network objectives is also an active interaction with standardisation bodies such as CEN TC251, IMIA and W3C. The research carried out under the auspices of this NoE will also address the need for approaches in Europe which will bridge language barriers and facilitate access for non-English native persons to the large scientific corpus of texts written in English.

SemanticMining is based on the partnership of 25 partners from 11 European countries with approximately 100 identified researchers (25 female) and 35 associated PhD students (10 female).

Link:
http://www.semanticmining.org

Please contact:
Hans Åhlfeldt, Coordinator,
Linköping University, Sweden
Tel: +46 13 227574
E-mail: hans.ahlfeldt@imt.liu.se

Major New Biotech Initiative to Strengthen Sweden's Uppsala Region

by Rhiannon Sanders

With 100 million Swedish crowns (ca. 11 million EUR) from the government's industrial innovation agency (Vinnova) and an equivalent amount from the local biotech industry, the two universities and the municipality, Uppsala BIO is powerful new initiative for promoting the long-term growth of biotechnology in a region that already has an uncommonly good track-record in the field. Close academic/industrial collaboration is a cornerstone of the initiative, and many scientific disciplines — including bioinformatics — are brought together.

The Uppsala Region

Uppsala, located about 40 minutes from Stockholm and 20 minutes from Stockholm-Arlanda international airport, is the fourth largest city in Sweden. Its steadily increasing population is now around 180,000. With a large proportion of Sweden's biotech sector in the shape of universities, industry and national agencies, the region is today a buoyant centre of biotechnology.

What characterizes Biotechnology in the Uppsala Region Today?

The trend was set in the early 1920s when Professor Theodor Svedberg of Uppsala University's Department of Physical Chemistry constructed the world's first ultracentrifuge and used it to show that the molecules of certain pure proteins are all of one size. When Svedberg's research assistant Arne Tiselius obtained his doctor's degree in 1930 with the thesis 'The moving-boundary method of studying the electrophoresis of proteins', this trend was confirmed (Svedberg was awarded the Nobel Prize in Chemistry in 1926, Tiselius in 1948.). Uppsala was to become a world-leader in developing tools for life science research — tools characterized by close collaboration between university and industry.

The patents and product development that led to the 1959 introduction of gel filtration as a technique for separating biomolecules resulted from direct consultations between Tiselius and nearby Pharmacia.

Today, that chromatography business is a vital part of GE Healthcare, the world's biggest biotech/diagnostic company. About 8% of the region's workforce is employed in biotech-related businesses. In addition, more than half of Sweden's biotech firms are located in the 70-kilometer corridor that extends between Uppsala and Stockholm.

Leading-Edge University Research

That the region has attained the dominant position it enjoys in biotechnology today is not due to the commercial sector alone, however.

Many of Uppsala's university campuses conduct leading-edge research in topics of great biotech interest; the Biomedical Center (pharmacology, bioinformatics), the Linnaeus Centre for Bioinformatics, the Ångström Laboratory (materials science, MEMS), the Rudbeck Laboratory (molecular genetics and medicine), the University Hospital (cancer, clinical research and trials) and the Genetics Center (plant biology and forest genetics), for example.

Linnaeus Centre for Bioinformatics

The Linnaeus Centre for Bioinformatics (LCB) is a good example of how Uppsala's two universities plan to maintain their position at the forefront of world research.

LCB is a recently-established joint initiative between Uppsala University and the Swedish University of Agricultural Sciences. The group aims to carry out cutting-edge research ranging from microbial and mammalian genomics via computational functional genomics to molecular evolution. This unique research platform, which also has access to Sweden's National Supercomputer Center, combines prominent bioinformatic research disciplines such as biology, computer sciences and mathematics.

A Data Warehouse (DWH) for storing, managing and analyzing gene expression data is the latest LCB initiative. The DWH is a microarray-experiment oriented warehouse for collections of expression data, integrated with gene annotation profiling, used to support genomic data mining. Figure 1 shows an example of a typical result.

A Culture of Collaboration

The interdisciplinary cooperation shown by Uppsala's two universities in setting up the LCB is typical for the region's actors on the biotech scene.

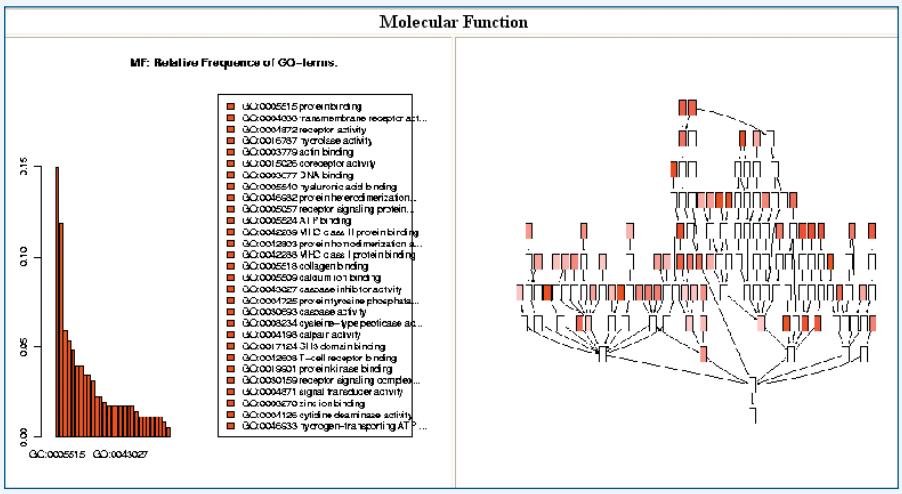


Figure 1: Images from the newly established Linnaeus Centre for Bioinformatics' data warehouse showing the visualization of gene function with the help of Gene Ontology.

Swedish industry and academia have always recognized the importance of working together, and have enjoyed much success in doing so. In Uppsala, the culture of collaboration is probably stronger than in the country as a whole.

Both universities have an open, international attitude, a strong life science history, and world-class biotech research. Industry provides innovative ideas and proven business skills. Put the two together in a tightly-knit region with strong networks and close personal ties and you have created a hotbed of interaction and growth.

Vision – One of the World's Top Five Biotech Regions

With such a strong biotech base today, one might wonder why Uppsala BIO is needed. The answer can be found in the initiative's vision: 'Uppsala is to be one of the world's top five biotech regions. Characterised by frontline biotech research, an outstanding innovation climate and close academic/industrial collaboration. Backed up with seasoned serial entrepreneurs and a full industrial/business infrastructure.' To achieve this aim, all parties involved have agreed on a number of strategic areas.

Figure 2: Uppsala's shopping malls were invaded by the region's universities, biotech companies and related bodies with information, competitions and hands-on experiments to attract young people to biotechnology.

Attracting more young people to science in general and biotechnology in particular is one challenge. Hence, on 20 November, Uppsala's shopping malls were invaded by the region's universities, biotech companies and related bodies with information, competitions and hands-on experiments for the local population to enjoy.

Another more mainstream approach is Uppsala BIO-X, a cross-disciplinary research effort focused on 'Tools for Life Science'. Uppsala BIO-X supports ambitious, world-class research by making supplementary funding and resources available, primarily for stimulating the formation of multi-disciplinary research teams.

Applications, which should have the potential to generate new opportunities for the life science industry, are welcome.

Links:
Uppsala BIO: http://www.uppsalabio.com

The Linnaeus Centre for Bioinformatics: http://www.lcb.uu.se

Please contact:
Rhiannon Sanders,
Uppsala University, Sweden
E-mail: Rhiannon.sanders@uppsalabio.com

CLEF: Joining up Healthcare and Biomedical Research

by Aniko Zagon, Adel Taweel and Alan Rector

CLEF (Clinical e-science framework) is a Medical Research Council sponsored project in the UK's e-science programme. CLEF is in its third year and is about to enter its second phase of development when it will begin translation of research into applications. CLEF aims to establish methodologies and a technical infrastructure for the next generation of integrated clinical and bioscience research.

The context for CLEF's work is provided by the current worldwide trend to create a single electronic healthcare record for each patient, eventually enabling a coherent medical care. This vision is currently being translated into reality in the UK in the government's 10-year close to £ 9 billion pounds National Programme for IT (NPfIT).

Information on the long-term course of patients' illnesses and treatments is needed both to improve clinical care and to enable post genomic research. CLEF's task is to develop a safe, generic, high quality, interoperable and pseudonymised information repository, derived from Electronic Healthcare Record (EHR) systems that can be queried just like any other GRID resource by healthcare and bioscience researchers without endangering operational security or data ethics.

Barriers to Improved Clinical Information

CLEF is developing methods for managing and using pseudonymised repositories of the long-term patient histories which can be linked to genetic genomic information or used to support customised patient care. It will enable ethical and user-friendly access to the information in support of clinical care and biomedical research.

CLEF's primary contribution will be removing key barriers to managing healthcare data repositories:

- *Compliance with privacy, consent, and security issues* at any time and at

all levels; policy, organisational structure, and technical implementation.

- *Comprehensive and intelligent Information capture*. Most of the clinical information is in the form of free text and not structured and systematic records. To address this problem, CLEF needs to design a 'context library' for data, which can be safely derived from existing notes to help to improve data utilization for researchers.
- *Information integration*. In their raw form, clinical records consist of hundreds of test results, medication and appointment notes. To make this data useful for clinical and bioscience researchers, a coherent 'chronicle' of events must be inferred from the records that summarises the key events from a single patients treatment records. This is a complex problem when the data concerned are clinical treatment records that may considerably vary from one patient to another...
- *Analysis and presentation of information for clinical and other scientific researchers*. The CLEF repository will be used by clinicians, medical researchers and a variety of other scientists, who will not be IT specialists.The questions that will put to the CLEF repository will be in a range of contexts and may require information from many sources from the GRID framework. The questions to be asked hence are difficult to preempt and creation of a query that will work to expectation can only be developed in collaboration with specialist groups of user beta testers.
- *Knowledge resources*. All of the above tasks require contextual data mining and recognition of implied meanings of the information. Since the number of knowledge resources is mushrooming, their coordination and management at the level of information integration is another complex task.
- *Standards*. Cooperation requires standards, which are only just emerging. Coordination of information gathering is also a serious security and ethical issue when medical data are involved and requires that data protection and system management guidelines are not only established but are also effectively enforced. Contribution in this international work is an integral part of CLEF.

Technologies in CLEF Solutions

Information Extraction from multiple texts. A typical lifetime record consist of 100- 200 text documents and even more laboratory, pharmacy or other structured data items. To improve the precision of information extraction, all available documents are used and cross-referenced during extraction.

At present, CLEF is using records of deceased patients where data confidentiality issues do not apply but the quality and complexity of the data are the same as in any other medical records and hence enable technological and regulatory/standardization processes to develop in parallel.

Information integration into standard healthcare record formats. CLEF draws on the work of OpenEHR which uses the new CEN standard for information interchange. CLEF's repository is built from standard 'archetypes' (reusable elements which facilitate interoperability and which can evolve) that were also adopted by HL7, the major standardization body in healthcare informatics.

'Chronicalisation'. The CLEF chronicle is an attempt to form a coherent view of the best inference about the course and choice of treatments in any single patient. Creation of a chronicle from 'index events' and their occurrences are extremely important source of information for those who study disease and treatment ontologies for research purposes. This is an extremely difficult transformation and thus the 'chronicle' will come into focus gradually as understanding improves.

Query formulation, WYSIWYM and Language Generation. For the CLEF repository to be useful to scientists and clinicians, it must contain data that are easily understandable to the majority of envisaged users.The interface to the repository of health records and chronicles is being designed around techniques from WYSIWYM –'What you see is what you meant' supplemented by various visual or graphical presentations.The next stage of the project will include user studies to ensure that the interface meets users' priorities.

The overall approach in CLEF is based on 'ontology anchored knowledge bases'. Some of the required information exists in established resources, such as the UMLS3, however, much of it needs to be compiled as CLEF repository develops. CLEF works with both myGrid and the new COODE project to develop usable/accessible knowledge resources and tools.

Metadata in the Repository. The CLEF repository requires at least four types of metadata:

- resource information
- provenance information
- usage and workflow information
- annotations on certainty and evidence.

While 1-3 are analogous to metadata within *myGrid* (see next article) and related projects, the 4th is more specific to CLEF.

CLEF is anchored in five prominent UK universities – UCL and the Universities of Manchester, Sheffield, Brighton and Cambridge. Since CLEF is working alongside an ambitious government project (NPfIT) it needs to develop a close an interactive relationship with the key industry initiatives to ensure that its research focus in closely aligned with the clinical system developments. CLEF also aims to provide research and development work for NPfIT using its technological expertise in both addressing healthcare informatics problem and interpreting clinical problems in the context of informatics.

From CLEF to CLEF Services

The aim of CLEF Services, which will be launched in January 2005, is to expand the clinical base for its developmental work. That will enable CLEF to begin to work with live data sets and step-up it's testing processes for user friendliness and faithfulness of data extraction and interpretation in face of complex queries.

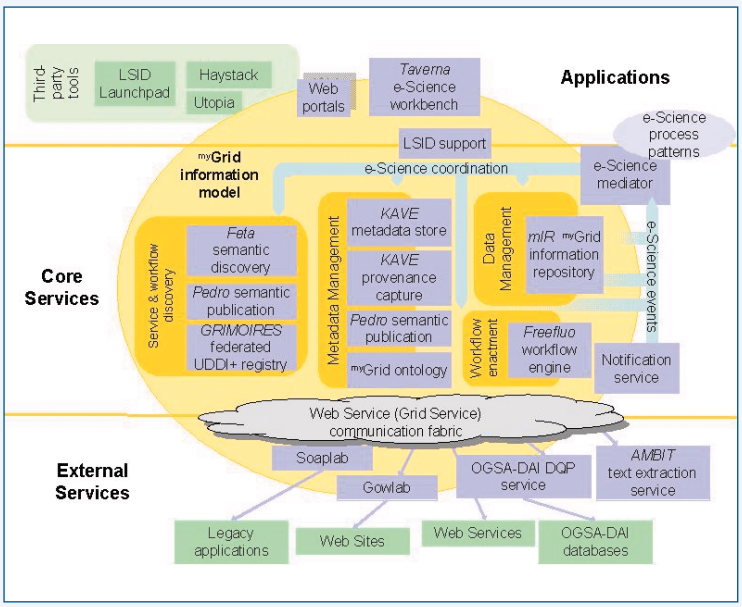
CLEF Services will also extend CLEF's ethical & security work with new partners, such as the Cathie March Centre at the University of Manchester and build closer links with *myGrid*, the GRID infrastructure and the NHS Care Record Service.

Please contact:
Adel Taweel, The University of Manchester, UK
Tel:+44 161 275 0659
E-mail: a.taweel@manchester.ac.uk

myGrid: Middleware for In Silico Experiments in Biology

by Carole Goble

Life science researchers traditionally chain together database searches and analytical tools, using complex scripts to overcome incompatibilities, or by manually cutting and pasting between web interfaces. These 'in silico' experiments are usually undertaken without support for the scientific process of managing, sharing and reusing the results, their provenance, and the methods used to generate them. The myGrid project has developed a comprehensive loosely-coupled suite of middleware components specifically to support data intensive in silico experiments in biology. Workflows and query specifications link together third party and local resources using web service protocols. The software can be freely downloaded and has been used for building discovery workflows by Williams-Beuren Syndrome and Grave's Disease Life Scientists collaborating with myGrid.



A screenshot of the Taverna interface showing a Williams-Beuren Syndrome workflow.

Bioinformaticians are knowledge workers, intelligently weaving together information globally available to the community, linking and correlating it meaningfully, and generating even more information. However, they are commonly forced to waste their time overcoming incompatible interfaces to applications and neglect basic scientific practice such as keeping records of the methods they used to integrate applications, how results link together, and even the goal of the experiment. The omission of these provenance records makes in silico experiments difficult to interpret and hard to repeat, and the methods difficult to reuse. This is not because bioinformaticians are unskilled but because the appropriate infrastructure to support them in their tasks are missing. The myGrid project has researched and developed open source high-level service-based middleware to support in silico experiments in biology, using workflows and semantic technologies.

The project, which has been running since late 2001, is a UK EPSRC-funded e-Science pilot project made up of a consortium of the Universities of Manchester, Newcastle, Nottingham, Sheffield and Southampton, IT Innovation, and the European Bioinformatics Institute, supported by nine industrial partners of whom GSK and IBM are the most significant. A third of the team are bioinformaticians with a Life Science background. The aim is to develop generic middleware driven by the real day to day problems of bioinformaticians, so the project has been firmly guided by strategic partnerships with UK-based Life Scientists in Newcastle-upon-Tyne investigating Grave's Disease, in Manchester investigating Williams-Beuren Syndrome and, more recently, in Liverpool investigating Trypanosomiasis in cattle. Third party developers of the EMBOSS, BioMOBY and SeqHound bio-service suites also use the middleware as a means of accessing and integrating their services. Currently myGrid supports access to over 600 public services.

The middleware is a toolkit of core components for forming, executing, managing and sharing discovery experiments. The components are intended to be adopted in a 'pick and mix' way by developers and tool builders to produce end applications. Bioinformaticians and service providers develop and run experiments via the Taverna workbench whereas biologists may use our configurable web portal to run them, examine results, and collaborate. Workflows that execute remote or local web services and Java applications are the chief mechanism for forming experiments. Legacy applications are incorporated using our Soaplab-Gowlab wrapper tools. Any web service can be incorporated – there is no restriction on the type of biology. The Taverna workbench is a GUI used for assembling, adapting and running workflows enacted by the Freefluo workflow enactment engine (see Figure 1). The project's Scufl workflow language is a user-oriented abstraction over general graph languages hiding the details of service invocation and control flow. Users can also configure support for fault management and service failover. In addition to workflows, databases may be integrated using the OGSA-Distributed Query Processor developed jointly with the UK OGSA-DAI project.

To support the scientific method the project has adopted a number of innovative technologies from the semantic web community. To enable service providers and bioinformaticians to publish, discover and match-make services, and bioinformaticians to publish and reuse workflows, our registry has mechanisms to support descriptions drawn from an ontology defined in the W3C standards

RDF (Resources Description Framework) and OWL (Web Ontology Language). Our Knowledge Annotation and Verification of Experiments component (KAVE) captures and stores provenance records of methods and purpose in RDF, and again semantically annotated by terms from an ontology.

The consortium has followed a twin-track of core-development and research-prototyping. Research efforts in semantic services and metadata management are currently being migrated to production as part of the core development. Through the use of standards such as Web services, RDF, OWL and Life Science Identifiers (LSIDs), the project has been able to leverage third party tools such as IBM's Haystack and LSID LaunchPad.

As the current funding finishes, many challenges remain. myGrid's ability to rapidly create and run in silico experiments is now being used by a range of newly funded Life Science projects such as PsyGrid, CLEF-Services, e-Fungi, ISPIDER, Integrative Biology and ComparaGrid. These and other projects, such as the EU FP6 STREP OntoGrid, also contribute to our continued middleware development. Further technical challenges in integrating disparate and distributed applications include: the incorporation of interactive, computationally intensive simulation services, such as heart models; the efficient management of very large data sets; the visualisation of results; and knowledge mining provenance metadata.

Links:
myGrid: http://www.mygrid.org.uk
OGSA-DAI: http://www.ogsadai.org.uk

Please contact:
Carole Goble,
The University of Manchester,UK
E-mail: carole.goble@manchester.ac.uk

Towards an Integrative and Context-Sensitive Approach to In Silico Disease Modelling

by Matej Oresic, Peddinti V. Gopalacharyulu, Erno Lindfors, Catherine Bounsaythip, Ilkka Karanta, Mikko Hiirsalmi, Lauri Seitsonen and Paula Silvonen

Historically, the scientific methods applied to biological problems have largely been limited due to the fact that it has been difficult to collect the data. Today, these scientific methods are challenged by the 'omics' revolution, which are empowering us with the ability to collect large amounts of data in parallel from a particular system. However, the development of efficient tools to exploit this data within the context of biological systems has been much slower. Due to this mismatch the knowledge acquisition in life sciences is actually increasingly difficult and new information technology solutions are needed to resolve this problem.

The overall objective of the project is to build an integrated system with explanatory and predictive abilities to represent biological and clinical level knowledge related to human diseases. In order to achieve sufficient focus, we will initially limit the project to domains of type II diabetes, cardiovascular complications associated with metabolic syndrome, and obesity. The project is a collaboration between the two VTT units, Information Technology and Biotechnology. We will also cooperate with other national and international experts in specific disease domains.

The primary reason for pursuing this goal is that a vast yet dispersed amount of information already exists about these diseases, and with advancement of new high-throughput technologies the amount of information is rapidly increasing. But this information will become knowledge only when it is mapped to a certain knowledge structure, ie, organized or linked together in a way that makes it accessible or interpretable by the users. In fact, the same pieces of information can yield different knowledge depending on the context and purpose,. Moreover, context-sensitivity means also that the 'emerged' concepts should match the community's consensus, ie, people working in the same field.

Project Structure

Our project can be divided into several sub-projects:

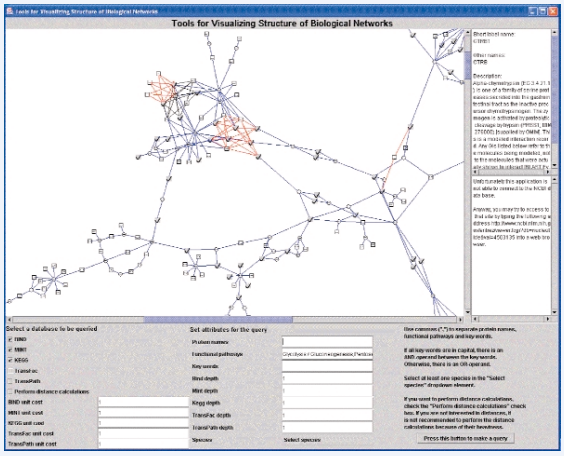
- develop multilevel bioinformatics data management framework based on XML and Semantic Web technologies
- develop software solutions to map the information to a knowledge structure, including the use of text annotation, common vocabularies and semantics
- implement a data and text mining solution for mining quantitative and qualitative relationships between entities of interest
- select most salient concepts in respect to different levels of the biological system to model disease, and set up an a priori framework for the ontology model linking the levels
- populate the model within the ontology framework, modify the model and framework and add new entities through validation.

In longer term, we wish to pursue the following goals:

- develop methodology to create predictive quantitative and dynamical models from the knowledge base
- validate the modelling tools by applying them to predict responses to specific interventions for which sufficient data already exists.

Handling Structured Data

Much of biological information, in particular at biomolecular level, is stored in databases, such as genome databases (eg GenBank), pathway databases (eg KEGG), protein databases (eg UniProt) or small ontologies like Gene Ontology (GO). Even at this level, a continuous frustration of researchers is the fact that it is not easy to map identical entities across multiple databases, due to different naming conventions. Therefore, any attempt at data integration should start with identifying the 'atoms of information' and creating solutions to resolve the names. XML and RDF are useful technologies for creating such identity-mappings across multiple data sources. There are ongoing efforts to standardise the life science data formats in order to facilitate the exchange of information and knowledge (eg W3C Semantic Web Life Science). For every database (either containing annotations or information about entity relationships) we create a simple XML schema that enables mapping to other databases.



Screen shot of Java based tool for investigating biological networks. Example of pathway retrieval query with nearest neighbour search in KEGG, MINT and BIND is shown. Multiple organisms can be queried simultaneously in comparative manner.

Textual Data and Concepts

Most of life science information is still available only in textual form. This is particularly true for information on relationships between the molecular level events and more complex concepts such as events related to a multifactorial disease. Text mining solutions are therefore needed to sift through the semi-structured disease-related data such as OMIM, literature (eg PubMed) and patent databases. Established vocabularies and concepts such as UMLS may be of help, but one should be aware that with rapid progression of life sciences and medicine today the new terms and concepts are rapidly emerging. There have already been efforts in the text mining area to infer knowledge from available sources in the domain, for instance the BioCreAtIvE community is set up to assess current text mining tools in the area of biology.

In our project, we use GATE as software platform, augmented with various scripts and programs. We extract (subject, predicate, object) triplets from the raw text mass. We use a seed vocabulary of terms – at least two terms of the triplet must be found in the vocabulary – and the vocabulary is then augmented with the third term, if needed. The extraction is based on shallow parsing: instead of forming complete parse trees from each sentence, we rather extract noun phrase and verb phrase blocks that are further processed to produce the triplets. In the end, the combination of all these triplets will form a conceptual graph that tells about the relations between the entities of interest.

Everything is Connected!

As a start towards these goals, VTT BEL has already integrated data from UniProt, KEGG, TransFac, TransPath, MINT, and BIND databases using XML; and developed a Java-based tool that allows parallel retrieval across multiple databases, including metabolic pathways, protein-protein interactions, signaling and regulatory networks. The results are then visually displayed as a network (see figure). Edge attributes contain information about type of relationship, possibly quantitative or semantic information (such as 'is located in' in case of linking a protein with a complex entity such as cell organelle, with information obtained by text mining).

Our aim is to connect entities (nodes), which can include molecular entities as well as more complex concepts such as insulin resistance or diabetes, with relationships (edges) which can either be direct physical interactions or more complex relationships. Nodes and edges may be clustered and mapped to an ontology-type structure. Specifically, we are also interested in retrieving quantitative information on relationships which can be used for predictive modeling in the future.

Multilevel Inference

Finding linkages between different entities in multifactorial diseases is a demanding task due to multiple levels of biological organization being involved in disease pathogenesis and progression. At the first phase of the project we want to accumulate and integrate the data on entities and relationships among them, which can be used for explanatory modelling, ie, for interpretation of results from experiments using clinical or genomic screening, transcriptional, protein, or metabolite profiling. Once we obtain the networks of associated entities across different levels, the inferences will be made based on the network's topology. Therefore, our initial computational efforts will focus on studies of network topologies and their associations with known experimental data and physiological conditions.

Links:

VTT: http://www.vtt.fi/

Quantitative Biology and Bioinformatics group at VTT BEL: http://sysbio.vtt.fi/qbix/

GenBank: http://www.ncbi.nlm.nih.gov/

UniProt: http://www.ebi.uniprot.org/index.shtml

Kyoto Encyclopedia of Genes and Genomes: http://www.genome.jp/kegg/

Gene Ontology: http://www.geneontology.org/

W3C Semantic Web for Life Sciences: http://lists.w3.org/Archives/Public/public-semweb-lifesci/

PubMed: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

Online Mendelian Inheritance in Man: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Unified Medical Language System: http://www.nlm.nih.gov/research/umls/

BioCreAtIvE: http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html

GATE: http://gate.ac.uk/

Please contact:
Matej Oresic, VTT Biotechnology, Finland
Tel: +358 9 456 4491
E-mail: matej.oresic@vtt.fi

The Prognochip Project: Transcripromics and Biomedical Informatics for the Classification and Prognosis of Breast Cancer

by Dimitris Kafetzopoulos

The completion of the Human Genome Project and the development of post-genomic applications have allowed new holistic approaches to disease analysis that will revolutionize biomedical research and health care. Consultation of both the comprehensive genotypic information of the patient and the detailed molecular classification of the disease will result in individualized treatments. The Greek 'Prognochip' project applies this approach in the field of breast cancer prognosis and treatment.

Breast cancer affects approximately one in ten women, and as such is one of the most common female malignancies. Breast cancer is both genetically and histo-pathologically heterogeneous, and the mechanisms underlying its development remain largely unknown. Breast cancer patients diagnosed with the same stage of disease often have remarkably different responses to therapy. Even with the strongest prognostic indicators, such as lymph-node status, estrogen receptor expression and histological grade, it is not possible to accurately classify breast tumours according to their clinical behavior. Genomic background and variations in the transcriptional programs account for much of the observed diversity. The 'Prognochip' project, funded by the Greek General Secretariat for Research and Technology, aims to identify and validate 'signature' gene expression profiles of breast tumours that correlate with other epidemiological or clinical parameters. This should provide a more accurate prognosis and prediction of response to therapy: a clear benefit to almost three out of four women who receive aggressive chemotherapy treatment, although they would have survived without it.

The major tasks within the Prognochip project are as follows:

Patients are informed and consent to the molecular and genetic data analysis of their tumour specimens and blood samples, provided that their anonymity is ensured. A tissue procurement protocol has been designed for tissue collection and storage and a tissue-bank system has been established for proper tissue filing and management. Patients with malignant tumours are staged according to the TNM system and a set of immunohistological markers are examined. A DNA microarray of long oligonucletide probes has been designed, representing all known human genes – approximately 35,000 different transcripts of 27,000 different genes. A common reference material has been chosen for the study, consisting of a defined set of cell-line extracts, thereby ensuring accurate quantitation of gene expression. After hybridization, fluorescence intensity images representing gene expression levels are acquired as 16-bit TIFF files and stored in BASE, a MySQL database developed by the University of Lund, Sweden. Special plug-ins have been created for data pre-processing (filtering, normalization) and analysis.

In general, two computational approaches from a suite of intelligent data processing tools are used for tumour classification. The first approach is the 'unsupervised' analysis, in which no source of knowledge is used to guide the analysis process. Instead, the data are searched for patterns with no a priori expectation concerning the number or type of groups (gene and tumour clusters) that might be present. The second is the 'supervised' analysis, in which we search for genes whose expression patterns (denoted X and Y in Figure 1) correlate with external parameters (denoted A and B). The 'supervising' parameters are clinical features such as the clinical outcome (including overall survival, relapse-free survival times, metastasis etc), other molecular markers, chromosomal aberrations, patterns observed with other diagnostic methods and responses to (chemo)therapy.

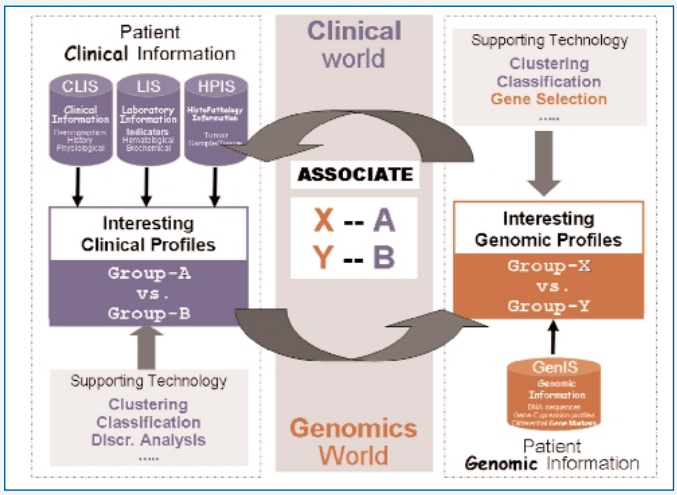


Figure 1: 'supervised' analysis for tumor classification.

In addressing classification, there are two issues: a) class discovery, the definition of previously unrecognized tissue

sub-types, and b) class prediction, the assignment of particular samples to existing classes (this could reflect current states or even future outcomes).

The main challenge of the Prognochip prospective study is the development of interoperable and efficient clinical and genomic information systems for the integration of heterogeneous data. In that context we are working towards the delivery of an Integrated Clinico-Genomics Information Technology Environment (ICG-ITE). The envisioned building blocks of the ICG-ITE include:

- a set of clinical information systems to store patients' clinical information
- an information system to store and manage the specifications of the respective microarray experiments, analyse measured bioassaysand store samples' genomic information
- a middleware layer for information/ data integration and intelligent processing.

The middleware layer is realized through a 'puzzle' of integrated software components that together enable:

- the seamless and efficient extraction of data from various sources (clinical and genomic)
- uniform information modelling through the utilization of standard clinical/genomic data models and respective ontologies
- uniform information representation through the utilization and appropriate customization of RDF/XML technology
- intelligent data processing and visualization through a suite of data mining components and tools.

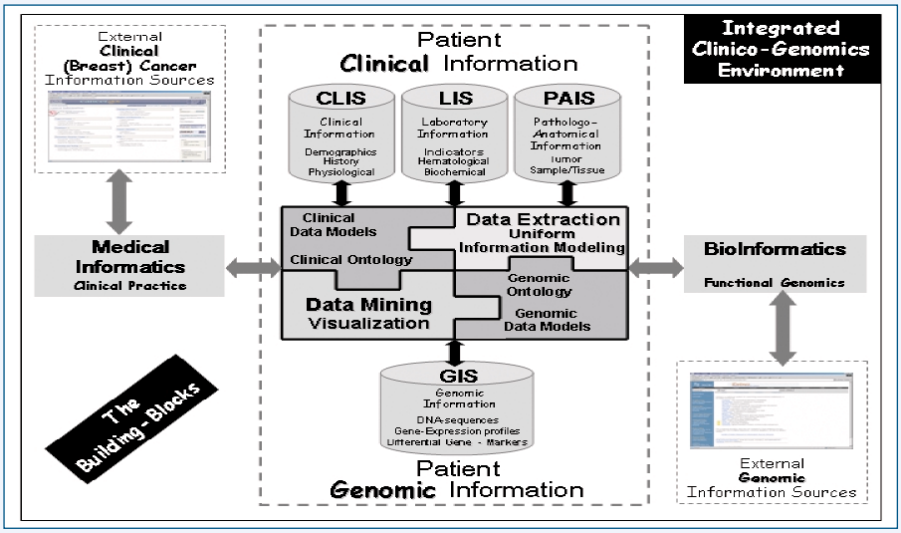


Figure 2: General layout of the Integrated Clinico-Genomics Information Technology Environment.

Since the integration of clinical and genomic data is such a demanding problem, there is a clear need to elaborate on the concept of Integrated Electronic Health Care Record architectures, utilize technological advances and extend the standard clinical data models to include and amalgamate genomic ones. Further, an equally important security and authorization infrastructure is employed. A general layout of the provisioned ICG-ITE is shown in Figure 2.

Please contact:
Dimitris Kafetzopoulos, IMBB-FORTH
Tel: +30 2810 39 1594
E-mail: kafetzo@imbb.forth.gr

Processing Multimedia Biomedical Information for Disease Evolution Monitoring

by Sara Colantonio, Maria Grazia Di Bono, Gabriele Pieri and Ovidio Salvetti

A methodology for the automatic monitoring of diseases using biomedical data in multimedia is proposed. We adopt a computational intelligence approach mainly based on a multilevel neural network architecture. This approach has been employed in applications for neurosignal and image categorisation.

Models that monitor disease evolution are often based on approaches to diagnosis and prediction that compare the information obtained from a set of diagnostic exams against a set of reference parameters. A set of rules then derives a prediction of the actual and future state of health of the patient. Such models are very rigid and do not easily adapt to variations in the evolution of diseases under different and/or aleatory conditions.

There are a number of other approaches to the monitoring of disease evolution. Model-based diagnostic approaches need accurate domain models and require a fixed number of diagnostic classes, which thus reduces their flexibility. In case-based approaches, the expert's knowledge is stored in a library of cases; however, the search for the best-matching case can be computationally expensive. Inductive learning, which can include decision trees, statistical classifiers and neural networks is also used.

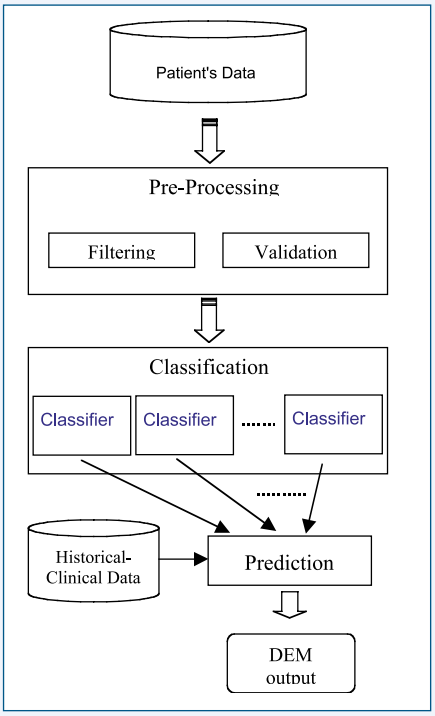


Figure 1: Architecture of the model.

We are currently studying a new methodology that belongs to the field of computational intelligence. Certain diseases can be monitored automatically by processing correlated biomedical data in different media. In particular, we are developing an innovative model for the realization of a valid diagnosis methodology that can also provide an evaluation of the course of the monitored disease, based on a Multilevel Artificial Neural Network (MANN) architecture.

Multimedia information related to diagnostic and/or therapeutic processes can be of different types and comes from several sources:

- images (eg CT scan, MR scan, etc.)
- signals (eg EEG, pressure, etc.)
- historical and clinical data
- laboratory data (eg blood or specimen analyses, etc.).

We are thus designing and developing a computational method to support disease diagnosis and evolution prediction that can process these different types of biomedical data.

For any given case under examination, a set of parameters, consisting of features extracted from the multimedia data available, is processed through a MANN. The MANN is organised in three hierarchical levels corresponding to three processing phases:

- pre-processing phase, dedicated to input data filtering and validation to eliminate incongruencies and to control consistency
- classification phase, performed on the output of the previous phase to identify the actual state of the disease
- prediction phase, to evaluate the course of the disease using the results of the previous phase together with patient's historical and clinical data.

The multilevel approach guarantees both specialisation and adaptability, since the single levels of the network can be optimised according to the specific characteristics of the problem to be solved, and the entire architecture can be easily integrated if the problem description changes, by simply training only those levels involved in the modification.

The advantages of using a MANN architecture can be summarised in two main points: the modular organisation facilitates analysis of the networks and the hierarchy enables the network to finely tune itself towards recognising the most promising directions to look for relevant patterns.

This multilevel architecture is shown in Figure 1. Multimedia data on the patient are acquired through several diagnostic modalities and passed to the pre-processing phase. This phase is dedicated to the selection of the relevant parameters (filtering) and to their validation with respect to known reference data. The pre-processed data are classified in the second phase in order to evaluate the patient's current state. The final phase provides a prediction of the evolution of the disease, using two different sets of data: the parameters obtained from the classification level, and historical and clinical data for other patients and diseases.

Our model has been applied to a real case study of a brain pathology, selected because of its clinical interest. Using a set of multimedia data composed of transcranial doppler (TCD) ultrasound signals, magnetic resonance (MR) neuro-images and other clinical data, the model has been instantiated for the monitoring of carotid and cardiovascular steno-occlusive diseases and the diagnosis and follow-up of cerebral anomalies. Figure 2 shows some of the information used to monitor this patient.

From the results of our first experiments, this model appears to be an effective tool for supporting the diagnosis activity and work is underway to integrate it into a hybrid system for medical decision support.

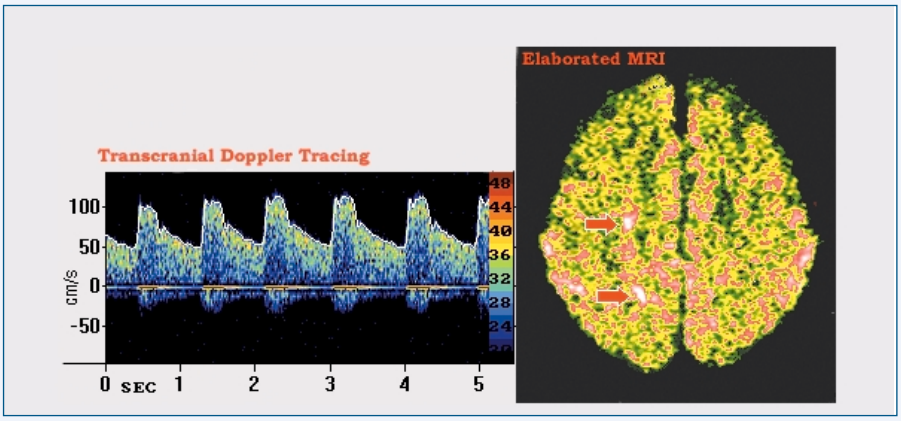


Figure 2: On the left, the transcranial doppler spectrum used to diagnose blood flow anomalies; on the right, a magnetic resonance image processed to differentiate different tissue densities (the arrows indicate two detected ischemic regions).

Please contact:
Sara Colantonio or Maria Grazia Di Bono,
ISTI-CNR, Italy
Tel +39 050 315 3146
E-mail: Sara.Colantonio@isti.cnr.it,
Maria.Grazia.Dibono@isti.cnr.it

Privacy Concerns in Biomedical Informatics

by Brecht Claerhout

Consensus exists on the fact that the integration of Medical Informatics (MI) and Bio Informatics (BI) will lead to unprecedented scientific opportunities. This evolution towards BioMedical Informatics (BMI) is supported by the vision that a combination of information on all levels (molecule, cell, tissue, individual, population) will lead to an improved individualized healthcare. However there is a downside: the increased collection and (combined) processing of sensitive personal health information raises serious questions regarding citizens' privacy.

(Bio-)Medical data usually have a sensitive nature and although generally used for the benefit of the community, this information is quite prone to abuse. There is an appropriate concern about the proper treatment of the increasing volume of sensitive data. Incidents of abuse have been previously reported in the public media, proving that the threat is genuine. It can be easily understood that abuse of sensitive personal healthcare information can lead to considerable financial gain for malicious people. Imagine the impact on society, when banks, insurance companies, employers, etc. could access healthcare data about their customers, revealing past, current and probable future (cf. genomics) health conditions. Indeed, abuse of medical data can affect us all, as at some point in life practically everyone is confronted with loan, insurances or job applications.

Public authorities are also aware of these repercussions and are putting considerable effort into privacy protection legislation (especially in Europe). The further elaboration and application of these laws is oriented towards technical means for protecting a person's privacy, instead of focussing merely in obtaining 'informed consent' and following guidelines.

A classical approach towards safeguarding confidentiality focuses on the creators and maintainers of the data, prohibiting them from disclosing the information to inappropriate parties. Basically, this comes down to the deployment of traditional security measures (access control, authorisation). A more advanced approach is incorporating real Privacy Enhancing Techniques (PETs) into biomedical data collection and processing systems. Complementary to standard security solutions, PETs can be defined as (according to J. Borking): "A coherent system of ICT measures that protects privacy by eliminating or reducing personal data or by preventing unnecessary and/or undesired processing of personal data, all without losing the functionality of the information system."

Over the last decade a lot of research related to Privacy Enhancing Technologies has been performed, both in the USA and in Europe. In Europe a large part of this research has been funded by the European Commission (the European interest in privacy issues can also be seen in the legislation effort). Some of these projects have focussed on PET solutions for biomedical applications, such as the PRIDEH-GEN project (Privacy Enhancement in Data Management in E-Health for Genomic Medicine), which studied privacy protection of genomic information in particular. Although all characteristics, typically associated with genomic data, that impact privacy, are also encountered in routine clinical information (from the EHR), they deserve special attention. These characteristics as listed below are usually not found all together or to such an extent in medical data as in genomic data:

- genetic data not only concern individuals, but also their relatives. A person's consent to release his or her genetic information constitutes a de facto release of information about other individuals, ie, his or her relatives. In the case of genomic medicine, there is a complex interaction between individual rights and collective requirements
- medical data deal with past and current health statuses of persons, whereas genetic information can also give indications about future health or disease conditions
- the full extent of the information included in the genomic data is not known yet, hence it is difficult to assess the full extent of disclosure
- genomic data are easily wrongly interpreted by non-professionals, 'susceptibility' to diseases can easily be mistaken with certainty of illness.

This research has first off all proven the use of PETs in practical situations (also because of EU funded take-up-measures), and given birth to commercial grade privacy protection services. Commercial grade, meaning generic solutions needing little or no customisation and providing transparency towards already used ICT tools (eg offering privacy protection in a service oriented way). These applications, mainly de-identification tools (such as pseudonymisation systems, privacy risk assessment tools, controlled database alteration algorithms, etc.) are now deployed by pharmaceutical companies (eg for post marketing follow up) and research institutes (eg for epidemiological studies, disease management studies, etc.).

Secondly, privacy protection remains an important topic within eHealth research, evolving in synergy with healthcare ICT. Networks of Excellence exploring the possibilities of biomedical informatics, such as INFOBIOMED (http://www.infobiomed.org/) are aware of the increased privacy risks associated with their applications and put a considerable effort in deploying technical means for privacy protection (through pilots). It is common belief that without this effort valuable data will remain unlocked for research (people will not be willing to share data unless their personal privacy is adequately protected).

Privacy Enhancing Technology can help maintain the balance between personal well-being (right for privacy) and collective benefit (sharing of sensitive data for research). The increasing commercial deployment and continuing research effort (eg privacy and security solutions for Biomedical GRID) will hopefully lead to a situation where PETs are used 'by default' in eHealth.

Please contact:
Brecht Claerhout, Custodix NV, Belgium
Tel: +32 9 210 78 90
E-mail: Brecht@custodix.com

GenoLink: Discovering Drug Target Proteins by Exploring Networks of Heterogeneous Biological Data

by Patrick Durand, Laurent Labarre, Alain Meil and Jérôme Wojcik

As ever-larger amounts of biological data are made available by genomic and post-genomic technology, the conversion of these data into valuable knowledge becomes crucial to the discovery of new drug target proteins. For this reason, the GenoStar Consortium has developed GenoLink, a software platform designed for the integration and exploration of complex biological datasets.

Genomic and post-genomic technology is producing a huge amount of heterogeneous data available for investigating the functions of proteins. Biologists are now faced with the problem of exploiting these raw data and extracting useful knowledge from them. An efficient way to perform this task consists in exploring the relationships between various kinds of biological data. As a very simple example, one can assign a function to a protein from a given organism knowing that this protein has a sequence similar to a well-known protein from another organism. In such a case, a network of objects (proteins and organisms) can be explored using relationships (eg a protein 'belongs' to an organism and a protein 'is similar' to another one). More generally, the genomic world can be viewed and explored using a complex network of biological objects and their relationships. Following this idea the GenoStar Consortium has developed GenoLink, a software platform dedicated to the exploration of complex networks of biological data. GenoLink is built on three modules that handle data modelling, data integration and data querying/visualization.

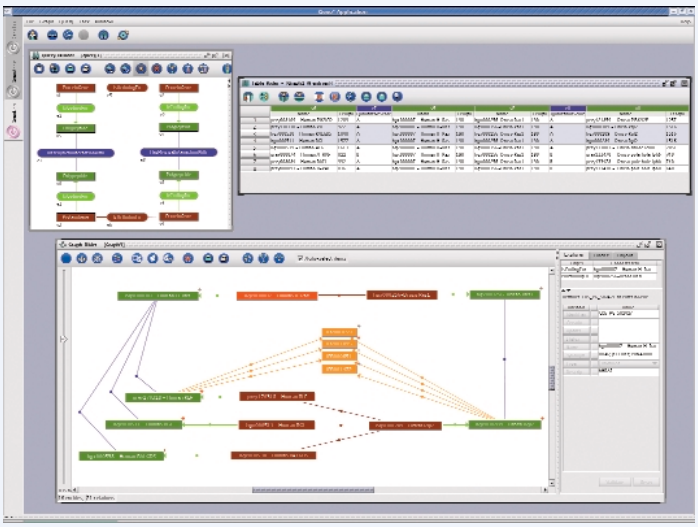
Data Modelling

GenoLink uses GenoCore to model and store networks of data. GenoCore is an advanced object-oriented knowledge management system developed at INRIA. Among other features, GenoCore provides two complementary entities of representation: classes and associations. GenoLink uses these to formally describe the objects and relationships within a network. As in any object-oriented system, a class or an association represents a set of objects or relationships respectively. Classes and associations can be organized in hierarchies, allowing inheritance and specialization mechanisms.

To discover new proteins of pharmaceutical interest we have set up a knowledge base describing a network of data centred on protein-protein interactions. For that purpose our data model can handle information coming from complementary sources of data, such as fully annotated genomes, pre-computed clusters of orthologous genes, functional classifications, protein domains and protein-protein interactions. It is worth noting that this data model can be updated to accommodate new sources of data.

Data Integration

To supply the knowledge base with data, GenoLink comes with a set of import tasks. These are organized to perform integration within a single data space. Data integration is carried out using unique object identifiers as they appear in the source databases. Going along with the idea of an extensible data model, new import tasks can be added to GenoLink using the application-programming interface provided with the software.



GenoLink at work on the exploration of conserved protein-protein interactions involving Ras network proteins from Homo sapiens and Drosophila melanogaster (data from Hybrigenics). Oncogenic mutations in this network are responsible for various cancers.

Data Querying and Visualization

To support the exploration of complex networks of data, GenoLink has a dedicated and original query system. Queries are network patterns built up from the classes/associations defined in the data model. For a given query, GenoLink will search for matching patterns by looking for sub-networks within the full knowledge-base network. It is then possible to further explore the network starting from a query result. Through this exploration process, the biologist may be able to infer new links between previously unrelated entities.

GenoLink was entirely written using Java technology. The creation of queries and the exploration of results is achieved in an easy-to-use graphical environment. It is worth noting that GenoLink also provides a full-featured graph query language that allows the design of complex querying strategies.

Applications

In addition to the identification of proteins of pharmaceutical interest in the fields of cancer, GenoLink has been used to annotate proteins from various bacteria genomes, and to infer protein-protein interactions in several eukaryotic genomes.

The GenoStar Consortium

GenoLink is one of the application modules of GenoStar, a bioinformatics platform for exploratory genomics. Other application modules are GenoAnnot (for the annotation of bacterial genomes) and GenoBool (a data mining application). The development of the platform was initiated by the GenoStar consortium at the end of 1999. Partly supported by the French Ministry of Research, it brings together four partners: two biotech companies – Hybrigenics (Paris) and Genome Express (Grenoble) – and two research institutes – the Pasteur Institute (Paris) and the INRIA Rhône-Alpes (Grenoble). The GenoStar platform is now maintained by the private company GenoStar. The platform remains freely available for the academic community.

Link:
http://www.genostar.org

Please contact:
Patrick Durand, IRISA-INRIA,France
Tel: +33 2 9984 7321
E-mail: Patrick.Durand@inria.fr

Jérôme Wojcik, Hybrigenics S.A., France
Tel: +33 1 5810 3862
E-mail: jwojcik@hybrigenics.com

Mining Genomic Data with Metaheuristic Techniques

by Carlos Cotta and Pablo Moscato

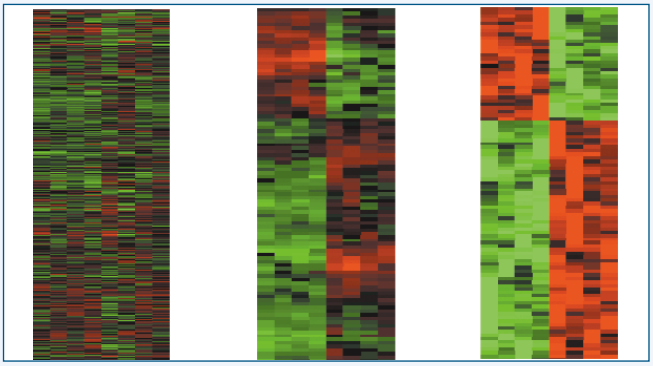
One of the major challenges in biomedical research is discovering the genetic basis of diseases such as Alzheimer's disease or schizophrenia. Research in the GISUM group at the University of Málaga is investigating the use of evolutionary metaheuristics for this purpose.

The pace at which research currently occurs is affecting most fields of science, not least the rapidly developing areas of molecular biology and genomics. The many initiatives in the life sciences that are planned or currently in execution are producing an unprecedented flood of data. As a result, many of the challenges in biology are increasingly becoming challenges in mathematics, and fundamentally in computing. The task of dealing with the large-scale combinatorial problems arising in bioinformatics is undoubtedly one of the greatest challenges facing computer science researchers, and new techniques and insights for algorithm design are required.

In molecular biology, the analysis of gene expression data represents one of these challenges. The difficulty of the problem lies in its computational complexity and the sheer amount of data that must be processed. For example, thanks to microarray technology we can monitor the activity of a whole genome in a single experiment. Huge amounts of data are becoming available thanks to this technique, providing access to a better picture of the simultaneous interactions of thousands of genes. The challenge now is to unravel the complex functional dependencies behind these data, and identify the linkage between genetic information and its phenotypical correlates.

From an algorithmic point of view, the data mining process can be shown in general to be hard according to the traditional P vs. NP scenario. For instance, assume that genomic data is available from both healthy individuals and patients affected by a certain pathology. Finding a minimal subset of genes such that the phenotypic status – healthy or ill – can be derived from their combined expression values is a problem for which no efficient (polynomial time) algorithm is known. This is just one example of the extremely hard tasks to be found in this domain.

In this situation, the classical approach is to define approximation algorithms. This is unpractical in many situations however, and has been superseded by two cutting-edge methodologies. In the first of these, parameterized complexity helps to identify tractable subclasses of these problems for a certain realistic range of some structural parameter of the problem; the resulting techniques are called fixed-parameter tractable (FPT) algorithms. In the second approach, modern heuristic techniques (metaheuristics) are employed to produce probably (though not yet provably) optimal solutions for these problems. The GISUM group of the University of Málaga (UMA) in Spain is working on this line of research in close cooperation with other centres worldwide, in particular the Newcastle Bioinformatics Initiative (NBI) in Australia. This cooperation has yielded numerous relevant results, and has been substantiated in an on-going project funded by the Australian Research Council on the area of genomic data mining with evolutionary algorithms (EAs).



Left: Microarray data comprising expression values of 2984 genes from eight individuals affected by two types of lymphoma. Middle: a subset of 100 genes found by an evolutionary algorithm that provide a robust classifier for the disease subclass. Right: the same subset of genes after gene expression values are renormalized using the thresholds provided by the evolutionary algorithm.

We have been able to show that minimum feature-set selection problems are intrinsically hard beyond the NP sense. More precisely, our results indicate that no FPT algorithm is possible for this problem, thus emphasizing the prominent role that metaheuristics must play in this domain. Furthermore, given that the data are inherently noisy and prone to measurement errors, robust feature identification methods are essential. Our research has provided evidence that the joint use of evolutionary algorithms with a reduction approach inspired by kernelization rules often used in the design of FPT algorithms can provide good solutions for this problem, eg for discriminating between different types of lymphoma (see Figure 1).

This methodology has also been deployed in conjunction with single-value decomposition and integer programming on microarray data from the brains of Alzheimer's patients and healthy patients used as a control. A clear pattern of differential gene expression is obtained, which can be regarded as a molecular signature of the disease. The results suggest that a unified approach may help to uncover complex genetic risk factors not currently discovered with a single method.

Our aims are now to formalize new problems in genomics as combinatorial, non-linear, mixed or multi-objective optimization problems and to study their computational complexity. Subsequently we plan to identify the best way of addressing and solving these problems using EAs and, where justified, hybridize the methods with exact algorithms and other types of metaheuristics. These techniques will be implemented in a unified software framework.

New collaborative initiatives are also underway. The most ambitious of these focuses on the use of bio-inspired techniques (subsuming EAs as well as artificial immune systems, or ant colony optimization) for mining genomic data, and involves institutions from Australia (NBI), France (Universities of Paris and Lille), the Netherlands (Free University Amsterdam), Spain (UMA) and the United Kingdom (University of Kent). We invite everyone to view our results and contact us with comments or suggestions for further collaboration.

Links:
http://www.lcc.uma.es/~ccottap
http://www.cs.newcastle.edu.au/~nbi/

Please contact:
Carlos Cotta
University of Málaga/SpaRCIM, Spain
Tel: +34 952 137158
E-mail: ccottap@lcc.uma.es

Mining Distributed and Heterogeneous Clinical Data Sources

by George Potamias

The HealthObs (Health Observatory) integrated environment offers seamless integration and intelligent processing of distributed and heterogeneous clinical information. This is achieved through the use of XML and data mining. Its aim is to assist healthcare professionals in coming to grips with the vast amounts of information and to enhance their decision-making capabilities.

HealthObs is an info-mediation and brokerage environment with 'knowledge discovery' functionality. It is composed of two synergistic layers (see Figure 1).

The Middleware Layer is a set of software components that enables (i) access to the distributed information and data sources regardless of platform, location and type of information system; (ii) fusion and semantic homogenization of information and data items, enabled by the uniform information modelling of the underlying information and data items and supported by advanced ontology and RDF/XML technologies; and (iii) intelligent information and data processing based on advanced data mining operations.

The Application Layer is based on the use and integration of the middleware components, for specific medical domains.

Central to the architecture is a single data-enriched XML file that contains information and data from remote (and potentially heterogeneous) clinical infor-

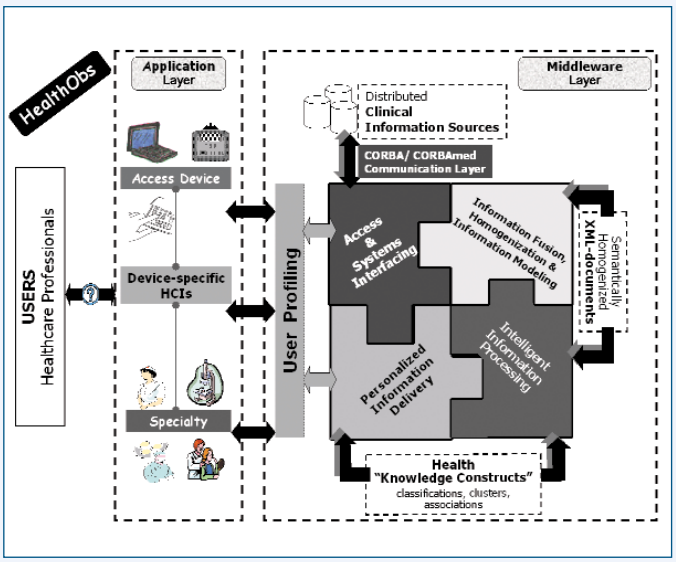


Figure 1: HealthObs synergistic layers.

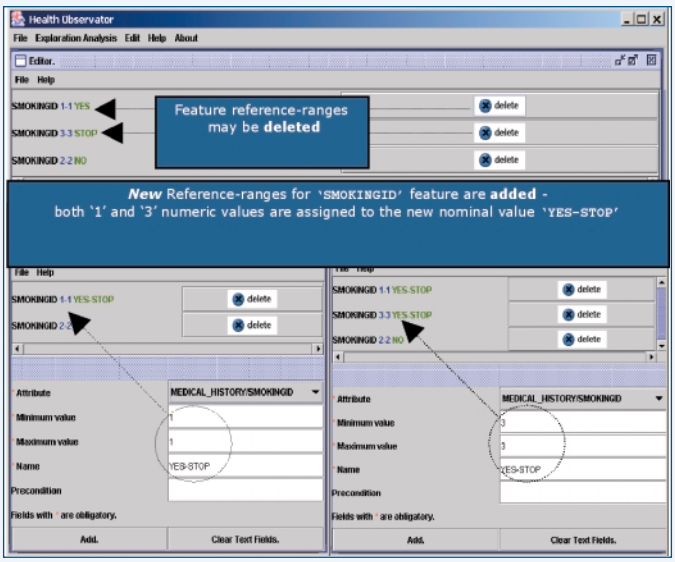


Figure 2: Domain Editor tool.

mation systems. To this end, the Integrated Electronic Health Care Record (IEHCR) services of HYGEIAnet (the Regional Health Information Network (RHIN) of Crete; http://www.hygeianet.gr) are utilized in order to query the federated clinical information sources and recall the relevant query-specific data items. For each query, and with the aid of custom-made filtering and formatting operations, a query-specific XML file is created. HealthObs initiates and bases its knowledge discovery (ie association rules mining) operations on such data-enriched XML files. In this respect, HealthObs falls into the category of XML-content mining tools.

Semantic Homogenization and Domain Adaptation

A domain-specific ontology is required in order to hide the heterogeneous nature of data. A special service known as the Common Clinical Term Reference service (CCTR) has been developed for the storage and retrieval of common and universally accepted names and codes of medical terms. This service uses terms and relations from the ICD9 (International Coding of Diseases) and ICPC (International Classification of Primary Care) standard coding schemes. Moreover, when dealing with clinical laboratory findings it is crucial to refer to qualitative, rather than numeric measurements. CCTR offers a means of assigning qualitative values to the different measurements. In HealthObs this is made operational via a special Domain Editor tool (see Figure 2).

Mining Interesting Clinical Associations

To enable the adaptation of Association Rules Mining (ARM) operations we have instituted two key conventions. In the first of these, each transaction corresponds to a specific patient encounter (ie identifiable visits of patients to a healthcare unit within the federation). Each encounter/visit is uniquely identified by reference to three attributes recorded in the respective clinical information systems; namely patientid, information-

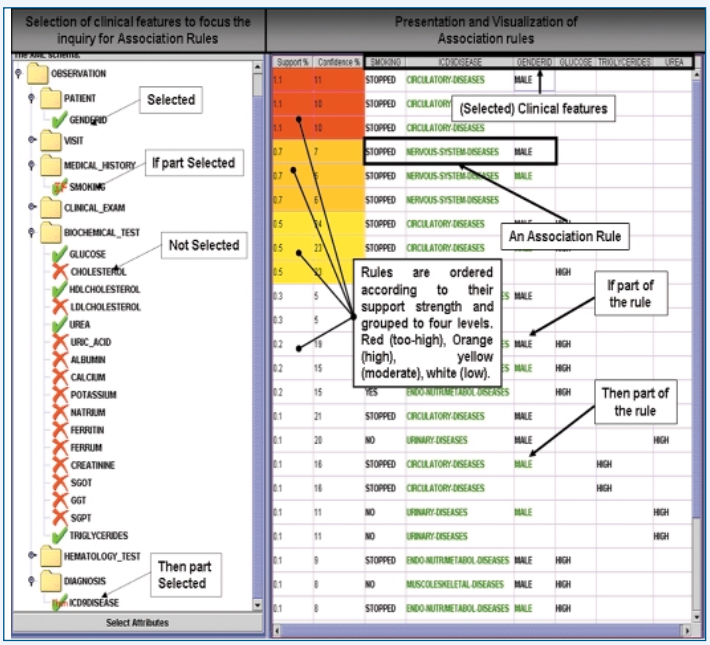


Figure 3:
Selection of clinical features (left frame):
- **if the user only checks a feature, then this feature may or may not be present in the rule, ie a non-obligatory feature (eg HDLCHOLESTEROL)**
- **if the user not only checks a feature but also posts an 'if' tick (eg GENDERID) or a 'then' tick (eg SMOKING) on it, then the presence of the feature is obligatory in the 'if' or 'then' part of the rules.**

Visualization of associations (right frame):
entries in black and green indicate the 'if' and 'then' composites of an association rule respectively; furthermore, rules are ordered according to their support strength, which is indicated by colour.

system-id, and encounter-id (or visit-id). Patient data are recalled anonymously and retrieved in a secure manner by security and role-based authorization services offered by the RHIN information infrastructure. In the second convention, each item is represented by the following triplet: '<SyntheticFeature, AtomicFeature, AtomicFeatureValue>', the entries of which correspond to instance elements present in the XML file to be processed. For example, a SyntheticFeature may correspond to the 'biochemical test' element; AtomicFeature to an element that stands for a specific biochemical test, eg 'glucose'; and FeatureValue to the value of the atomic feature, eg '68.0' for glucose.

With the aid of the ARM component, HealthObs is able to form significant and useful associations of the type: $X_1, X_2, ... X_n \Rightarrow Y_1, Y_2, ..., Y_m$, where each pair $<X_i, Y_i>$ corresponds to specific clinical features and findings. For example, "*IF GLUCOSE = high and UREA = high THENISCHEMIC-HEART disease [Support: 1%, Confidence: 60%].*"

In HealthObs, both the selection of clinical features to focus the knowledge discovery operations, and the visualization of discovered association rules are achieved via a specially designed graphical user-interface (GUI). An indicative screen-shot of this interface is shown in Figure 3.

HealthObs system has been successfully used for various knowledge discovery tasks in the context of the HYGEIAnet RHIN. Furthermore, work is in progress to extend the system's information and data model towards the incorporation and management of genomic information (eg patients' samples, gene-expression profiles, gene markers etc). The work is being undertaken in the context of the PrognoChip project funded by the Greek Secretariat for Research & Technology, the aim of which is the discovery of reliable prognostic molecular (ie gene) markers for breast cancer and their linkage and validation with identifiable clinico-histological patient profiles.

Please contact:
George Potamias, ICS-FORTH, Greece
E-mail: potamias@ics.forth.gr

Validation of Clustering Techniques for Microarray Gene Eexpression Data

by Nadia Bolshakova and Pádraig Cunningham

Recent advances in microarray technology have enabled the measurement of the simultaneous expression of thousands of genes under multiple experimental conditions. The methods implemented in this research may contribute to the validation of clustering results and the estimation of the number of clusters. For instance, these tools may be used for the identification of new tumour classes using gene expression profiles. The results show that this estimation approach may represent an effective tool to support biomedical knowledge discovery and healthcare applications.

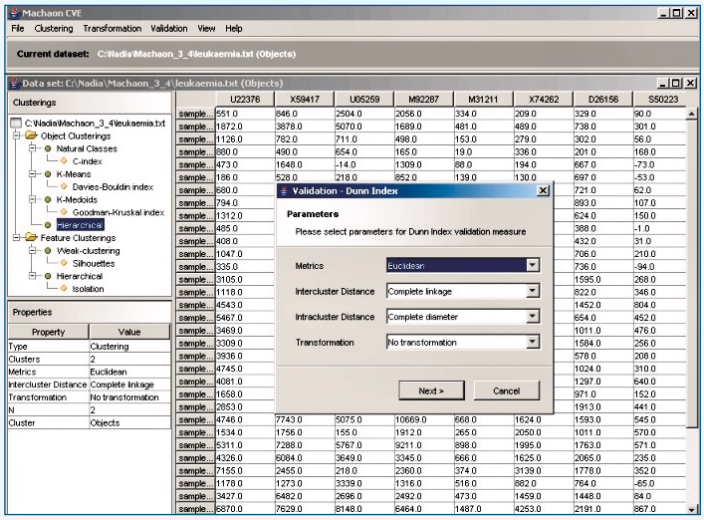
One of our major tasks is to advance data analysis and integration capabilities in genomic expression pattern discovery and classification. It has consisted of the implementation of algorithms and tools to organise and categorise genome expression data. It has integrated and improved a number of machine learning techniques, which may aid in the identification of relevant features for diagnostic, prognostic and system biology studies. These tools may also be applied to other information management domains such as biomedical informatics. Moreover, automated discovery solutions may assist the design of novel techniques for intelligent information retrieval and knowledge and meta-knowledge representation, which are crucial aspects for the integration of information over the global network.

An important step in the analysis of gene expression data is the detection of samples or genes with similar expression patterns. The accurate classification of tumours is essential for a successful diagnosis and treatment of cancer. One of the problems associated with cancer tumour classification is the identification of unknown classes using gene expression profiles. Several clustering algorithms have been developed for gene expression data. Also techniques to systematically evaluate the quality of the clusters have been presented. The prediction of the correct number of clusters in a data set is a critical problem in unsupervised classification. Various cluster validity indices have been proposed to measure the quality of clustering results. It is useful not to rely on one single clustering or validation method, but to apply a variety of approaches. Therefore, a combination of these methods may be successfully used for the estimation of the number of clusters. It has been shown that these methods may support the prediction of the optimal partition and computational diagnosis.

We have developed the Machaon Cluster Validation Environment (Machaon CVE) for the application of different clustering and validation algorithms to experiment on gene expression data. This tool may improve the quality of the data analysis results, and may support

the prediction of the number of relevant clusters in the microarray datasets. The major stages of the system can be summarised as follows:

- Clustering. In this step we extract clusters that correspond to the pre-defined number of clusters for a particular dataset. It offers a number of the well-established clustering methods that are available in the literature as well as some recently developed ensemble techniques.
- Validation of clustering techniques. The clustering methods can find a partition in a dataset, based on certain assumptions. Thus, an algorithm may result in different clustering schemes for a dataset assuming different parameter values. Machaon evaluates the results of clustering algorithms based on quality indices and selects the clustering scheme that best fits the data. The definition of these indices is based on two fundamental criteria of clustering quality: cluster compactness and isolation.



A screenshot of the Machaon CVE.

To support biomedical knowledge representation in gene expression data analysis, a new markup language for microarray data clustering and cluster validity assessment has been developed. cluML, a free, open, XML-based format has been designed to address some of the limitations observed in traditional formats, such as inability to store multiple clustering (including biclustering) and validation results within a dataset.

To enhance the predictive reliability and biological relevance of the validation results, a knowledge-driven cluster validity assessment approach for microarray data clustering has been implemented. It consists of validity indices that incorporate similarity knowledge originating from the Gene Ontology (GO), which is a structured, shared vocabulary that allows the annotation of gene products across different model organisms.

The methods performed in this research may bring contribute to the evaluation of clustering outcome and the prediction of optimal cluster partitions. The described estimation approach represents an effective tool to support biomedical knowledge discovery in gene expression data analysis. Despite the fact that Machaon CVE was developed for DNA microarray expression analysis applications, it may be effectively used for clustering and validating of other biomedical and physical data with no limitations.

Link:
http://www.cs.tcd.ie/Nadia.Bolshakova/Machaon.html

Please contact:
Nadia Bolshakova,
Trinity College Dublin/IUC, Ireland
Tel: +353 1 608 3688
E-mail: Nadia.Bolshakova@cs.tcd.ie

Network Visualization in Biomedical Informatics

by Alkiviadis Symeonidis and Ioannis G. Tollis

Since a picture is worth a thousand words, then, most probably a thousand pieces of data can be expressed succinctly by a picture. We describe directions for visualizing data that result from biomedical applications.

The use of computers in biological and medical sciences has lead to great advancements because many activities, such as searches, simulations and data manipulation (just to mention a few), can be performed easier and faster. Since the amount of produced data is huge, the visualization of the data could have significant impact in biomedical informatics. An effective visualization can show details and relational information that was not known before. In the Biomedical Informatics Program at the Institute of Computer Science at FORTH we deal with various visualization problems, including microarrays, gene correlation networks, and patient networks.

There have been many attempts to visualize all kinds of medical information. From the simple structure of compounds up to protein structures, almost everything can be shown on a computer screen. For attributes with clearly defined structure and relationships like DNA the visualization approach is rather straightforward. One can simply draw what one sees. Visualization is a challenging task however, if the data have no clearly defined structure, or when dealing with abstract entities such as regulatory networks of genes. We can detect a metabolic pathway and record the set of reactions involved but we cannot have a visual representation in the laboratory. This is where visualization is

needed most. Given a visual representation of a metabolic pathway a researcher can discover interesting features that are hard to find otherwise. For example we can find the longest sequence of reactions. Another example where visualization may prove useful is when we want to compare two objects such as DNA sequences or proteins.

The importance of visualization is clearer when dealing with abstract entities such as 'correlation' or 'reaction', as in the case of metabolic pathways. All we need is some assumptions. For example we can consider a 'reaction' to be an arrow from one compound to another. Compounds can be visualized as circles or rectangles, labeled or unlabeled. Furthermore, in an image different scales of colors can be used to give more information. We can use different colors to denote for example, how much time a reaction takes.

Another area where visualization is of high importance is when we consider data obtained from a microarray. A microarray can give us the expression of genes of a patient; in fact we use microarrays to obtain the expressions of genes for a set of human samples. Most of the times we know the medical status of the samples and use a microarray to obtain the gene expressions in order to apply some statistical methods on them in order to get results that can be of help for prediction. A visual representation of the information obtained from a microarray is shown in Figure 1. Rows represent genes and columns human samples. The value of each cell is the respective gene expression and the color (green/blue) shows whether the value is positive or negative while the scale of the color reflects the absolute value (the lighter the color, the greater the value).

Researchers use statistical analysis on the above data and obtain useful information. For example, they identify genes with similar behaviour. Furthermore, knowing the medical status of the samples concerning a specific disease, like cancer, one can identify genes with clearly different behavior (expression) in two cases: healthy or unhealthy tissue.

Figure 1: Visual representation of the information of a microarray.

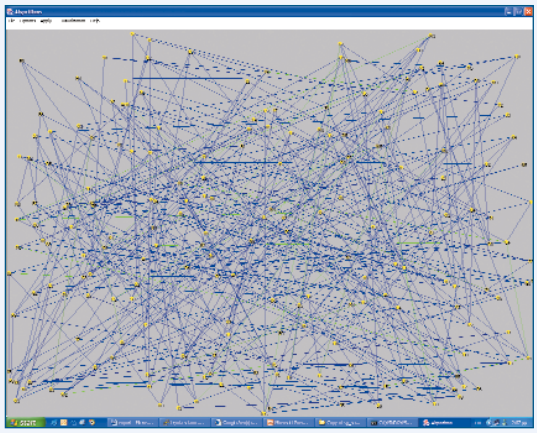


Figure 2: random placement of the information described.

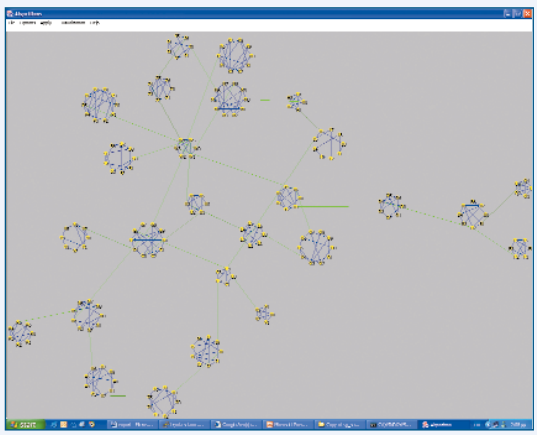


Figure 3: Sophisticated visual representation.

Another interesting problem is to visualize genes according to the relationship to each other, regarding a specific disease. This is done for a set of genes that are typical of a specific disease. In this case, a signal-to-noise procedure is usually used and a respective value is calculated for every gene. During this procedure the expressions of every gene are studied and a value of distinctness between the two categories is assigned to every gene. This method limits the number of genes by selecting only genes with high signal-to– noise score, ie, genes whose expressions can provide information. The next step is to determine genes with similar behavior. This can be done by calculating a correlation coefficient for every pair of genes and group sets of genes that are highly correlated. These groups of genes are shown together in a visual representation. Each gene is represented as a small circle and all the genes belonging to the same group are placed on the periphery of a large circle. Furthermore, we may assume that two genes are highly correlated if their correlation coefficient exceeds some threshold. We can visualize this high correlation by adding a line between two genes. Once again color can be used to show how high the correlation between two genes is. In order to have a clear and comprehensive representation we need to have an picture that is as clear as possible. In order to achieve this, special attention must be paid to the ordering of the genes that appear on the periphery of a circle.

Using this and some other aesthetic aspects we obtain the picture of Figure 3 from the graph (network) that is shown in Figure 2, where the nodes are placed randomly. The green lines join two genes that are not in the same group but are still highly correlated. This hidden information can be shown only using an effective visualization algorithm.

In conclusion we believe that the visual representation of biomedical information is very important. It allows researchers to visualize abstract entities and relationships between them, and understand much better specific functions, even hidden ones, that were previously unnoticed.

Please contact:
Ioannis G. Tollis, ICS-FORTH, Greece
Tel: +30 2810 391671
E-mail: tollis@ics.forth.gr

Dispensation Order Generation for Pyrosequencing

by Mats Carlsson

With the huge increase in the use of DNA technology in fields such as forensic analysis, mutation analysis, antibiotic resistance studies, clinical genetics and pharmacogenetics, it is becoming ever more important to optimize the throughput of DNA sequencing equipment. In a project at SICS, constraint programming was used to optimize the instruction sequence driving DNA sequencing based on the Pyrosequencing principle.

The main application area of Pyrosequencing is the analysis of polymorphic stretches of DNA sequence in the context of stretches of known sequence. The method is based on the principle of sequencing by synthesis. That is, a single strand of DNA is used as a template for synthesizing its complementary strand. The synthesis proceeds by incorporating one nucleotide at a time. In each reaction cycle, a nucleotide is dispensed, ie added to the reaction. There are two possibilities:

- it matches the current nucleotide in the template, it is incorporated into the complementary strand, and a chain of reactions leads to the emission of quantitatively detectable visible light
- otherwise, no incorporation takes place and no light is emitted.

In either case, certain enzymes ensure that any surplus reagents are degraded, making the equipment ready for the next cycle.

Thus, the equipment is driven by a dispensation order, that is, a sequence of instructions. An instruction is one of the DNA nucleotides A, C, G, T. The picture shows how a cyclic dispensation order (A, G, T, C, A, G, T, C, ...) can be used to analyze an unknown sequence. However, if most of the sequence to analyze is known, this is wasteful in terms of reagents and time. By cleverly taking into account the known parts of the sequence, and the known variants of the polymorphic parts, a dispensation order that allows for an optimal throughput of the analysis can be computed.

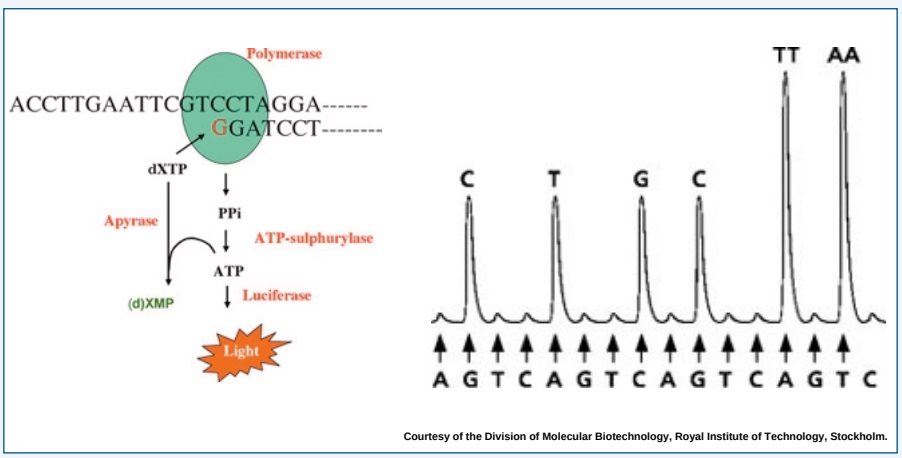
The task of finding such a dispensation order may seem relatively straightforward. However, humans and most other higher organisms are diploid, ie we have all inherited one copy of each gene from our mother and one from our father. These two copies may be identical or may represent different variants in the polymorphic parts. The dispensation order must allow for determining unambiguously and quantitatively what sequence(s) are present in the sample being analyzed.

To further increase throughput of the method, it is often possible to multiplex several Pyrosequencing reactions in the same reaction well. This significantly increases the complexity of the problem of finding an optimal dispensation order, or even a feasible one.

Biotage AB (formerly Pyrosequencing AB) is a Swedish corporation manufacturing Pyrosequencing equipment. A project with SICS was set up, where the task of SICS was to study this problem and come up with an algorithm to produce an optimal dispensation order given a formal description of the sequence(s) to analyze.

From a computer science perspective, this was a clean yet challenging problem. To successfully address this challenge, a host of computational techniques was brought to bear, including logic programming, term rewriting, nogoods, and constraint programming over finite domains.

The technical part of the project was finished in a matter of weeks. Since the final algorithm was far from trivial due to the advanced computer science techniques used, the best method of communicating the algorithm to Biotage became an issue. At the time, Biotage's technical staff included computer science engineers, but they were not specialists in the techniques used in the algorithm. We solved this issue by delivering in addition to the code itself an intensive course in logic programming, in constraint programming, and in the details of the algorithm.



A cyclic dispensation order (A, G, T, C, A, G, T, C, ...) is used to analyze an unknown sequence.

Link:
http://www.sics.se/isl/bio

Please contact:
Mats Carlsson, SICS, Sweden
Tel: +46 18 572361
E-mail: Mats.Carlsson@sics.se

Exploring Genomes with the Self-Organizing Map

by Shaun Mahony
and Aaron Golden

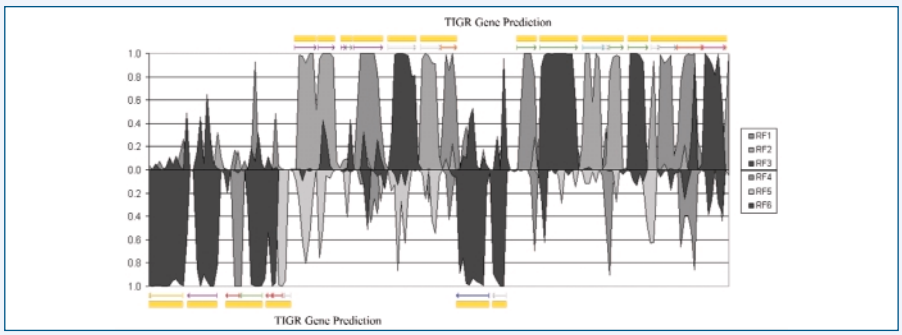


Figure 1: Gene prediction in the B. suis genome (region 450Kbp-475Kbp) using a SOM trained on ORFs of over 1200bp in length. Probability scores for reading frames 4, 5 & 6 were multiplied by –1 for clarity. The yellow bars above and below the graph are regions that TIGR has annotated as being protein-coding regions.

The use of pattern recognition software is not new in the field of bioinformatics, but is perhaps not as developed as it should be. With the growing amounts of data that are being produced by various micro-array technologies and other devices on the one hand, and an appreciation of the fact that there is more to the vast amount of 'non-coding' DNA than meets the eye certainly for H. sapiens on the other, an ability to apply unsupervised algorithms to winnow down the enormous variety of parameter space to a sub-domain that is contextually meaningful is becoming more and more relevant. At the National Centre for Biomedical Engineering Science, National University of Ireland, Galway researchers have been exploring the capabilities of the Self-Organizing Map (SOM) algorithm to this end.

The Self-Organizing Map (SOM) neural network is based around the concept of a lattice of interconnected nodes, each of which contains a model. The models begin as random values, but during the iterative training process they are modified to represent different subsets of the training set. The algorithm effectively performs a map from the high dimensional input vector space to a low dimensional representation whose nodes are characterised by these subsets. Optimum use of the SOM necessitates defining the descriptor that encapsulates the 'pattern' we wish to classify. We have utilised well known descriptors whose representation in the conventional probabilistic frameworks would be difficult if not computationally impracticable - this is one particularly compelling advantage of the SOM algorithm. We highlight two examples of our work to date:

Automatically Generating Multiple Gene Models with RescueNet

The study of codon usage variation in coding as against non-coding regions of DNA has been scrutinised for many years, and whilst there is considerable evidence arguing for clear deviations in usage patterns between the two cases, our actual understanding of the nature of synonymous codon usage is still limited, with the likelihood that more subtle characteristics have yet to be uncovered. Using synonymous codon usage as the descriptor, this SOM variant captures such high dimensional diversity. RescueNet allows the user to identify clusters of genes that have similar codon usage patterns, to identify genes displaying atypical codon usage patterns and perhaps most interestingly, to analyse a contiguous genomic sequence for areas displaying similar codon usage patterns to the major patterns found in a training set, thus making it a valuable annotation tool. In figures 1 and 2 we outline examples of the algorithm's use in these contexts.

Transcription Factor Binding Site Identification with SOMBRERO

Every eukaryotic genome sequenced thus far has shown vast quantities of DNA which do not appear to contain protein-coding regions. Although such non-coding DNA can play important structural roles, much of it also harbors intricate gene regulatory information, including short (6-20 bp) motifs that serve as transcription factor binding sites (TFBS). Cracking the so-called 'cis-regulatory code' has become an important goal in the decoding of genomic data, and an integral part of this challenge is the identification of TFBS. Their identification however is complicated their short size, the inherent degeneracy, and their location ranging by three orders

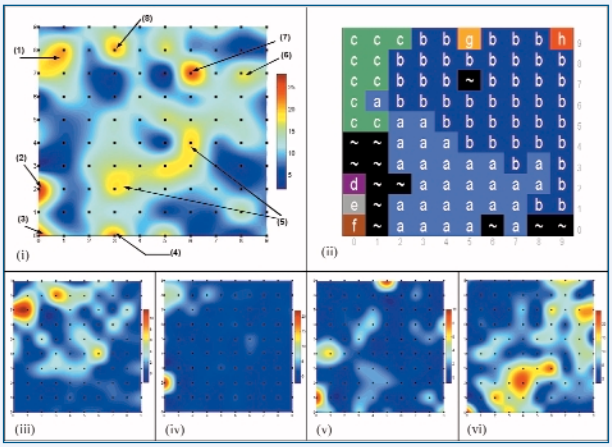


Figure 2: Cluster Analysis in N. meningitidis:
(i) nodes on the output layer responding to all known genes. Legend shows number of times each node responds to the dataset
(ii) groups of similar weight vectors on the output layer; the symbol '~' denotes outlying nodes that are dissimilar to their neighbours
(iii) distribution of the energy metabolism functional group genes
(iv) distribution of the protein synthesis functional group genes
(v) distribution of the cellular processes functional group genes
(vi) distribution of the high scoring hypothetical genes.

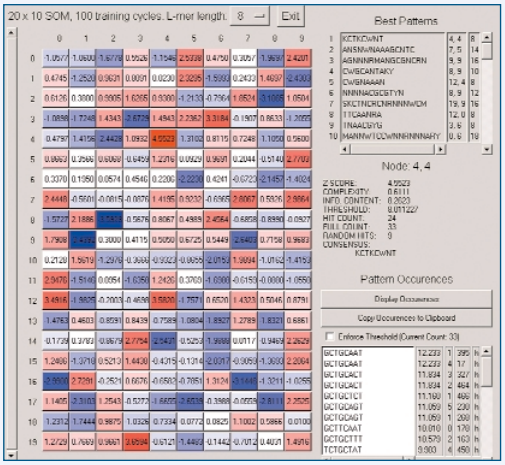


Figure 3: The SOMBRERO results viewer. In this example, SOMBRERO has been trained on genomic sequences from S. cerevisiae that contain binding sites for the transcription factor mcb. Separate SOMs were trained for each even sub-sequence length from 8 to 18, and each SOM is accessible from the results viewer. The SOM shown here is a 20x10 node SOM trained using length 8 subsequences. The SOM nodes are color-coded according to the z-score of the motif contained in the node (red nodes having the most significant motifs). A list of the most significant motifs across all trained SOMs is displayed in the top right hand corner of the results viewer. Information can be displayed for any motif, including a display of each instance of the motif on the input sequences.

of magnitude up and downstream of a given gene's transcription start site. On the plus side, there are many incidents of TFBS classes in the genome, and many correspond to fundamental processes common to us and other species, so hunting in conserved regions of DNA between human and say, mouse, would be expected to boost the signal to noise nature of the data so presented. We have developed a SOM algorithm, SOMBRERO, that uses as its descriptor of TFBS diversity the position weight matrix (PWM), perhaps the most effective way to characterise the degenerate set of a given TFBS class. Our initial experiments have indicated that SOMBRERO yields advantageous performance over more conventional probablistic/statistical mechanical techniques. In Figure 3 we show SOMBRERO-Viewer, which allows one to examine the results of the trained SOM.

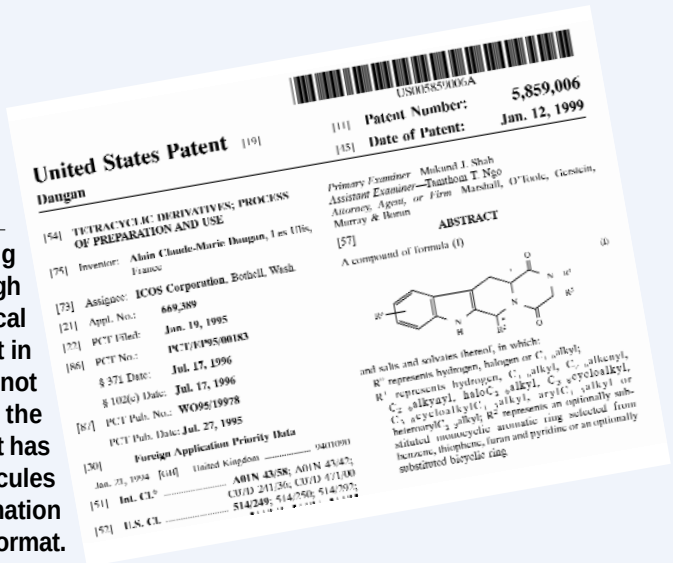
Thus far our work has been devoted to developing working SOM models that are applicable in specific areas and that function on a par with existing probabilistic/statistical algorithms. We aim to go further and to incorporate the SOM algorithms into more contextually defined formalisms, such as the known clustering of TFBS within eukaryotic promoter regions, thus improving the efficacy of the technique. As the data loads facing researchers grow, along with our appreciation of the inherent complexity of regulatory networks we need to decode, we believe appropriate applications of the SOM algorithm will become more and more necessary.

Link:
http://bioinf.nuigalway.ie/shaun.html

Please contact:
Shaun Mahony, National Centre for Biomedical Engineering Science, National University of Ireland, Galway/IUC
Tel: +353 91 512074
E-mail: shaun.mahony@nuigalway.ie

Combating Illiteracy in Chemistry: Towards Computer-Based Chemical Structure Reconstruction

by Marc Zimmermann, Le Thuy Bui Thi and Martin Hofmann

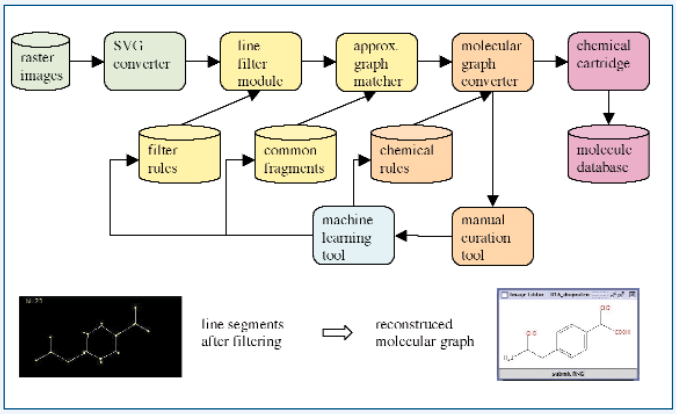


The majority of chemical structure information in the literature (including patents) is present as two-dimensional graphical representations. Although these chemical structures have usually been produced with a chemical drawing program, the machine-readable compound information is lost in the course of publication. Currently, 90% of published structures are not in a chemical file format that can be interpreted by a computer. At the Fraunhofer Institute for Algorithms and Scientific Computing, a project has been initiated to extract chemical information from depictions of molecules in the public literature. The goal of this work is to gather the rich information from pharmaceutical patents and transform it into a machine-readable format.

Our research group in the Department of Bioinformatics at Fraunhofer SCAI is working on the automated extraction of information from biomedical literature. In this highly interdisciplinary domain, interesting information is often presented as a combination of text and graphics. Based on our experience in the field of biological information extraction (eg protein-protein interaction networks), we recently extended the scope of our research towards chemical entity recognition and chemical structure reconstruction.

The information described in patents is extremely valuable to researchers involved in the design of potential drugs.

A patent comprises information on the complete drug-design process, including the disease and indication area, the target protein, the chemical structure of the drug molecules and assay information. Chemical information is often available in printed form or as bitmaps in on-line resources. Reproducing this information by redrawing the structure with a computer program is time-consuming and prone to errors. This process could be greatly improved by a system that was able to read and understand chemical drawings, and could automatically extract the information necessary for a public database.



Software modules for chemical structure reconstruction. First the extracted raster images are converted into line segments (green). These line segments are then processed by a filter module and an approximate graph matcher which identifies chemical fragments, eg a phenol (yellow). All fragments are then combined into a molecular graph using a set of chemical rules (orange).

State-of-the-Art

While the problem of chemical structure recognition and reconstruction is not new, it has not yet been solved. The first attempt at chemical structure recognition resulted in CLiDE (Chemical Literature Data Extraction, developed by the University of Leeds in the beginning of the nineties). We have tested CLiDE in an extensive validation study with drawings of pharma-relevant molecules. The results of the study are based on an assembled test-set of 100 'blockbuster' drug (best-selling drugs, see www.rxlist.com/02top.htm) from 2002. The structures were extracted from the Merck Index and were also manually drawn using Chemdraw (a widely used drawing tool for molecule structures). These 200 files were submitted to CLiDE and the output was manually inspected. Almost 50% had at least one error and often the recognition process was interrupted and asked for manual interference (on average twice per molecule). Unfortunately, there seems to be no further public or commercial effort to improve CLiDE (eg enabling the tool to learn from human intervention).

Scientific Challenge

Following our validation study, we have taken up the challenge to develop a prototype for a new structure recognition tool that would enable true chemical structure reconstruction. In order to overcome the shortcomings of CLiDE, we propose to link modern supervised machine-learning algorithms (eg Support Vector Machines) with cheminformatics similarity searching and reaction-planning algorithms. The following scientific challenges will be addressed:

- conversion of the raster image format into vector graphics
- elimination of 'graphic noise' through a filter module
- development of a graph grammar to handle contours, additional or missing lines or nodes and intersections
- matching of small sub-graphs to a database of consensual chemical fragments
- merging of identified sub-graphs into a single molecular graph based on chemical synthesis rules (ie a chemical grammar)
- training and machine intelligence.

First Steps

For this final challenge, a rule-based approach to the mapping of overlapping fragments from the graph matcher can increase the reliability of the reconstructed molecule. The filter rules, the collection of common fragments and the synthesis rules will initially be trained on a test corpus. Afterwards the whole process can be further improved by automatically retraining each module with the output of the manual curation tool.

In order to solve the problem of recognizing and learning chemical structures in image documents, our system combines pattern recognition techniques with supervised machine-learning concepts. The method is based on the idea of identifying from depictions the most significant fragments of small molecules. We have therefore recursively decomposed a diverse collection of 'blockbuster' molecules into smaller sub-graphs in an off-line process. At runtime, the input image of a chemical structure is converted into a vector graphic. A graph representation of this vector graphic is then generated by combining line segments. In this pre-processing step, the trained system detects and corrects any common errors that occurred during the vectorization procedure or were due to a flawed scanned image. The next step uses sub-graph isomorphism to detect known fragments in the assembled graph. Finally all sub-graphs are combined using a chemical knowledge image (ie proposing chemically suitable corrections) in order to reconstruct the most likely structure represented in the input. If the input graph cannot be matched completely, the user will be asked to specify the solution with a manual curation tool; the solution is then added to the knowledge space of trained molecules. In this way we expect the system to steadily improve its performance.

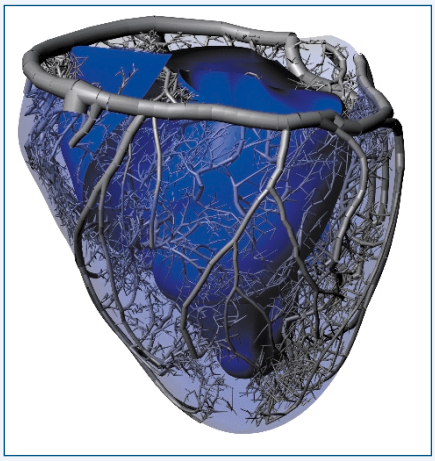
We have already assembled a test corpus of blockbuster molecules and their fragments. We have integrated potrace (http://potrace.sourceforge.net) and autotrace (http://autotrace.sourceforge.net) as SVG converter tools, and the pre-processing and the graph-matching module have been implemented. The next steps are to separate atom labels from chemical bonding lines in the input image and to recognize them. The final step is then to convert the reconstructed graph to a chemical file format.

Please contact:
Marc Zimmermann, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Fraunhofer ICT Group, Germany
E-mail: marc.zimmermann@scai.fraunhofer.de

Integrative Biology — Exploiting e-Science to Combat Fatal Diseases

by Damian Mac Randal, David Gavaghan, David Boyd, Sharon Lloyd, Andrew Simpson and Lakshmi Sastry

Heart disease and cancer are the two biggest diseases, and obviously the focus of intense work within the biomedical community. One aspect of this work is the computer simulation of whole organs based on molecular and cellular level models, and this requires the application of large scale computational and data management resources. The Integrative Biology project, funded by EPSRC, will build a customized Grid framework for running large scale, whole organ HPC simulations, managing a growing database of simulation results and supporting collaborative analysis and visualization.



Whole heart model.

Approximately 60% of the UK methodologies tools and standards to population will die from either heart disease or cancer. Computer simulation of whole organs based on molecular and cellular level models offers the potential to understand better the causes of these conditions and eventually to develop new treatment regimes and drugs to reduce their threat to life. The Integrative Biology (IB) project brings together a team uniquely qualified to tackle this problem, the universities of Oxford, Auckland, Sheffield, Leeds, Birmingham, Nottingham and UCL together with CCLRC and the support of IBM. The project is a second generation UK e-Science project, funded by EPSRC, which will build on the output of first round projects and integrate these and other new developments into a customised Grid framework for running large scale, whole organ HPC simulations, managing a growing database of simulation results and supporting collaborative analysis and visualisation.

Three major cornerstones of the project are the cellular models of cardiac electrophysiology developed over many years by Denis Noble's group at Oxford, the extensive work underpinning computational modelling of the whole heart by Peter Hunter's group in Auckland, and Grid software already developed by several of the partners, particularly CCLRC.

Figure 1 shows a view of a whole heart model. The long term goal driving the project is development of an underpinning theory of biology and biological function capable of sufficiently accurate computational analysis that it can produce clinically useful results.

e-Science Challenges

The e-Science challenges for this project are:

- to provide transparent, co-scheduled access to appropriate combinations of distributed HPC and database resources needed to run coupled multi-scale whole organ simulations
- to exploit these resources efficiently through application of computational steering, workflow, visualisation and other techniques developed in earlier e-Science projects
- to enable globally distributed biomedical researchers to collaboratively control, analyse and visualise simulation results in order to progress the scientific agenda of the project
- to maintain a secure environment for the resources used and information generated by the project without inhibiting scientific collaboration.

Scientific Agenda

The scientific agenda being addressed includes:

- developing integrated whole organ models for some of the most complex biological systems in the clinical and life sciences
- using these models to begin to study the development cycle of cardiac disease and cancer tumours
- bringing together clinical and laboratory data from many sources to evaluate and improve the accuracy of the models
- understanding the fundamental causes of these life-threatening conditions and how to reduce their likelihood of occurrence
- identifying opportunities for intervention at the molecular and cellular level using customised drugs and novel treatment regimes.

Beneficiaries

In both cancer and heart disease, Integrative Biology will improve the design and understanding of new drugs as well as enabling optimisation of novel treatments such as gene therapy or cancer vaccines which might complement conventional cytotoxic drugs. The tools developed by the project will improve the productivity of clinical and physiological researchers in academia and the pharmaceutical and biotechnology sectors. The UK e-Science community will benefit from access to new tools developed by the project and from the example of an integrated computational framework that the project will develop. This will be useful in other areas requiring a total system approach such as understanding environmental change processes. But most importantly, the ultimate beneficiaries will be patients with heart disease, cancer and, eventually, other potentially fatal diseases.

Project Organisation and Software Architecture

The Integrative Biology project team is organised into three main groups charged with developing:

- the modelling and simulation codes
- the computational framework for simulation and interaction
- the security infrastructure required by the project.

Within these groups, cross-institutional teams are working on specific technical areas including heart modelling, cancer modelling, molecular and cellular modelling, testing tuning and running simulations, data management, computational steering, workflow, visualising data and user interfaces.

Portal technology will be used to provide users with a lightweight interface to the Integrative Biology front end services and will support collaborative access to ongoing simulations and results. The services available can be grouped into four main categories:

- job management (including deployment, co-scheduling and workflow management across heterogeneous resources)
- data management (from straightforward results data handling and storage to location and transformation of experimental data for model development and validation)
- computational steering (both interactive for simulation monitoring/control and pre-defined for parameter space searching)
- analysis and visualization (not only of results, but also of interim state, parameter spaces, etc for steering purposes).

Underpinning the entire system are three overriding considerations:

- standardization (in particular OGSA and/or Web Service compliance)
- scalability
- security (covering confidentiality, integrity and accessibility of data and resources).

Many of the underlying components will be adopted from existing projects, and adapted if necessary in collaboration with their original developers. Currently exploited projects include Reality Grid, gViz, myGrid and of course the middleware being developed by the OMII.

Link:
http://www.IntegrativeBiology.ac.uk

Please contact:
Damian Mac Randal, CCLRC,
E-mail: d.f.mac.randal@rl.ac.uk
David Gavaghan, University of Oxford
E-mail: david.gavaghan@comlab.ox.ac.uk

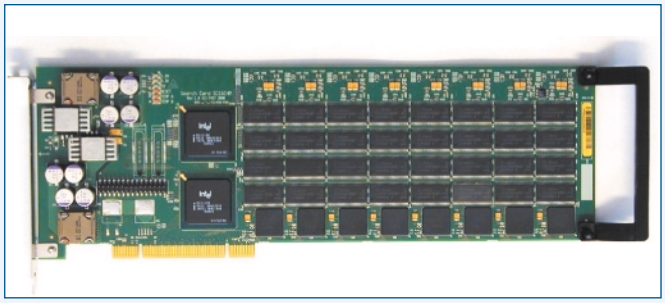
Bioinformatics in the Fast Lane

by Finn Drabløs and Ståle Fjeldstad

By using special-purpose search processors, a standard PC can be accelerated to perform complex pattern-matching at 10 teraoperations per second. This platform is used by the Norwegian University of Science and Technology (NTNU) and Interagon AS for biomedical research.

The flow of data generated by genome-related research is growing exponentially, and this trend is likely to continue. Full genome sequencing is becoming a standard approach even for large genomes. Several research groups are developing novel sequencing techniques, trying to bring the total sequencing cost for a human genome under US$1000. This will open up new opportunities for personalized medicine, with treatment being tailored to specific genetic profiles. Other techniques are also contributing to this data flow, particularly from areas such as the analysis of gene products by proteomics, microarray analysis of gene expression patterns and the mapping of genetic variations, as in Single Nucleotide Polymorphisms (SNPs). At the same time, increasingly complex approaches are being developed to extract essential information from these data. More computing power is therefore needed for such large-scale analysis of genome data.

The Interagon Pattern Matching Chip (PMC) is a special-purpose search processor, capable of searching for complex approximate patterns in arbitrary data. The architecture is massively parallel, thus making it possible to simultaneously search the same data stream with a large number of queries. Query features include regular expressions, with additional functionality such as proximity, adjacency and order conditions, as well as alphanumerical comparisons and approximative matching at both the character and expression levels.



Interagon Pattern Matching Chips on a PCI card.

Sixteen PMC chips, each with its own dedicated memory, are mounted on a PCI-compliant plug-in card. Up to six cards can be inserted in a standard workstation, turning ordinary PCs into high-performance search tools. This enables a single standard PC to perform close to 10 teraoperations (10^{13} operations) per second for pattern-matching purposes. Our in-house PMC-equipped Linux cluster is capable of 80 teraoperations per second.

Novel software tools have been developed in order to make use of this immense computational power for biomedical research. The main software component is based on evolutionary algorithms, where Darwinian principles are used to identify essential

information in large and complex data sets. Important examples are SNP analysis and siRNA design.

An SNP is a genetic variation at a single position (nucleotide) in the genome. This variation may affect gene regulation or gene product properties. The total effect of a large number of variations makes up our genetic 'personality', that is, our genetic disposition to cancer, strokes, adverse drug effects or a long life. Genetic variation alone does not determine an individual's medical history, but it has an influence on their risk of being affected by diseases for which there is a genetic component. However, the correlation between genetic variation and disease risk is complex and difficult to identify. Large data sets with genetic data from both patients and non-affected controls are needed in order to identify significant correlations. We have used PMC technology and evolutionary algorithms on data sets containing several hundred SNP candidates, in order to identify SNP subsets that are associated with specific clinical outcomes. This is an important contribution towards personalized medicine and a better understanding of complex diseases.

Small interfering RNA (siRNA) is a novel technique for the selective blocking of gene expression. It employs a natural defence mechanism against foreign RNA, where short (21-23 nucleotides) double-stranded RNA is incorporated into a silencing complex. This then cleaves messenger RNA (mRNA) with complementarity to the short RNA fragments. Since the mRNA is the intermediate in the synthesis of protein from genomic information, siRNA represents a flexible mechanism for selective silencing of specific genes. This can be used in research, but also has potential as a gene-based therapy. However, it is a prerequisite that the siRNA is designed with high efficacy and specificity, enabling the targeted gene to be selectively knocked down without side effects from non-selective binding to mRNAs from other genes. PMC and genetic programming are used to predict the efficacy of existing siRNA designs, and have shown that many existing siRNA designs may knock down more than one gene. This approach has also been used to design improved siRNAs.

Interagon and the NTNU bioinformatics laboratory are a part of FUGE, a national initiative for functional genomics in Norway. FUGE is coordinated by the Norwegian Research Council. The initiative includes, in addition to bioinformatics, laboratories for proteomics, structural biology, microarray work, biobanks, SNP analysis and molecular imaging.

Link:
http://www.interagon.com/

Please contact:
Finn Drabløs, NTNU
E-mail: finn.drablos@ntnu.no
Ståle Fjeldstad, Interagon AS
E-mail: steel@interagon.com

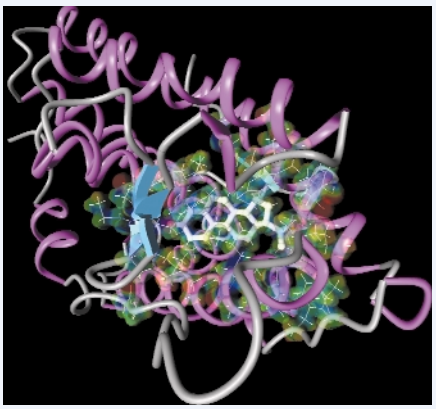
In Silico Virtual Experiments

by Wanda Andreoni

Using massively parallel computers and sophisticated mathematical modeling, scientists at IBM Zurich Research Lab have been able to shed new light on the binding of progesterone with its human receptor. Unraveling how progesterone binds to its human receptor is a first step in understanding that could lead to the development of more effective medication.

Progesterone, 'the hormone of pregnancy', is often the starting point for the design of so-called fertility drugs. For progesterone to have an effect it needs to get into cells. It can either diffuse through the cell membrane or bind to a protein in the cell membrane called a receptor. Controlling, activating, or blocking the action of this hormone amounts to modifying the binding with its receptor. So far, no methodology has been able to show unambigiously how this binding takes place, constituting a real puzzle.

To study intermolecular interactions, the traditional tools are 'in vitro' (in a Petri glass) or 'in vivo' (in a living organism). Our in-silico method is effectively a simulation, based on classical molecular dynamics enhanced with quantum chemistry (quantum-refined-force-field molecular dynamics). The approach enables scientists to study the all-important role of water in the binding process,



Progesterone and its human receptor.

to analyze chemical and physical processes over time, and to show precisely how a given molecule (be it a hormone or a potential drug) and its biomolecular target (be it a protein or DNA) dynamically interact at the level of their component atoms.

The molecular structure of progesterone has two end groups that according to chemical intuition could contribute to the binding to its receptor. Contrary to indications from 'mutagenesis' analysis that confirmed this view, the X-ray structure of the complex has led to the belief that only one end was participating in the binding. This could be neither confirmed or refuted on the basis of traditional modeling.

Our in silico simulations have solved the puzzle and revealed that indeed in a typical aqueous solution both ends of progesterone play a crucial role in the binding in agreement with the results of mutagenesis analysis, and also revealed which parts of the protein (amino-acids) act as primary binding partners. They also showed the highly dynamic nature of the binding process, which is mediated by water molecules present in the pocket. The surprising results obtained from the X-ray structure could thus be ascribed to the sporadic presence of water in the crystal structure and also to the inability to reflect the flexibility of the chemical groups involved. Traditional modeling on the other hand lacked the accuracy that was necessary to represent the interatomic interactions of the real system.

The case of the progesterone-receptor complex is certainly not unique. Virtual experiments based on accurate simulations at the molecular level will increasingly be required to gain insight into the nature and functioning of complex systems like those we find in biology, and may indeed become of critical importance, given the unique information they can provide. As such they have also the potential to enhance rational drug design and thus to become a unique auxiliary tool in the development of pharmaceuticals, which could help reduce the time and expense associated with bringing new drugs to market.

It was fascinating to gain insight into the action of the hormone that determines human birth. Although still in its infancy, we believe that this synergetic approach of high performance computing and accurate methodologies for molecular simulations will lead the path into the unraveling of other mysteries of our life. Our research continues.

Link:
http://www.zurich.ibm.com/deepcomputing/scientific/projects_biochem.html

Please contact:
Wanda Andreoni,
IBM Zurich Research Lab/SARIT
E-mail: and@zurich.ibm.com

Modelling a Living Cell — Mathematics to Model Metabolic Pathways

by Joke Blom and Annette Kik

Can biologists describe a living cell in a computer? Mathematicians of the Centrum voor Wiskunde en Informatica (CWI) cooperate within the Silicon Cell project, where biologists and mathematicians try to reach this goal. Their aim is to relate theory and models closely with data from biological experiments. In future this research could give an indication whether new medication will be effective and how we can keep food good and tasteful as long as possible. This shortens research time and saves expensive biological experiments.

The twentieth century saw many inventions in the frontier areas of life sciences, physics, chemistry, and mathematics. Biological experiments improved: It even became possible to look inside a living cell. Biologists would gladly model its functioning. This is precisely the aim of the Silicon Cell Consortium Amsterdam, a cooperation of the Institute for Molecular Biological Sciences (IMBS) of the Vrije Universiteit in Amsterdam, the Swammerdam Institute for Life Sciences (SILS) and the Section Computational Science (SCS) of the University of Amsterdam, and CWI.

In the living cell, processes of metabolism — metabolic pathways or networks — can be described as partial differential equations (PDEs), where concentrations can vary in space and time: a reaction-diffusion model. There are two lines of research, the qualitative and quantitative one. Qualitative research could describe a process as: 'If cell-size increases it will affect the functioning in such a way' or: 'If the concentration of protein A becomes infinite, then that effect will occur'. In practice the situation is often more balanced and the parameters are limited, because the cell will not survive infinite values.

With quantitative research, realistic parameters for living cells are used. The values for constants and parameters are determined in real biological experiments, for example on metabolic pathways in yeast cells and the E. coli bacterium. One of the recent analytic results – a contribution to the theory of Metabolic Control Analysis - is a new type of control coefficient summation theorem, which relates the control by the membrane transport, the diffusion control, and the size of a cell. Quantitative numerical experiments indicate that diffusion through a membrane in the E. coli bacterium is no limitative factor for the uptake of glucose. However, if the bacterium were ten times larger, the sugar admission from its living surroundings (human intestines) would falter.

Besides reaction-diffusion models for cellular pathways, developmental regulatory networks are studied. CWI develops models for simulating networks that are capable of quantitatively reproducing expression patterns in developmental processes. One example is the embryonic development of fruit flies, where biologists supply the experimental data. Mathematically speaking this research adds to a continuum-discrete hybrid model where discrete, moving and deformable objects, in

which biochemical reactions take place, exchange species with the surrounding environment modelled as a continuum.

For the mathematicians, one challenge is to relate models that work on the different scales that can be measured in biological experiments by now or in the near future: From nanoseconds to weeks and from nanometers to millimeters. Current research at CWI is built around the concepts of simplification and integration. Simplification is essential: a straightforward simulation of a model comprising all the biochemical knowledge is computationally too demanding. Depending on the cellular phenomenon considered, models and methods of appropriate temporal and spatial scales will be developed: ordinary differential equations for simple cells that can be described as homogeneous objects, partial differential equations for moderate spatial and temporal variations, and particle methods for even smaller structures.

Furthermore, techniques for modularisation, model reduction, flexible gridding and flexible time discretization are developed. An effective use of these techniques requires that they are adaptive. Variations in spatial, temporal, and chemical complexity are handled most efficiently if the simplification techniques can be adjusted accordingly. Therefore it is necessary to integrate different model descriptions and numerical approximations into an aggregate simulation without sacrificing reliability. The dynamic heterogeneity of the living cell and the flexibility of the simplification techniques imply that the degree and type of simplification will vary in space and time, thus placing further constraints on the integration.

Another challenge for the mathematicians is to perfect the models to a level that shows in advance whether it is useful to perform a certain experiment. This way, expensive and elaborate biological experiments could perhaps decrease by half.

Links:
http://www.cwi.nl/htbin/pdels/frame?SiC
http://www.siliconcell.net/sica/
http://homepages.cwi.nl/~gollum/

Please contact:
Joke Blom, CWI, The Netherlands
Tel: +31 20 592 4263
E-mail: Joke.Blom@cwi.nl

Integrative Biology at CCLRC

by Daniel Hanlon, Lakshmi Sastry and Kerstin Kleese van Dam

The Integrative Biology project is a collaboration between researchers in diverse disciplines who are applying the developing 'Grid' metaphor to tackle two of medicine's biggest killers — cardiac disease and cancer.

Many physiological models of the heart exist, but they are often restricted in scope by a lack of available computing power. In the case of tumours, a systematic modelling framework is yet to emerge. The biological modelling community is not like that of Particle Physics where the use of high performance computing is commonplace. It is not unheard of for a researcher's laptop to be the most powerful machine on which a simulation is ever run. The relative youth of the physiological simulation community means that collaboration between researchers is not as routinely adopted as it might be - to the detriment of the science.

Integrative Biology (IB) is an inter-disciplinary project that aims to establish an e-Science/Grid based e-Infrastructure for the advancement of biological modelling in general, using, as test cases, studies to understand heart arrhythmia and tumour growth. IB uses simulation codes from teams around the world: Universities of Oxford, Sheffield and Nottingham and overseas from Aukland, San Diego and elsewhere. E-Scientists in the UK from the CCLRC e-Science Centre, Leeds and UCL are working with these numerical physiologists to develop an e-Infrastructure that facilitates the collaborative development of these codes and a secure exchange of research results.

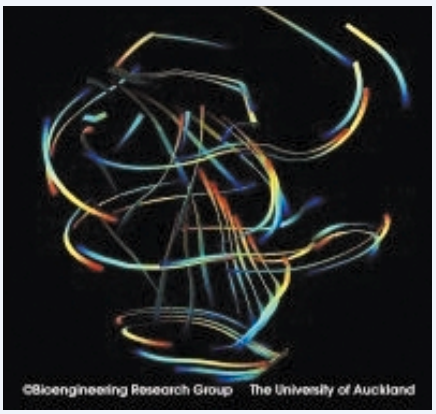


Figure 1: Structural fibres of the heart.

The approach taken by Integrative Biology has two main themes:

- to facilitate the extension of existing physiological codes to cover biological processes on different scales
- to deploy codes within an Integrative Biology grid environment where scientists can experiment with the various physiological models to further their understanding of the systems under examination.

Key to these goals is the creation of a stable middleware environment which is compatible with existing codes but which is not held back from embracing the Grid philosophy. The IB software environment is starting with the synthesis of the proven technology components from a number of existing e-Science projects. This approach aims to maximise code re-use and to fully take advantage of the existing expertise. The grid infrastructure is being built on a number of key pillars:

Security

The Grid Security Infrastructure (http://www.globus.org/security/overvie

w.html) is employed throughout IB. Using the UK e-Science Certification Authority's X509 certificates and this tried and tested architecture, the IB environment gains strong user authentication and encryption capabilities.

Data Management

The Storage Resource Broker (originally developed by the San Diego Supercomputing Centre) forms the basis of the IB data management provision by virtualising file location and providing sophisticated access control mechanisms. Once a file is put into 'SRB space' authorised users can access it via a variety of different client tools from software APIs to web browsers anywhere in the world. File owners retain complete control over who can see or alter their files. A metadata catalogue based on the CCLRC Scientific Metadata Format with extensions from the UK myGrid Project provides the information and annotations necessary to make the data accessible for reuse and sharing with fellow researchers. Finally the CCLRC DataPortal will provide access and search capabilities to the both the data and accompanying annotations.

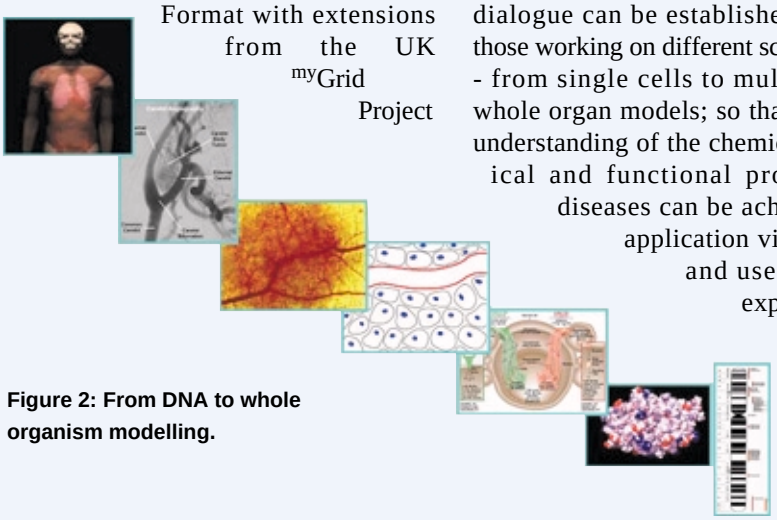


Figure 2: From DNA to whole organism modelling.

Visualisation, Interaction and Computational Steering

Fundamental to the successful analysis and extension of existing models is the scientist's ability to interactively control, or 'steer', simulations as they are executing. This experimentation increases their understanding of the parameters under investigation and provides an opportunity for collaboration with colleagues and peers, promoting the sharing of knowledge and understanding of the models and the results being produced. A useful dialogue can be established between those working on different scale systems - from single cells to multi-cell and whole organ models; so that a holistic understanding of the chemical, biological and functional processes of diseases can be achieved. The application visualization and user interface experts within the IB consortium are focusing on these issues, using a range of commercial and public domain problem solving environments such as Matlab and VTK. The methodologies being employed are based upon those developed for generic visualization on the Grid as part of the CCLRC e-Science Centre's core activity (http://www.e-science.clrc.ac.uk/web/projects/gaptk/) and libraries from the eViz and RealityGrid projects.

Long Term

Current development is aimed at a middleware suite specifically tailored to the IB community. In the future however, the Integrative Biology infrastructure will be deployed as a demonstrator within a more generic Virtual Research Environment. Its functionality will be exposed within the portlet development framework of the Open Grid Computing Environment (http://www.collab-ogce.org). The aim of this interface is to make available the plethora of tools available from a host of disparate scientific disciplines such that the researcher can use any that are appropriate within their own experimental area, accessed with a simple web browser from any networked computer.

Link:
http://www.integrativebiology.ox.ac.uk/

Please contact:
Daniel Hanlon, CCLRC e-Science Centre,
Daresbury Laboratory, UK
Tel.: +44 1925 603683
E-mail: d.hanlon@dl.ac.uk

Applying Complex Models on Genomic Data

by Patrick Durand, Dominique Lavenier, Michel Leborgne, Anne Siegel, Philippe Veber and Jacques Nicolas

The Symbiose team at IRISA-INRIA is involved in large-scale genome studies using complex models. These encompass genome modelling, development of dedicated hardware and gene networks modelling. Symbiose is also in charge of the bioinformatics platform of the OUEST-genopole research structure.

Symbiose is an INRIA project in bioinformatics, and comprises 25 people. Our research includes large-scale genome studies and complex pattern filtering methods. Principal areas of focus are genome modelling with formal languages, development of dedicated machines and gene networks modelling. Applications include the discovery of target proteins, the study of mutations of toxic bacteria such as Staphylococcus aureus, the prediction of disulfide bonds involved in processes such as protein aggregation, high-speed searching of huge databases and signalling of TGF-beta in liver cancer.

More Expressive Models on Sequences

Locating sequences of medical interest in genomic databases can be efficiently

performed using a sequence model (also called a pattern). Such a model describes a region of biological interest characteristic of a family of sequences using various classes of languages such as regular expressions, definite clause grammars and string variable grammars (SVG). These provide a high degree of expressiveness, allowing the modelling of high-order sequence organization (such as copies, palindromes or even structural properties like stem-loop or pseudo-knots). In this context, we have developed STAN, a piece of software that provides the researcher with a SVG-based search system. To perform efficient scans, STAN uses a suffix-tree data structure. Since the usage of SVG syntax is quite difficult, we have also designed ModelDesigner, a graphical software platform that allows the user to create a model graphically, without the need for a programming language.

An Application in Anti-Microbial Discovery

In collaboration with INSERM U435 laboratory, the Symbiose team scanned the human genome to search for new anti-microbial peptides known as defensins. These promising compounds are components of innate immunity that are produced naturally by vertebrates. We first defined a model of the genes in question, by extracting the characteristic signature of genes in this family from known defensin sequences. The human genome was then scanned with the translation of the protein model. The STAN and ModelDesigner software previously described found forty new human defensins, of which thirteen putative defensins have now been synthesized and tested with success. This research is leading to the development of a spin-off biotechnology company.

More Computing Power for Genomic Applications

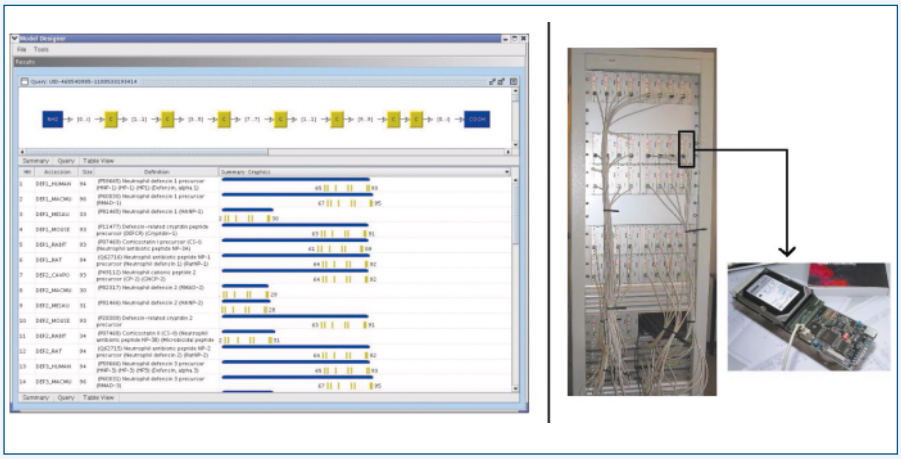
Genomic databases nearly double every year. Applying complex models on this huge data set is a continuous challenge. Our approach mixes both parallelism and specialized computer architecture. We have assembled a 48-node cluster called RDISK (Reconfigurable Devices for Intelligent Storage toward Knowledge-discovery), to perform very fast scanning of genomic databases. Each node houses a hard disk drive tightly connected to reconfigurable hardware (FPGA technology). Depending on the pattern, a custom hardware filter is mapped (in less than one second) to the hard disk output, which allows data to be processed on the fly, with relevant information then being reported.

Discovering Olfactory Receptors

In collaboration with CNRS UMR6061 laboratory, RDISK has been used to scan the draft genome of the dog (36 Gbp consisting of 36 million unassembled sequences) for olfactory receptors. Models created from well-known receptors were translated to automata hard-wired in RDISK nodes. We found a total of 1122 olfactory receptors, including more than 400 new ones. The scan process takes ten minutes on the RDISK cluster. In comparison, it takes two days on a standard Pentium 4 (2.5 GHz) PC.

From Genes to Cells: Dynamics of Biological Networks

Most of a cell's phenotypes result from the intricate interactions of a large number of cellular constituents. Modelling the dynamics of such systems is a challenging issue, since the lack of quantitative data means a qualitative modelling approach is required. Qualitative differential models appear to be appropriate for studying genetically regulated metabolisms such as lipid metabolism: it has been proven that this system changes its functioning mode by equilibrium shifts instead of the multi-stationary equilibrium switches that usually appear in genetic networks. As a complementary approach, we study the combinatorics of incoming signals in discrete event networks. For example, TGF-_ is a transcription factor implied in the development of fibrosis, often leading to liver cancer. In collaboration with the INSERM U456 laboratory, we are identifying signals that spread through the TGF-_ signalling network and lead a cell to a given state. These two approaches are embedded in a graphical analyser tool called GARMeN.



On the left, an example of using ModelDesigner to search the UniProt-SwissProt 3.0 database for a well-known defensin pattern. The pattern is taken from the Prosite database of patterns (entry PS00269). On the right, the RDISK cluster and a detail of a node.

A Stimulating Research Environment

All our collaborations benefit from the OUEST-genopole structure, part of a French Government initiative in favour of genomics and post-genomics studies. Founded in January 2002 in the west of France, it represents a network of more than fifty laboratories with experience in marine science, agri-business and human health. It offers technological platforms and is linked to a cancéropole structure. Symbiose is in charge of the bioinformatics platform (http://genouest.org/).

Link:
http://www.irisa.fr/symbiose

Please contact:
Jacques Nicolas, IRISA-INRIA Rennes, France
Tel: +33 2 9984 7312
E-mail: Jacques.Nicolas@inria.fr

BAIT: Bacteria – Antibiotic Interaction Tool

by Grainne Kerr

ALIFE is the name given to the discipline that studies natural life by attempting to recreate biological phenomena using computers and other artificial media. Through the study of the simplest organisms we are able to gain fundamental knowledge about the underlying mechanisms of life. BAIT (Bacteria Antibiotic Interaction Tool) was developed to provide a simulation tool for such a study. Using only simple rules governing the behaviour of individual bacteria cells with each other and with the environment population behaviour can be investigated.

Complex agent based systems consist of many similar and simple components. The system as a whole often has complex behaviour, which is more than the sum of its constituent parts. Agents can be used to conceptualise and implement such a software modelling application. The fundamental unit of bacteria life is the cell; it therefore seems appropriate to model each bacteria cell as an agent. BAIT was developed to provide a platform whereby the underlying mechanism governing bacteria growth could be investigated. Using the agent based approach the bacteria can be treated as individual entities and hence the model is more general and quantitative. Without specifying a global algorithm and using only rules at a local level of each agent, the growth patterns of bacteria will emerge. By specifying rules for bacteria interactions with the environment and obtaining the correct growth curves we can get an understanding of the underlying principals of bacteria behaviour.

Bacteria growth can be divided into four distinct phases – lag phase, exponential phase, stationary phase and death phase. How the bacteria thrive within a particular environment is largely dependant on the classification of the bacteria. Bacteria can be classified under a number of headings. A bacterium can be classified according to optimum growth temperature (Mesophile, Thermophile, psychrophile), optimum ph (acidophile, neurtophile, alkaliphile), gram staining (gram positive, gram negative), motility (high, low) etc. These factors influence how a bacteria strain will grow and survive within a particular environment eg an acidophile bacterium will not thrive in a alkaline environment, not all bacteria are motile.

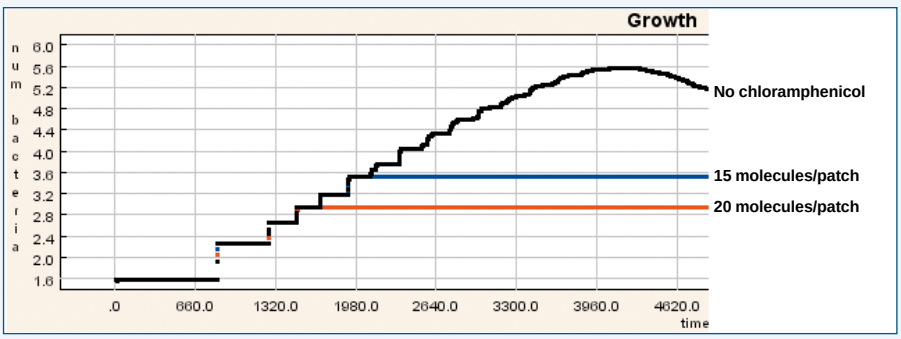


Figure 1: Effects of bacteriostatic antibiotic on bacterial growth on the graphs.

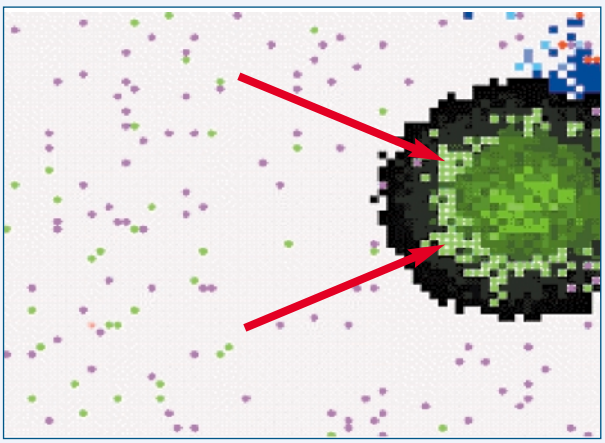


Figure 2: Zone of inhibition, arrows pointing to dead cells (light green), shows a clear area in which no bacterial growth is achievable. Scientists use the size of this zone of inhibition to determine whether a strain of bacteria is sensitive to a particular antibiotic or not and to what degree.

The affect of an antibiotic on a particular bacterium can be classified as either bactericidal or bacteriostatic. A bactericide is a substance that has the ability to kill bacteria. In contrast bacteriostatic antibiotics, while not directly lethal to bacteria, significantly hamper their growth by interfering with processes involved in protein production, DNA production and cellular metabolism. Antibiotics can also be classified according to mode of action and effective range. Effective range refers to the range of bacteria that a particular antibiotic is effective against eg wide spread will effect both gram positive and gram-negative bacteria. Mode of action refers to the method of inhibiting the cells natural working processes. Two main types: cell wall inhibitors and protein synthesis inhibitors.

BAIT was developed using the rules derived from microbial literature. Each bacterium is represented as an individual agent, all having the same potential behaviour and conforming to the same rule principals. Associated with each bacterium is a set of parameters that model certain aspects of the agent such as temperature classification, ph classification, metabolism, motility, minimum energy required for cell maintenance etc. By setting the relevant parameters for the individual agents a range of bacteria

types can be modelled. Implementing as simple rules as possible for the interactions of the agents with each other and with the environment the emergent growth patterns can be observed eg when replication stock reaches a pre defined threshold, divide.

Implementing 'simple rules' for the behaviour of bacteria within an environment growth curves exhibiting the four characteristics stages of microbial growth can be observed. When antibiotic is included in the environment the effects are evident not only on the graphs generated (see Figure 1) but also visually in the virtual world (see Figure 2). In the case of the bactericidal antibiotics in particular, the bacteria manifest themselves in the form of a zone of inhibition around the diffusing mass of antibiotic molecules.

Using the agent based approach the four distinct phases are produced for population growth of bacteria while only specifying bacterial action at a local cell level. When the environment is altered the growth of the bacterial population responds in a manner predicted from in vitro experimentation. This simulation offers a flexible approach to viewing the effects of individual behaviour on bacterial growth and provides an excellent mechanism for understanding the behaviour of individual bacteria. Modelling the effects of antibiotic further advances the simulation. The user can localize the antibiotic in a particular area and allow it to diffuse out into the agar. Using this tool the effect of antibiotic under a wide variety of environmental conditions can the studied.

BAIT was developed at Dublin City University. Contributers are Marian Duggan, Grainne Kerr, Christopher Pender and Ronan Winters.

Please contact:
Grainne Kerr,
Dublin City University/IUC, Ireland
E-mail: grainne.kerr@computing.dcu.ie

Agent-Based Modelling applied to HIV/AIDS

by Ashley Callaghan

Through the use of agent-based computer simulations we hope to model both the spread of HIV/AIDS throughout the population and aspects of the immune response to HIV infection.

The devastating effects of the AIDS epidemic are compounded by its complex patterns of transmission. The rate of transmission and the demographic spread of the disease are influenced not only by direct factors such as age, marriage rates, number of sexual partners, sexual preferences, frequency of extramarital liaisons, etc., but indirectly by factors such as migration patterns, economic conditions, government policies and ultimately by the complex interactions between all these factors. It is these interactions and their evolution over time that provides a major difficulty in trying to predict the spread and social impact of the disease.

The rapid spread of HIV/AIDS throughout the world particularly in areas of sub-Sahara Africa has resulted in a desperate global need for an AIDS vaccine, but to date there has been little success in finding candidate vaccines that stimulate effective neutralising antibodies. In order to develop a viable vaccine the dynamics of the immune response to HIV infection need to be fully understood. In recent years much progress has been achieved in understanding HIV. However the exact mechanisms by which HIV causes AIDS remains unclear. When compared to other viral diseases, HIV infection and the progression to AIDS displays an unusual time evolution (i) initial contact with the virus results in a normal primary immune response, however instead of the virus being completely defeated it remains present in a low concentration; (ii) this phase is followed by the chronic phase. A long period of latency (2-10 years) during which time the CD4 T cell level slowly decreases; (iii) finally when the CD4 T cell count drops to approximately 20%-30% of the normal value, the immune system becomes unable to defend itself its host from opportunistic diseases and the patient generally dies within 2-3 years. If the mechanism that accounts for the slow decrease in CD4 T cells during the latent phase was properly understood it could be of enormous benefit towards designing an effective vaccine.

Agent-based simulations attempt to model the system under consideration by means of interactions between agents at a local level. An agent in an epidemic model would be an individual person, whilst in a model of the immune response an agent would represent an individual cell, whether that cell be a T cell, B Cell, virally infected cell, or any of the other cells that participate in the immune response. The agents then interact with each other based on simple rules. A simple rule in an epidemic model could be of the following form: a single woman will seek a single man

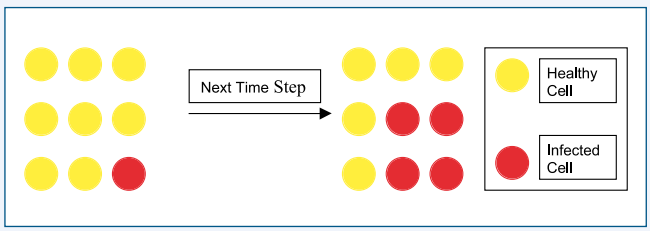


Illustration of a simple rule applied to the immune system, whereby a cell becomes infected at time step t+1 if at least one of its neighbours was infected at time step t. For the purposes of this illustration the immune system is viewed as a 2-D grid. Note that in an actual simulation there would be more than one rule.

within the same age group and then stay together and have children together with a certain probability. This may seem trivial but it allows us to model family structure, something that isn't realistically possible using mathematical techniques. In modelling the immune response to a virus a simple rule could be of the form: if a healthy cell has at least one infected neighbour it too becomes infected (see Figure). Out of such simple local interactions complex global phenomena can emerge.

Until relatively recently large scale simulations such as these weren't computationally feasible, but the advent of faster cheaper machines has lead to an increased level in the use of techniques such as parallel computing with the computational load spread amongst different processors. The processors may all reside on one machine such as a supercomputer or be coupled together in the form of a cluster.

This research is at a very early stage, but we hope to gain insight both into the population dynamics of the epidemic and the immune response to HIV infection. Hopefully this will enable us to determine useful control strategies against the spread of the epidemic, as well as giving us a greater understanding of the immune response. To date we have a simple model that allows us to model such things as family structure, extramarital liaisons, sexual preferences, migration, average lifespan, and different population growth rates. This model is implemented on a dedicated cluster consisting of 17 machines each with four processors. The area being modelled is divided into a number of regions, each of which represents a different geographical area. Each area is then simulated on its own processor with a control program operating on one of the processors to co-ordinate the work and to transfer data between the different processors. Distributing the work among a number of processors enables us to model large populations that wouldn't be possible on a standard desktop machine due to memory and speed constraints. This work will be continuing for the next three years and we hope to have interesting results within the next year.

Please contact:
Ashley Callaghan,
Dublin City University/IUC, Ireland
E-mail: Ashley.Callaghan@computing.dcu.ie

Regulatory Compliance of Pharmaceutical Supply Chains

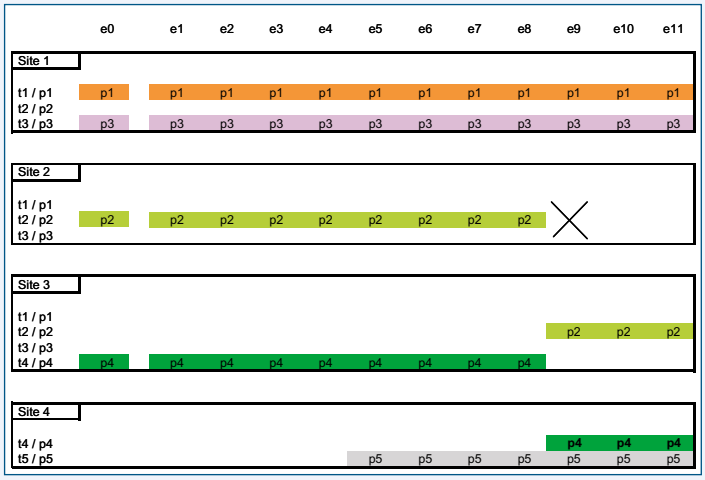
by Eleni Pratsini and Doug Dean

The U.S. Food and Drug Administration (FDA) launched a major initiative to modernize the regulation of drug manufacturing and product quality in 2002. This program calls for the application of modern risk and quality management techniques, the use of engineering knowledge and new manufacturing technologies, and the application and demonstration of cutting-edge science throughout the entire production process. The IBM Zurich Research Lab in collaboration with IBM's Business Consulting Services in Life Sciences have developed a tool to assist pharmaceutical companies quantify their risk exposure to the new regulations and subsequently restructure their supply chains and manufacturing assets to minimize exposure to regulatory risk and at the same time maximize their revenue.

Driven by the increase in the number of adverse events and drug recalls in recent years, FDA has changed its method of monitoring drug manufacturing. They introduced systems thinking, quality by design and related processes that assure the quality of any product in manufacturing. This resulted in a new FDA initiative 'Good Manufacturing Practices (GMP's) for the 21st century'. The initiative requires pharmaceutical companies to comprehensively manage patient risk, base new drug submissions and manufacturing approaches on demonstrable scientific principles, and simultaneously implement inspection-ready 'GMP Systems' which embed compliance and quality in their operations. Of particular significance is the risk transfer among products resulting from the new approach of GMP Systems inspections. All products produced at a facility may be considered 'adulterated' if any one GMP system fails inspection. It is possible that a single low-revenue high-risk product may wipe out an entire facility's revenue. The assets and infrastructure of pharmaceutical companies were designed to meet the 25-year old GMP regulations and are now exposed to the new risks brought up by the FDA programs related to quality by design.

One of the practical implications of the FDA's challenge is the restructuring of a company's supply chain. A company will need to manage its risks across products, technologies, and sites, and ensure that its drugs are safe at the point-of-use. Using a novel statistical approach, we developed a model to quantitatively measure a company's current exposure to compliance hazards for each of the risk sources based on historical performance. It is based on a retrospective analysis of non-conformance issues and their root causes of failure, leading to a view on their exposure to risk. These risk indices are then fed into a mathematical programming model that determines the

sequence of corrective actions the company can take in order to minimize its exposure to risk, minimize operational and action-induced costs and maximize its revenue. The solution is, of course, constrained by budgetary considerations, physical restrictions and a defined rate of change. The problem is represented as a mathematical model with the combination and interaction of risk indices resulting in a nonconvex, nonlinear integer formulation. Through variable reformulation and logarithmic transformation, a mixed integer convex nonlinear model was obtained which was then linearized. Typical corrective actions are risk mitigation for any of the sources, closing down a technology or site, delisting a product, introducing a new technology, and others. Both company-internal and company external actions can be considered.



Example of supply chain restructuring for minimization of risk exposure. Product p2 is a risky product using an aging technology (t2) on site 2. In period e8 the risk of this technology is high enough to force the product to move to site 3 where a less risky technology can be used. In order to avoid 'adulteration', product p4 moves to another site.

Based on examples generated by real data, the model was tested and the output analyzed. The risk interactions generated solutions that were not intuitively obvious. The sequence and timing of actions depended on budgetary as well as risk considerations. Products were moved to other technologies if their technology was aging and thus became too risky. In certain instances products were moved to other sites which in turn exposed all products on the new site to new levels of risk. This 'adulteration'produced the subsequent transfer of other products to other sites (see Figure for an example). By considering the risk transfer in the optimization model, a company can reduce its exposure to business risk by 40% while increasing its secured profitability by 30%.

The model can be used as a simulation engine for testing various alternative solutions, or for partial optimization when the end state is fixed but the sequence of actions to get to that state is optimized.

The mathematical model has a large number of binary variables resulting in an NP hard problem. The complexity of the optimization model is due in part to the infinite combination of corrective actions that can take place. Furthermore, considering multiple periods in a planning horizon further complicates the model. Current research investigates heuristic procedures for obtaining near optimal solutions for very large problems in minimal computation time.

Please contact:
Eleni Pratsini
IBM Zurich Research Lab./SARIT, Switzerland
E-mail: pra@zurich.ibm.com

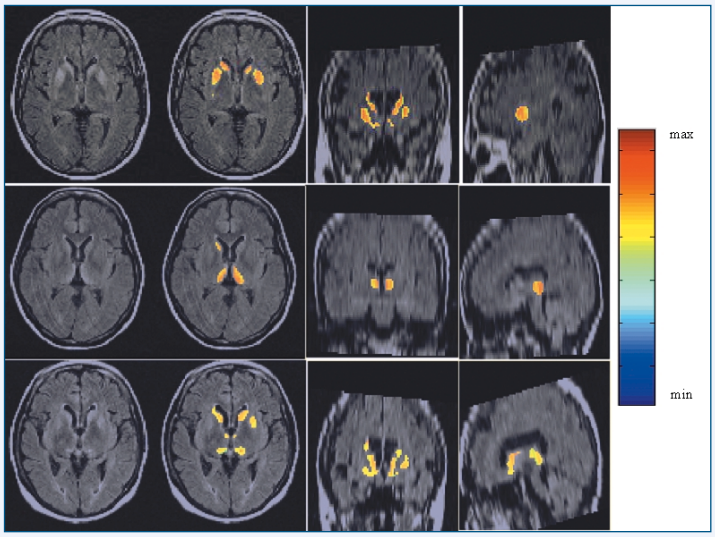
Doug Dean, IBM Business Consulting Services, Switzerland
E-mail: doug.dean@ch.ibm.com

Definition and Evaluation of MRI-Based Measures for the Neuroradiological Investigation of Creutzfeld-Jakob Diseases

by Marius George Linguraru and Nicholas Ayache

The identification of diagnosis markers is a major challenge in the clinical care of patients with Creutzfeldt-Jakob Disease (CJD). A study was carried out as part of the French 'GIS-Prions' project to detect and quantify CJD-specific abnormal intensities in Magnetic Resonance Imaging of the brain.

Of great concern has been the occurrence in the United Kingdom in the 1990s of variant CJD (vCJD), a form of human environmentally acquired CJD, also known as 'mad cow disease'. Although the number of vCJD cases has decreased since 2001, a new risk exists of transmission by blood transfusion. The disease raises a number of challenges for neuroradiological centres, since there exists only limited knowledge of how it might be detected through medical imaging. We have studied the detection and quantification of CJD-specific abnormal intensities in magnetic resonance imaging (MRI) of the brain in order to diagnose CJD patients and differentiate vCJD from sporadic CJD (sCJD) cases. MRI is commonly used for non-invasive examinations of patients with neurological diseases. Some recent studies have found strong correspondences between the diagnosis of CJD and the detection through MRI of signal abnormality in the deep grey matter internal nuclei of the brain. However, observations describing the ability of MRI to assist in the diag-



Results on patient data. We present on each row on the left an axial cross-section of the FLAIR MR data with abnormal hyperintensities in the internal nuclei; next to it we have the CJD detection map with corresponding intensities, as seen in the attached colourmap; further to the right, we present a coronal cross-section and a sagittal coronal cross-section with their detection maps. The top row shows a typical sporadic case with stronger hyperintensities in the head of caudate and putamen (the upper nuclei in the axial view). The middle row shows a typical variant case with stronger hyperintensities in the thalamus(the lower nuclei in the axial view). The bottom row displays an ambiguous case, which was difficult to interpret visually by the clinician, as it has strong hyperintensities in all internal nuclei; our algorithm classified it correctly as vCJD.

nosis of CJD are at an early stage, and the visual interpretation of MR images performed by physicians can be challenging. At present, MRI is not included as a diagnosis criterion for sCJD, but it would be certainly be useful for this and for vCJD. It is therefore necessary to further explore the advantages of Computer Aided Diagnosis (CAD) techniques in the MRI clinical environment.

Our method used analysis of deep grey internal nuclei of the brain (namely the head of caudate, putamen and thalamus) to accurately detect human spongiform encephalopathy in multisequence MRI of the brain. T1, T2 and FLAIR-T2 MR sequences were used for the detection of intensity deviations in the internal nuclei. The use of a priori anatomical knowledge in the form of an accurately segmented and labelled atlas facilitated the precise segmentation, while a probabilistic atlas allowed intra- and inter-patient analysis. A feature detection technique based on a model of the human visual system was employed for the depiction of hypersignals. The varieties of human prion disease (sCJD vs vCJD) were differentiated using newly defined MR measures based on the lesions' topographical distribution.

Our database comprised fifteen CJD cases (ten sCJD and five vCJD) and eight healthy controls of the same age range as the patients. All patients showed abnormal intensities in the deep grey nuclei, which were correctly detected by our algorithm. We diagnosed all fifteen prion disease cases with no false positives amongst the controls. The results are robust over the patient data and in accordance with the clinical ground truth.

The caudate nuclei are highlighted as the main area of diagnosis in sCJD, in agreement with the histological data. The algorithm permitted the classification of abnormal signal intensities in sCJD patient FLAIR images with a more significant hypersignal in the caudate nuclei (10/10) and putamen (6/10) than in the thalami. In vCJD patients, we found more significant hyperintensities in the pulvinar than in the other internal nuclei, which confirmed the visually based radiological observations related to CJD. Defining normalized MRI measures of the intensity relations between the internal grey nuclei of patients, we differentiated without ambiguity all CJD cases (sCJD and vCJD) from healthy controls and further classified the CJD patients into two subgroups, sporadic and variant. This is to our knowledge the first attempt towards an automatic classification tool of human spongiform encephalopathies.

The algorithm also allowed the study of asymmetries in CJD MR hypersignals, which has long been a subject of debate by neuropathologists. Using brain internal nuclei masks, we also noted that hypersignals are inhomogeneous over the nuclei.

Our method proved as reliable as the visual interpretation of radiologists for the detection of deep grey internal nuclei hypersignals in MRI of the brain. Moreover, it allowed quantitative data to be automatically obtained from MR patients with CJD, which could be used to follow up patients and evaluate the efficiency of therapeutic procedures. Our study demonstrates the value of MRI as a potential non-invasive diagnostic tool for sCJD and for the characterization of prion diseases.

This work was completed as part of GIS-Prions, a project funded by the French Ministry of Health. The collaboration included the EPIDAURE Research Group at INRIA-Sophia Antipolis, CNRS UPR640-LENA, Paris, the Department of Neuroradiology, La Pitié-Salpêtrière Hospital, Paris, CRMBM UMR CNRS 6612, Marseille, INSERM U360, Paris, and R. Escourolle Neuropathological Laboratory, Paris. The goal was to perform a prospective study of sCJD and vCJD and to develop techniques for the detection and classification of various types of CJD. The database includes CJD patients from two main neuroradiological centres in Paris and Marseille in the period from 2002 to 2004. Contributing members include Miguel Ángel González Ballester, Eric Bardinet, Damien Galanaud, Stéphane Haïk, Baptiste Faucheux, Patrick Cozzone, Didier Dormont and Jean-Philippe Brandel.

Please contact:
Marius George Linguraru, Nicholas Ayache, INRIA, France
Tel: +33 492 38 76 60
E-mail: Marius.Linguraru@sophia.inria.fr, Nicholas.Ayache@sophia.inria.fr

Biomedical Imaging for Enhanced Genetic Data Analysis

by Thanasis Margaritis, Kostas Marias, Manolis Tsiknakis and Dimitris Kafetzopoulos

The Institute of Molecular Biology and Biotechnology in collaboration with the Institute of Computer Science - FORTH, have initiated several joint efforts towards the development and implementation of advanced techniques for biological data analysis and management. One of the most promising research areas is biomedical imaging, a multidisciplinary field aiming at effective analysis and processing of biological/genetic data. It can also be considered as one of the emerging areas in biomedical informatics, emanating from synergies amongst medical informatics, medical imaging and biology. The aim of this field is to build on previous engineering knowledge, in order to develop robust analysis frameworks for maximizing the information content of biological data.

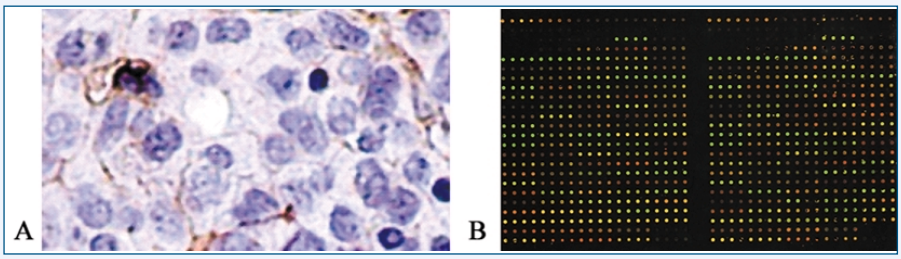


Figure 1: A: ImmunoHistoChemistry of HER2 protein (brown color) in mammary tumor; B: A Microarray imaging experiment showing differential gene expression.

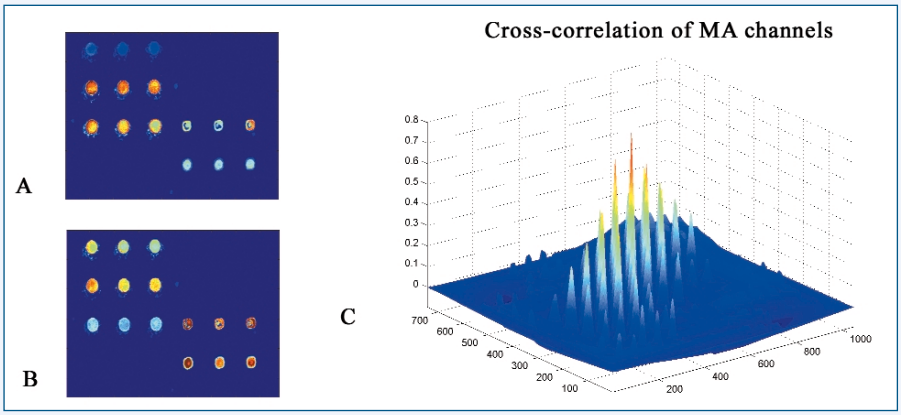


Figure 2: Registration of different microarray channels (A,B) using cross-correlation (C).

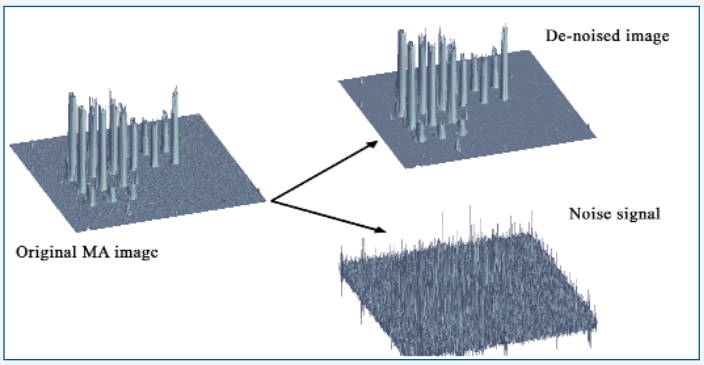


Figure 3: Wavelet decomposition of the original microarray image (left) to the denoised image and the noise image.

Traditionally, medical imaging has focused in providing anatomical information, mainly imaging human bones, dense tissue and arteries. Several advances, especially positron emission tomography (PET) and functional magnetic resonance imaging (MRI), allowed the study of various pathophysiological processes via radiolabeled tracers (PET) or pharmacokinetic models in contrast enhanced MRI. Current gamma camera, PET, MRI, and optical imaging paradigms have been demonstrated for monitoring molecular-genetic processes (via direct and indirect approaches), rather than anatomy. At the same time, newly developed imaging techniques provide very important information regarding gene and protein expression and function. Some of the most interesting techniques used include:

- Microarray imaging: An array of DNA reporters is hybridized with labeled probes to study patterns of gene expression. It is based on confocal laser microscopy.
- Fluorescence in situ hybridization (FISH) techniques: A method for determining the cytogenetic location of a cloned fluorescent-labeled segment of DNA. The site of chromosomal DNA-probe hybridization is determined by fluorescent microscopy.
- Spectral karyotyping (SKY) or multiplex fluorescence in situ hybridization (M-FISH) of human chromosomes.
- ImmunoHistoChemistry (IHC): A method of detecting the presence of specific proteins in cells or tissues.

Here we describe several common analysis issues of microarray, FISH and IHC image-data analysis. In particular, microarray analysis is one of the first domains that clearly addressed the need to develop synergies that extended beyond the limits of previous fields, such as medical imaging and bioinformatics. Thus, gene expression imaging can be regarded as a biomedical imaging discipline that deals mainly with the following data analysis problems:

Registration: This is a typical problem for microarray imaging, as well as in M-FISH and IHC imaging, because images are collected in more than one channel. Due to the differences in the optical

paths of color channels in acquisition of images and the inherent 'chromatic aberration' of the optical system, the same object in the specimen changes its off-axis position when imaged with different wavelengths. In order to align multi-spectral images a plethora of 'classic' algorithms is available. We have used an image similarity approach based on similarity measures (eg cross-correlation) in order to geometrically align microarray images(see Figure 2).

Segmentation: It is particularly crucial to be able to define the regions of interest both in microarray imaging and in FISH/M-FISH/IHC. In FISH imaging it is crucial to segment dot signals that are used then to quantify the gene duplication or deletions caused by chromosome abnormalities. Similarly, it is important to segment microarray spots from the background in order to define the relative statistical distributions of values and perform normalization. Popular methods derived from pattern recognition/neural networks include k-mean, fuzzy clustering and Markov random fields. We have used wavelet filtering as segmentation pre-processing step by separating the signal from noise (see Figure 3).

Image Feature Quantification: Following image preprocessing steps, it is crucial to derive quantitative information related to gene expression/function. In this context, important features or morphology in images need to be extracted and quantified. In an interphase FISH-labeled cell, a number of features are measured to verify if FISH dot signals are well separated dots, overlapping dots, contiguous or spread signals, or split signal dots. This quantitative information can be related to the characterizations of gene duplication or deletions, associated with several genetic diseases (eg neuropathy). On the other hand, in microarray image analysis it is important to define the actual quantitative relationship of two information channels (eg healthy tissue RNA labeled with Cy3 and a diseased tissue RNA labeled with Cy5). This way, differential expression of genes can be detected on the basis of 'differing' from the overall trend that characterizes the relationship of gene expressions. This is illustrated in Figure 4, where a LOWESS fit has been used to estimate the overall trend of intensity pairs in a microarray experiment. This way subtle differences that can be potentially attributed to differential gene-expression are preserved (see encircled regions).

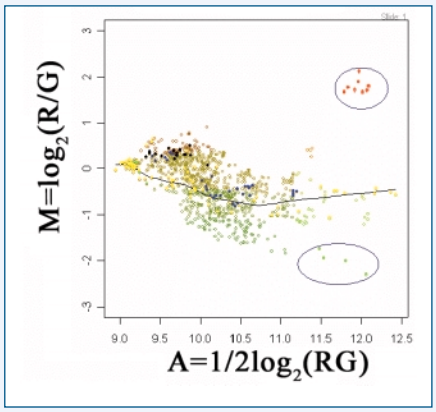


Figure 4: Plot for microarray normalization. It is crucial to estimate the overall trend using methods that preserve outliers (eg lowess).

We presented some basic research directions in biomedical imaging, with special focus in our work on imaging of gene-expression. By developing synergies within FORTH, we aim to expand the present capabilities of biological data analysis by applying modern imaging and computational algorithms, inspired from a deep understanding of the underlying biological phenomena.

Please contact:
Dimitris Kafetzopoulos
and Manolis Tsiknakis, FORTH, Greece
Tel: +30 2810 391594
E-mail: kafetzo@imbb.forth.gr,
tsiknaki@ics.forth.gr

Virtual Tissue Matrix: A Pathologist Aid in Tissue Microarray Analysis

by Catherine M. Conway, Graham Dodrill, Darragh Lawler and Daniel G. O'Shea

Advancements in telepathology have resulted in the ability to review image of glass slides on-line. The development of Tissue Microarrays in the late 1990's, has significantly increased the throughput of biopsies analysed using immunohistochemistry, compared to more traditional methods. The Virtual Tissue Matrix was created to combine these two technologies, to allow virtual review of Tissue Microarray slides, and the recording of review data generated in a relational database.

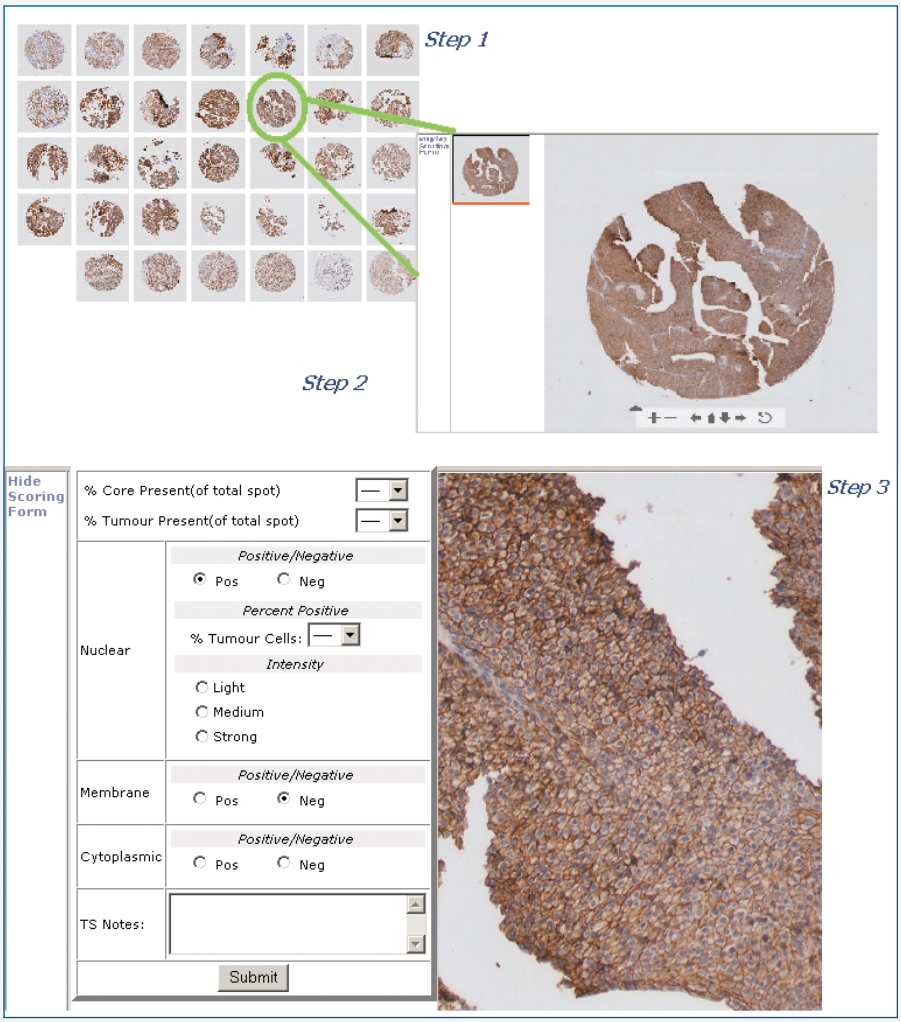
Tissue Microarray's (TMA's) are glass slides, which permit simultaneous large-scale analysis of small tissue samples at DNA, RNA and protein expression. TMA's are manufactured by extracting cores, from donor tissue blocks. These cores are transferred into a TMA recipient block, which can house hundreds of cores from multiple donors. The TMA recipient blocks are then sectioned, and are transferred onto glass slides. One recipient block can create multiple glass slides, all of which have tissue spots arranged on the surface of the glass. The throughput of biopsy tissue reviewed, is greatly increased using TMA's, compared to more traditional methods.

TMA review involves manual examination of hundreds of tissue spots per slide, ranging in diameter from 0.6mm to 2mm

per spot, using a conventional microscope. This process is labour intensive, and requires large amounts of time and concentration. There are a number of problems associated with glass slide reviews. Under a microscope, it is difficult to track which spot is under examination, at a given time. Pathologists can sometimes confuse their position on the slide, and record results associated with an incorrect spot. It is common practice for results to be handwritten at the time of the review, and subsequently entered into an Excel spreadsheet. Retrieval of results from these flat file spreadsheets can be difficult, due to the object oriented nature of the data. Technological developments in the area of telepathology have resulted in, reviews of histopathology slides being performed on line. This liberates the pathologist from the microscope, and also reduces the degree of human error associated with the scoring of Tissue MicroArrays using conventional microscopy.

Virtual Tissue Matrix (VTM) is a web application that allows, the review of TMA slides at remote locations. It consists of a user interface that provides images for review, and a relational database to store the generated results. The Medical Informatics Group, Dublin City University, first created the VTM in 2003.

Once a user has successfully logged into the site, they are presented with a list of TMA recipient blocks available for review. On selecting one, a list of the associated TMA slides are presented. Once a TMA slide has been selected, an overview image of the slide appears. A larger image of each spot can be obtained by clicking on the image. The user can then scroll around the image, and change magnification as needed. Once a user is ready to record results for a spot, they click on an option button, which displays a pre-defined scoring form. The scoring page is a drop down form, which hovers over the image. It allows the user to record their results and submit the data to a relational database. The interface is designed to minimise data entry. The coordinating pathologist completes a form at the start of a study, which details the information they want to record. This information creates the pre-defined scoring form for each study.



Step 1: Overview image of TMA Slide.
Step 2: Spot selected, image is enlarged.
Step 3: Image has been zoomed in; user has selected the scoring option.

The VTM was created using PHP, JavaScript, and an Oracle database. In order to allow the user to change the image magnification, the VTM uses a tool called Zoomifyer EZ. This allows high-resolution images to be viewed on a web page.

Future activities of the VTM will include the incorporation of an analysis package, which will assist in a pathologist interpretation of recorded data. Our aim is to have a tool that can run concurrently with the reviewing process. That will interpret trends and patterns in the results, and will highlight these areas of interest to the pathologists.

In order to validate the VTM, extensive testing of the system needs to take place. This will be done in conjunction with Beaumont Hospital Pathology Department and the Conway Institute, UCD.

Link:
http://www.telepathology.dcu.ie/VTM
To log into the site use the username "ERCIM", password "ERCIM"

Please contact:
Catherine Conway,
Dublin City University, Ireland
Tel: +353 1 5005281
Tel: catherine.conway3@mail.dcu.ie

Spatio-Temporal Analysis in 4D Video-Microscopy

by Charles Kervrann, Jérôme Boulanger and Patrick Bouthemy

Recent progresses in cellular, molecular biology and light microscopy make possible the acquisition of multidimensional data (3D+time) and the observation of dynamics of fast cellular activities. Image processing methods for quantitative analysis of these massive movements have been limited. Automatic techniques to extract information about dynamics from image sequences are therefore of major interest, for instance, to assess the role of Rab GTPases, a large gene family, involved in membrane transport. We have recently developed methods to perform the computational analysis of 4D image sequences.

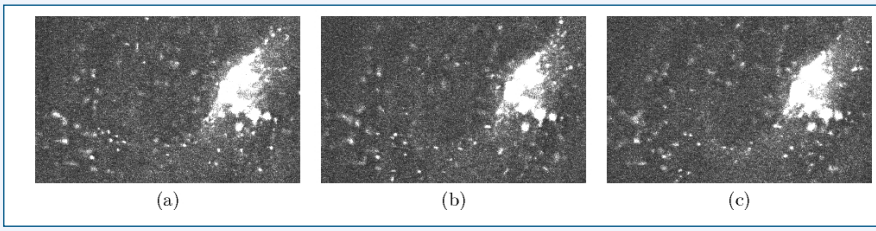


Figure 1: Three images showing the spot positions at time t = 5s, t = 10s and t = 15s. the dynamic of fluorescent tags was recorded by fast 4D deconvolution microscopy. The large white region depicts the Golgi apparatus while small spots are vesicles moving with a high average speed (~ 10 pixel/frame).

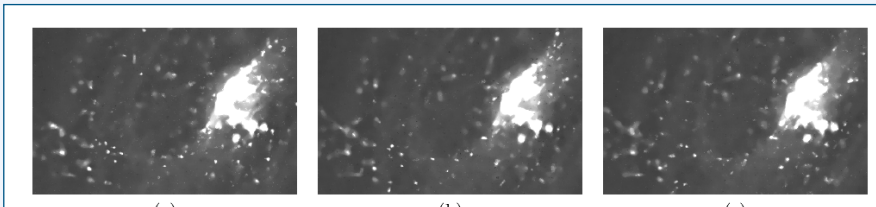


Figure 2: Three images extracted from the restored sequence using the spatio-temporal window adaptive estimation approach (Figure 1). We can see that discontinuities are well preserved while flat regions are drastically smoothed.

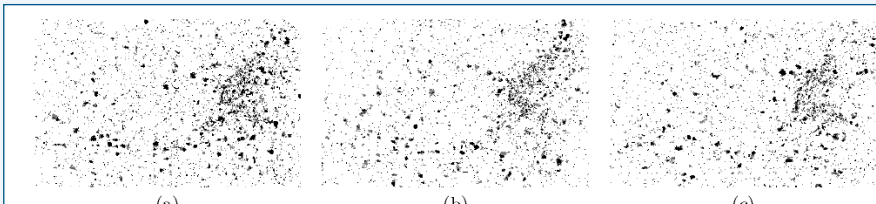


Figure 3: Three images depicting the size of spatio-temporal estimation window considered at each pixel. Black values represent small windows which are associated with spatio-temporal discontinuities into the space-time signal while white values represent larger windows used to estimate nearly constant regions.

This project aims at extracting major motion components by statistical learning and spatial statistics from the computed partial or complete trajectories. We mainly focus on the analysis of vesicles that deliver cellular components to appropriate places within cells. Applications of the proposed image processing methods to biological questions should provide a new and quantitative way for interpreting motility of membrane transport vesicles. The challenge is to track Green Fluorescent Protein (GFP) tags with high precision in movies representing several gigabytes of image data and collected and processed automatically to generate information on complex trajectories. Quantitative analysis of data obtained by fast 4D deconvolution microscopy allows to enlighten the role of specific Rab proteins. The role of Rab proteins is viewed as to organize membrane platforms serving for protein complexes to act at the required site within the cell. Methods have been developed for a target protein - Rab6a' - involved in the regulation of transport from the Golgi apparatus to the endoplasmic reticulum. Typically, the state of Golgi membranes during mitosis is controversial, and the role of Golgi-intersecting traffic in Golgi inheritance is unclear. In Figure 1, three types of objects are observed within the cell: 1) organelle (Golgi), 2) rapid transport intermediates (moving vesicles) and 3) elongated structures (membrane tubulations).

Work began with an analysis of possible methods to improve the quality of images and their adaptation to 4D imaging. We present a spatio-temporal filtering method for significantly increasing the signal-to-noise ratio in noisy fluorescence microscopic image sequences where small particles have to be tracked from frame to frame. New video-microscopy technologies allow to acquire 4-D data that require the development and implementation of specific image processing methods to preserve details and discontinuities in both the three x-y-z spatial dimensions and the time dimension. Particles motion in such noisy image sequences cannot be reliably calculated since objects are small and untextured with variable velocities; the S/R ratio is also quite low due to the relatively limited amount of light. However, partial trajectories of objects are line-like structures in the spatio-temporal x-y-z-t domain. Image restoration can be then achieved by an adaptive window approach which has been already used to efficiently remove noise in still images and to preserve spatial discontinuities, for image decomposition into noise, texture and piecewise smooth components. The proposed 'adaptive window approach' is conceptually very simple being based on the key idea of estimating a locally regression function with an adaptive choice of the space-time window size (neighbourhood) for

which the applied model fits the data well. We use statistical 4-D data-driven criteria for automatically choosing the size of the adaptive growing neighbourhood. At each pixel, we estimate the regression function by iteratively growing a space-time window and adaptively weighting input data to achieve an optimal compromise between the bias and variance. The proposed algorithm complexity is actually controlled by simply restricting the size of the larger window and setting the window growing factor. Global statistics such as velocities and directions could be then extracted in order to measure the general behaviour of vesicles in the 4D volume.

We have applied this method to noisy synthetic and real 4-D images where a large number of small fluorescently-labeled vesicles move in regions close to the Golgi apparatus. The S/R ratio is shown to be drastically improved resulting enhanced objects which even can be segmented. The objective is to report evidences about the lifetime kinetics of specific Rabs in different membranes may be similar within on cell type. This novel approach can be further used for biological studies where dynamics have to be analyzed in molecular and subcellular bioimaging.

This work was performed in collaboration with the MIA Unit (Mathématiques et Informatique Appliquées) from INRA, Jouy-en-Josas, Curie Institute - UMR 144 - CNRS ('Compartimentation et Dynamique Cellulaires' Laboratory), Paris and UMR 6026 ('Interactions Cellulaires et Moléculaires' Laboratory – 'Structure et Dynamique des Macromolécules' team), Rennes. The Vista team is the prime contractor of this project (MODYNCELL5D).

A second applicative project will concern the CLIP 170 protein involved in the kinetochores anchorage (in the segregation of chromosomes to daughter cells, the chromosomes appear to be pulled via a so-called kinetochore attached to chromosome centromeres) using new fluorescent probes (Quantum Dots). New image analysis methods should be developed for tracking fluorescent molecules linked to microtubules.

Please contact:
Charles Kervrann, IRISA/INRIA, France
Tel: +33 2 99 84 22 21
E-mail: ckervran@irisa.fr

Improved Quantification of the Heart by Utilizing Images from Different Imaging Directions

by Jyrki Lötjönen, Juha Koikkalainen and Kirsi Lauerma

Since cardiovascular disease is the most common cause of death in the Western countries, there is a strong need to diagnose and to treat cardiac diseases. The cardiac imaging techniques, such as ultrasound and magnetic resonance imaging (MRI), have improved considerably during the last years providing nowadays detailed anatomical and functional information on the heart. We have developed a novel method for extracting quantitative measures of the heart from differently oriented image directions.

In clinical practice, the interpretation of the images is still often performed visually due to lack of automatic tools for extracting quantitative measures from images. Although visual inspection of images provides in many cases enough signs for the interpretation of images, more objective and accurate quantitative measures are needed, for example, for detecting subtle indicators related to early phases of diseases or comparison of different populations in drug-discovery studies.

Although MRI can be regarded as a golden standard for extracting three-dimensional (3D) anatomical and functional information on the heart, commercially available products provide tools only for the analysis of the left ventricle, and the tools require often a substantial amount of manual interaction. In addition, the quantitative analysis is often performed only for short-axis images (see figure). The slice thickness is usually several times higher than the resolution in the image plane, especially in functional cardiac images. For this reason, it is difficult to localize accurately different structures on the thickness direction, such as the apex and the valve levels from short-axis images. Long-axis images, which are orthogonal to the short-axis images, are often acquired in clinics providing accurate data on the problematic direction in the short-axis images.

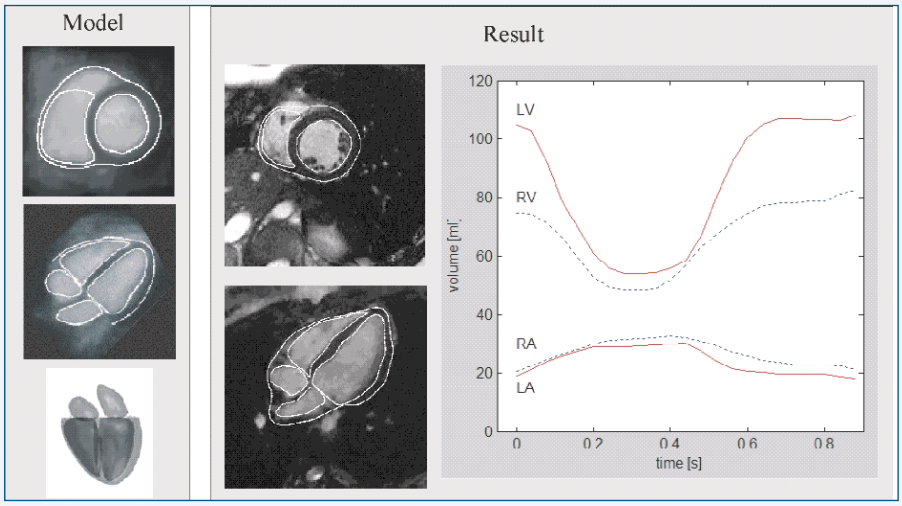
We have developed a technique where image series from two or more imaging directions are used for extracting quantitative measures from images. In this work, the technique was applied for computing volumetric measures from the ventricles and atria of the heart. The procedure consists of the following steps:

- Define the spatial relationship between the images from different imaging directions. The location of each pixel in the co-ordinate system of an imaging device can obtained from image headers. However, breathing causes movement artefacts (even more than 2 cm) which need to be corrected separately. We have developed a technique where the location of each short-axis image is optimized based on the information on long-axis images and vice versa.
- Segment the structures from images by registering non-rigidly (warping) an a priori model (template) to all image series simultaneously. Our model was

built from a database of 25 healthy volunteers; the model represents the mean shape and appearance of a healthy human heart (see figure). The similarity between the model and the data to be segmented was maximized. In addition, the typical variation of the shape was also modelled from the database and used to constrain the deformation.

- Compute the volumes of the objects of interest.

The figure represents segmentation results of one subject using the technique described above. The volumes of left ventricle (LV), right ventricle (RV), left atrium (LA) and right atrium (RA) at different phases of the cardiac cycle are also represented. Two major improvements can be emphasized: 1) The automatic definition of the volumes of atria has not been reported earlier. The volumetry of atria is interesting because it is known that the symptoms in some diseases appear first (at early phase) in atria and later in ventricles. 2) The volumes of ventricles are defined more accurately than earlier. Up to 30 % (average 10-15%) differences were detected in the volumes of ventricles, as the results using only short-axis images and using both short- and long-axis images were compared.



Model: The mean model consists of short- and long-axis image series computed from 25 healthy subjects and a 3D surface model of atria, ventricles and epicardium. The white curves on the images show the surface model superimposed on the mean grey-scale images. Result: Segmented short- and long-axis images and the volume of atria and ventricles from one subject during a cardiac cycle.

In the near future, the software tool will be installed into clinical environment and the user interface will be modified to fulfil the requirements of medical experts.

Please contact:
Jyrki Lötjönen,
VTT Information Technology, Finland.
Tel: +358 3 316 3378
E-mail: jyrki.lotjonen@vtt.fi

Juha Koikkalainen,
Helsinki University of Technology, Finland.
E-mail: juha.koikkalainen@hut.fi

Kirsi Lauerma,
Helsinki University Central Hospital, Finland.
E-mail: kirsi.lauerma@hus.fi

Analogic CNN Computing Fosters Detecting Stroke Signs

by Tamás Szabó and Péter Szolgay

Computationally intensive medical image processing applications typically require computational power usually not available with traditional computers. The CNN Applications Laboratory at the University of Veszprém, Hungary recently demonstrated analogic cellular neural network algorithms offering parallel cellular methods as well as wave methods for image enhancement and segmentation on computed tomography images.

In spite of the struggle against stroke, the frequency of cerebrovascular diseases which is ranked third in lethality worldwide has been increasing gradually all around the world. One of the most important reasons for the high lethality is the lack of efficient treatment within a critical period of three hours after the first observation of symptomes. As a result of fundamental and applied research we have developed a prototype of a cellular nonlinear network (CNN) based system with special analogic algorithms for two purposes. One is the enhancement of characteristics of diagnostic image features on computed tomography, CT images, and the other is the detection of stroke signs on CT images which can be misappreciated by a neuroradiologist. This system is integrated in a telemedical stroke expert network among hospitals and medical centers in Hungary.

A very high-speed computer architecture (four Tera equivalent floating point operation per second per square centimeter) with special algorithms reduces the computing time for a neuroradiologist to pre-process the images. A CNN may be embedded in a CT machine.

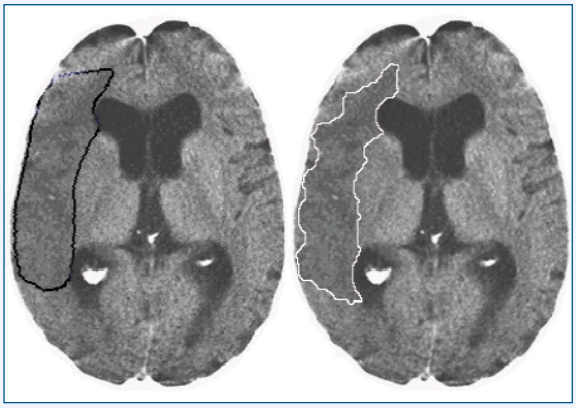
The Neurorad CNN System

CNN is a new computing paradigm of interdisciplinary interest, especially of two-dimensional biomedical imaging. The Neurorad CNN is a CNN-based system that is used to process CT scans.

CNNs are parallel-computing, analog arrays, with very high computational speed. The cells are nonlinear dynamic processing elements. The connectivity among them is local in space. The program of the network is determined by the pattern of the weights of local interactions, the templates. The time-evolution of an analog transient represents the computation on a CNN. The CNN Universal Machine architecture invented in 1992, is a novel spatio-temporal array computer. The term analogic CNN computer is also used since analog spatio-temporal dynamics is combined with logic operations inbuilt in a stored programmable framework. Pixels of a CT image are mapped to the input of the cells. CT images have relatively low pixel and grayscale resolution. In the Neurorad CNN system several analogic algorithms are implemented.

Modification of the pixel-value distribution (the histogram) is done by differential equations, so the image contrast is enhanced while simultaneously reducing noise. This method can be improved for local contrast enhancement.

After separating the gray from the white matter, the average grayscale value of each hemispheres can be compared. Moreover, an active contour with initial position located on the central symmetry line of the CT image evolves to a final position, and its deformation is proportional to the quantified measurement of the hyperdense or hypodense areas of either of the two hemispheres.



Experimental results of segmentation. From left to right: parenchymal hypodensity, a kind of stroke sign, can be seen as a region of low grayscale values marked up by a neuroradiologist; the same sign detected by an analogic CNN algorithm.

The mathematical model we use for the segmentation of CT images is underlain by a nonlinear diffusion equation which is a partial differential equation, a PDE supplemented with reactive terms. Reactive terms are responsible for forcing the evolution-based process to stop at the edges that are unfortunately ill-defined in most of the CT images. This model — as being a more general model — displays the common qualitative characteristics of remarkable segmentation models.

A mask is constructed exactly by binary mathematical morphology for the segmented structures of interest. These structures are marked up afterwards in the original image. Later, erosion and dilation are used. The boundary of the mask is superimposed on the original image. Each result is verified by a medical expert. If no verification is gained, the process resumes to a new segmentation kernel.

Detecting Stroke Signs

One possible stroke sign, a hypodense region detected by a neuroradiologist and by the proposed system is shown in the figure.

In the early hours, following stroke onset changes indicating the nature and extent of the stroke may be very subtle and difficult to see on a CT scan. To detect these early signs is important during future activities.

Please contact:
Tamás Szabó,
University of Veszprém, Hungary
Tel: +36 88 624799
E-mail: tszabo@almos.vein.hu

Péter Szolgay, SZTAKI, Hungary
Tel: +36 1 2796128
E-mail: szolgay@sztaki.hu

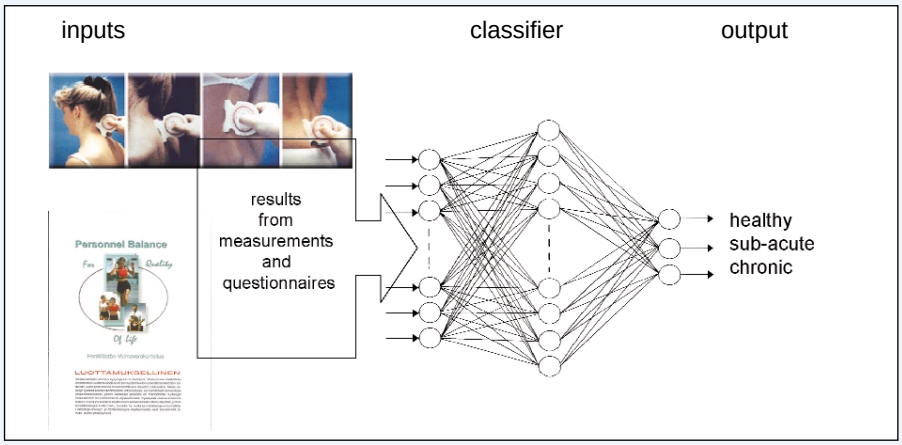
A System for Automated Back-Pain Disorder Classification

by Mark van Gils, Juha Pärkkä and Juho Merilahti

VTT Information Technology has developed a computer-based classification system for back pain problems that can quickly interpret examination and questionnaire results and suggest a suitable rehabilitation program.

Back pain is one of the most common problems in healthcare, and has tremendous financial implications on society due to costly rehabilitation programs, lost working days etc. Early and accurate assessment of the need for a suitable exercise or rehabilitation program would save substantial amounts in health care costs. However, making an early and accurate diagnosis has proven to be deceptively complex. Choosing the right plan for rehabilitation in case of back pain complaints is a non-trivial task that is typically done on the basis of (partly subjective) psycho-physiological expert examinations. There is a need to develop a computer-based system that can interpret examination/questionnaire results quickly and objectively and suggest a suitable rehabilitation program in an early phase, rather than having a patient undergo a sequence of consultations and examinations.

Because much of the knowledge used to select a rehabilitation program is implicitly present in the clinician's experience it



Schematic representation of the system. Data in the form of functional measurements and questionnaire answers are provided as input to a classifier (in this case an artificial neural network). The desired outputs are diagnoses as provided by a clinical expert. The classifier learns to perform expert-like input-output mapping by processing a large number of example cases.

is not necessarily easy to use straightforward rules or algorithms. Problems with these types of properties are typically addressed by statistical methods or learning paradigms that use a 'learning from example cases' approach. In this study we aimed to develop a classification system for back pain problems by defining an optimal set of variables and evaluating different classification paradigms.

The Back-Pain Monitor System

This study used data collected with the so-called back-pain monitor system (http://www.e-healthbase.com) provided by BPM-Group Oy (Helsinki, Finland). Data for the variables are obtained via questionnaires and functional measurements and entered by physicians or nurses in a Web-based system. The data consist of:

- functional measurements performed by a clinical expert (performance tests, assessment of muscle tightness, measurements of flexibility, mobility, posture etc)
- questionnaires filled in by the subject (questions concerning activity, subjective experience of pain, working load, alertness, stress, mood etc).

Data sets contain in the order of 60-100 recorded variables per patient. A human expert interprets the patient state (on a scale from 'excellent' to 'weakened') mainly on the basis of experience. This decision-making has the disadvantage that it is subjective, and does not give an insight into the relevance of all the other variables. The developed system combines variables into an index, which provides us with an objective, automatically calculable, and immediately available assessment of the patient's status along with suggestions for possible treatment (see Figure 1).

Data and their Classification

Two data sets were used to develop classifiers. Set A contains both functional measurements and questionnaire answers (2400 subjects; 43% healthy, 51% sub-acute, 7% chronic cases), while Set B contains questionnaire answers only (1063 subjects; 42% healthy, 58% sub-acute cases).

Set A represents data as it has been collected traditionally during a health check performed by a physician; this requires about one hour (56 scalar and 13 category variables) to complete. Set B represents data that will be increasingly collected in the future, provided by the subject through filling in a questionnaire. It requires between ten and twenty minutes to complete (8 scale variables and 90 categorical/nominal variables).

As the distinction between healthy and sub-acute is the most relevant in practical terms (ie detecting whether follow-up checks are necessary), our research focused on the separation between those two classes. One quarter of the sets was used for testing, the remaining three quarters for training. Performance was assessed using cross-validation.

For set A, principal factor analysis with varimax rotation was performed to assess how many, and which, variables would provide a good subset containing most of the variation in the data. On this basis, ten variables were retained that covered around 90% of the total variation in the data. These variables were used as input to different classifiers (eg linear discriminant, logistic regression, back-propagation and radial basis function neural networks). For set B, cross-tabulation analysis was used to process the mainly nominal and categorical data. Seventeen variables that showed significant differences in their distributions for different classes were used as input to classifiers; linear discriminant functions and logistic regression classifiers.

Results

When used in a linear discriminant function, the ten-variable set used for classification tasks in set A leads to an acceptable specificity (87%) but unacceptable sensitivity (65%). A non-linear approach using an artificial neural network trained with the back-propagation algorithm reaches a sensitivity of 81% and specificity of 74%.

For set B the cross-tabulation analysis reveals that such variables as those representing feelings of stiffness and clumsiness in body, tiredness and sleeplessness are highly associated with occurrence of back-pain disorder whereas for example the weekly amount of sport activities does not. A logistic regression classifier using the appropriate variable set reaches a sensitivity of 92% and specificity of 81%, which is acceptable for practical use.

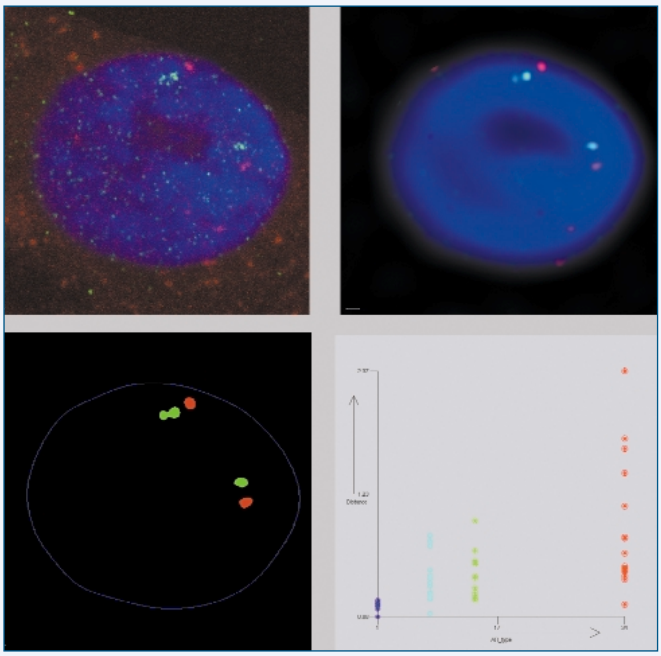
Relevant subsets of variables were found that could successfully be used for classification purposes. For data from set A, a neural network approach delivered acceptable classification results, while for set B a logistic regression classifier proved to be effective. At the moment we are implementing the classification methods in software modules that communicate with the existing Web-based database. The modules will be able to use the measurement and questionnaire data as input and present a report containing the classification results as output.

Please contact:
Mark van Gils,
VTT Information Technology, Finland.
E-mail: Mark.vanGils@vtt.fi

Computer at the Microscope: Visualization and Analysis of Three-Dimensional Microscopy Data

by Wim de Leeuw

An important topic in biomedical research is the three-dimensional organization of cellular structures. Novel developments in microscopy, such as the ability to make three-dimensional scans, play a crucial role in this research. However, manual interpretation and analysis of such scans is difficult and time consuming. Imaging and image processing can play an important role in the analysis of such scans. CWI developed the Argos system, in cooperation with the Swammerdam Institute of Life Sciences (SILS). It helps biologists in the interpretation of the microscopy data using a combination of image processing, feature detection and visualization. The system can be applied to the detection and analysis of tumour cells and to research into chromatin structure. The new combination of quantitative analysis and visualization techniques is a powerful tool for the understanding of microscopy data.



Stages in the analysis process. From top left to bottom right the raw data, filtered data, detected features, and a scatterplot showing feature attributes. The image presents a three-channel scan of a cell nucleus. Blue shows the shape of the nucleus, red and green show DNA-probes. Data courtesy: Swammerdam Institute of Life Sciences (SILS) Amsterdam.

Laser scanning confocal microscopy (LSCM) allows biologist to scan fluorescently labeled specimen in three dimensions. The Argos system aims at the analysis of three dimensional volume data such as produced by LSCM microscopy. The system integrates quantitative analysis and visualization. Visualization is used to present extracted data as well as intermediate results. This enables close monitoring of the process. The analysis is based on the extraction of biologically relevant features from the data, using image processing techniques and attributes of these features such as size and roundness.

The analysis of biological data poses specific demands on automatic processing. Noise due to the imaging process and biological variation necessitates the analysis of collections of data sets and the ability to closely track the process. The software allows the inspection of intermediate results. Interactive feedback allows the user to trace back the results back to the raw data for example to spot problems in the used feature filters or an anomaly in a particular nucleus. For typical experiments tens of gigabytes of raw data are produced. During processing, relevant information with regard to the performed operations is stored for interactive inspection later on. Argos offers image processing, feature detection and quantitative information extraction tools for the analysis of large volume data collections integrated with visualization. Using this combination of image processing and visualization, novel insights into the data can be gained.

An application of the system is the automated tumour detection in tissue samples. Currently, examination of tissue samples for tumours is done manually. The nuclei of malignant tumours can differ in various ways from the nuclei of benign tumours. The structure of the nucleus, for instance, can be more or less granular and those grains have different dimensions and densities. Argos uses image-processing techniques to quantify these characteristics based on images of the nuclei from the tissue samples. Pathologists can make much cheaper and more accurate diagnoses with these quantified characteristics than with images alone. Noise in images and natural differences between nuclei make a correct diagnosis difficult. The visualization possibilities of Argos make it possible to present various features of the investigated nuclei and to discriminate between healthy and sick cells.

Another application in which the system is used is the quantitative analysis of scientific experiments to determine structural properties of spatial cellular structures such as chromatin. Subtle differences in structure only can be deducted by the study of a collection of microscopic images. Such experiments involve a large number of scans, which have to be analyzed in a consistent manner. The analysis consists of a number of steps: filtering out the relevant features, detection of the features, calculation of attributes and statistical analysis of these attributes. Combining these steps in a single system allows close control over the entire process and tracking of the effects the various steps have on the final result.

Please contact:
Wim de Leeuw, CWI, The Netherlands
Tel: +31 20 5924320
E-mail: wim.de.leeuw@cwi.nl
http://homepages.cwi.nl/~wimc

Time Profiles Rreveal the Structure of Sleep Stages in the Neonatal EEG

by Vladimír Krajča, Svojmil Petránek, Karel Paul and Miloš Matoušek

In an article published in ERCIM News 29 the authors described the methodology for a computer-assisted long-term electroencephalogram (EEG). In this contribution, they show how this methodology could be used for structural description of sleep stages in a neonatal EEG. The procedure is based on the processing of time profiles computed by adaptive segmentation and subsequent classification into several classes. Time profiles reflect the EEG structure over time and may be used for detection of changes in the neonatal sleep stages.

EEG is one of the most important methods of studying the maturation of the child brain. The aim of the study was to evaluate the possibility of distinguishing between the EEG activity of quiet and active sleep in both full-term and pre-term infants using the automatic method. This is based on adaptive segmentation, feature extraction and subsequent classification.

The EEG recordings of 21 healthy sleeping newborns (ten full-terms and eleven pre-terms) were processed. To take into account the non-stationary behaviour of the signal, the features were extracted from piece-wise stationary segments detected by an adaptive segmentation algorithm, which used the two sliding windows to detect the change of stationariness in the signal. The most important parameters for distinguishing the EEG activities between both sleep states were the amplitude variance, the parameters describing the duration of the segments, and power in the frequency bands delta, theta, and alpha. The classification was performed by cluster analysis (k-means algorithm). The time profiles, that is, the function of class membership over time, were plotted for each EEG channel. The EEG was also evaluated by an experienced electroencephalographer by visual inspection.

The example of the signal processing of ninety minutes of neonatal EEG (time profiles and two pages of recordings) is shown in Figure 1. The segment boundaries are indicated by short vertical lines showing the change of stationariness of the signal. The segments are identified by a colour and relevant class number.

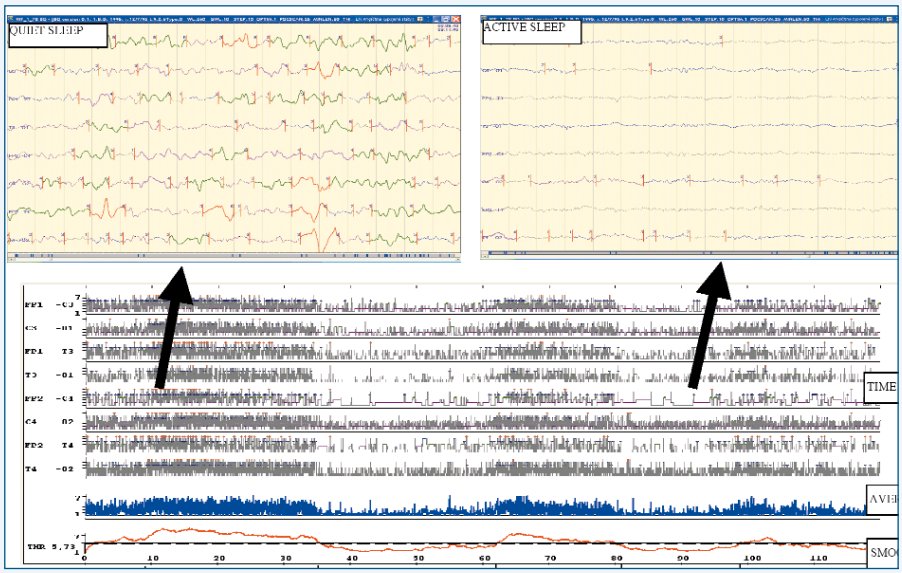


Figure 1: Ninety minutes of neonatal EEG (time profiles) with the neonatal EEG processing (eight channels). Shown above are the parts of the EEG corresponding to the time profile position below (indicated by arrows).

The time profile is the function of class membership of each EEG segment depending on the time. The awake period and quiet and active sleep stages can clearly be seen. The signal graphoelements were divided into several classes reflecting the occurrence of the basic EEG graphoelements over time. Later we also exploredwhether such time profiles can be used for detection of neonatal sleep stages. In order to obtain the detection curve, the time profiles (their class membership) were first averaged (the blue curve in Figure 1). The resulting curve was then smoothed by a simple moving average filter (the red curve), producing the final detection curve, which can be compared to the threshold.

The study revealed that this method of automatic analysis describes the electroencephalogram of neonates with sufficient accuracy. The time profiles reflect the difference in EEG activity between quiet and active sleep. The structural description of the multichannel EEG of the newborns provides a good basis for a fully automatized method of sleep-stage identification. This work was supported by grants IGA MZ _R NF7511 and NF7586.

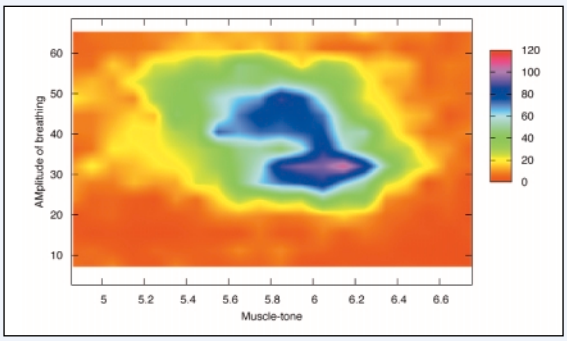
Link:
http://www.ercim.org/publication/Ercim_News/enw29/kraica.html

Please contact:
Vladimír Kraj a, Faculty Hospital
Na Bulovce/Department of Neurology,
Prague, Czech Republic
Tel: +420 2 6608 2307
E-mail: krajcav@fnb.cz

Data Mining in Children's Hypnograms

by András Lukács and László Lukács

The Informatics Laboratory of SZTAKI, the Eötvös Loránd University, Budapest, and the Madarász Street Children's Hospital, Budapest, are collaborating in infant sleep research, using state-of-the-art data mining techniques to improve the efficiency of diagnostic activity of the hospital's Sleep Laboratory and to conduct basic research in physiology.



The histogram of muscle-tone and the amplitude of breathing. The peaks on the plot correspond to different sleep phases, the upper one with lower tone and more active breathing presumably to the paradox phase, the lower one to the deep sleep.

Our project aims primarily at developing a diagnostic tool for infant and child breathing anomalies during sleep. The tool is intended for clinical use, while also addressing the needs of basic research in physiology. In order to find out whether children are threatened by breath-disorders during their sleep, expensive and time-consuming polysomnographic examinations are carried out. We intend to develop a diagnostic software which can expose the signs of danger even if no or only few apnoeas are found in the recording, thus making diagnose less expensive and more accurate.

Adult sleeping disorders are much more examined and known than those of children, although it is evident that in some aspects they are very different. The main breathing problem of adults is snoring, mainly caused by overweight or hypotonicity, whereas the childrens' main problem is apnoea which may be caused by disturbance of development or diseases of the gullet or the nervous system. The polysomnographic data analysed in our project are collected in the Sleep Laboratory of the Madarász Street Children's Hospital. The Sleep Laboratory has been working for seven years and all routinely collected diagnostic data have been archived. There are both data of healthy and diseased children in the dataset, which are also provided by the original diagnoses.

So far, data recorded at night are searched manually for apnoeas. The diagnosis is based on the number and type of apnoeas found. Because this approach takes into account only a very small part of the information collected, better visualization and automatic feature extraction tools are needed to improve diagnostics and to incorporate all possible information in decision making. Setting up standards for children's sleep using data mining techniques will make it easier and more accurate to point out the problematic phenomena in the sleep recordings.

The analysis is carried out at the Data Mining and Web Search Group, Informatics Laboratory of SZTAKI in co-operation with the Department of Physiology and Neurobiology at Eötvös Loránd University. The project started three months ago and is in its initial phase. We are building a unified database from the collected data and are creating the algorithmic tools needed to extract features from the huge amount of raw data. Input data consist of 6-8 hours long continuous polysomnographic records on 10-15 channels, which we analyse breath-to-breath and calculate descriptors for each breath-cycle. These descriptors include the time elapsed since the beginning of the sleep, length and shape of the breath-cycles measured in different channels, distribution of the lengths of the heartbeats, muscle-tone, intensity and frequency of the eye movements, and the frequency distribution of the EEG. The framework is flexible, any further descriptors suggested can be included into the model.

We use several data mining techniques to extract knowledge from this derived data set. In the individual records we use clustering algorithms to find sleep phases, and sequence mining to explore typical changes. We find patterns emerging from the sequence of breaths on two scales. First, a large scale analysis targets the total sleep-duration, and roughly identifies the sleep phases (see Figure), and then a detailed analysis finds specific changes in the individual phases. Differences among the different sleep periods provide important sources of information, as both cardiovascular problems and apnoeas occur most often close to the end of the sleep. The local analysis is done at the level of dozens of breaths close to the transitions between the sleep phases in order to reveal the order of quick events around these changes.

Our preliminary results show that data mining techniques are suitable tools to reveal an internal structure of these multi-channel recordings and to find characteristic features identified by physiologists and physicians; ie they can be used to extract useful information from this huge array of individual data. Promising findings show that pathological changes in the record could also be found and identified by software based on data mining techniques. Using these patterns, we can automatically find anomalies in the sleep records even if there are no apnoeas present. This tool can easily be implemented in clinical practice. Our aim is to offer the expert a comprehensive set of information derived from the recording and combine them in the most suitable way.

Links:
http://www.ilab.sztaki.hu/~lacko/sleep
http://www.ilab.sztaki.hu/websearch

Please contact:
András Lukács, SZTAKI,Hungary
Tel: +36-1-279 6169
E-mail: alukacs@sztaki.hu

László Lukács,
Eötvös Loránd University, Hungary
E-mail: lukacs@bolyai1.elte.hu

Articles in this Section

The Quest for Novel Computational Paradigms and Machines

by Christof Teuscher

The quest for novel and unconventional computing machines is mainly motivated by the man-machine dichotomy and by the belief that meeting tomorrow's complex real world challenges will require new paradigms and new engineering methods to organize, train, and program such machines and to interact with them.

Ever wondered how tomorrow's computers might look like? Ever wished scientists and engineers would come up with smarter, more robust, and more autonomous computers able to reliably operate in and adapt to their complex and unpredictable natural environments? Well, that's what my and many other researcher's work is focused on!

At the time of Alan Turing, computing machines were merely considered as devices capable of doing what the human mind can do when carrying out a procedure. Turing's hope that "[...] machines will eventually compete with men in all purely intellectual fields" is far from being fulfilled and the man-machine dichotomy remains more than obvious. For example, one of the keys to machine intelligence is computers that learn in an open and unrestricted way, and we are still just scratching the surface of this problem. Something must be wrong!

Although there is no reason and evidence to believe that any physical system, including the brain, can perform computations a computer cannot, there is a growing interest in computational architectures that go beyond the classical paradigms and that offer novel properties. For example, many man-made systems have become so complex that they are virtually uncontrollable and incomprehensible: self-design, self-adaptation, self-diagnosis, and self-repair will become of paramount importance.

The research community investigates novel computational paradigms along several directions and on various different levels. My own research is influenced by a number of principal insights. First, I believe that we have to go beyond the omnipresent sequential paradigms in computer science if we want to come closer to machines with properties and architectures similar to the brain. However, it is not my goal to 'copy' nature but only to draw valuable inspiration from it and to apply adapted paradigms to the design of artefacts. Second, novel approaches have to be integrative, all-embracing, and should start at the level of the computational substrate. Simulating a parallel machine on a sequential computer might be practical, but is certainly not efficient and by no means a scalable approach. Third, the capacity to adapt and learn should go in hand with the capacity to gradually create more complex and hierarchical systems, as I believe that our abilities to program complex systems are simply not keeping up with the desire to solve complex problems.

The Logic Systems Laboratory at the Swiss Federal Institute of Technology in Lausanne (EPFL) was a pioneer in biologically-inspired computing machines and hardware. For my PhD thesis, I investigated an unconventional reconfigurable architecture that is based on an amalgamation of a particle-based and randomly interconnected substrate, membrane systems, and artificial

© Alain Herzog, EPFL

The BioWall: A self-repairing cellular computing machine built at the Swiss Federal Institute of Technology in Lausanne (EPFL).

chemistries in combination with an unconventional adaptation paradigm.

The proposed reconfigurable architecture relies on a simple, irregular, inhomogeneous, locally interconnected, asynchronously operating, and imperfect particle-based substrate, not unlike an amorphous computer. However, communications are wire-based and the main part of each particle consists of a reactor for artificial chemistries. The only way to build a perfect machine out of imperfect components is to make use of redundant spare components and therefore a fine-grained particle-based implementation is beneficial. Scalability is assured by avoiding central control and by using local interactions only. An additional goal was to make the basic hardware component as simple and universal as possible. The creation of hierarchical organizations was inspired by biological membranes and membrane systems (P systems). Artificial chemistries represent, if appropriately used, an ideal means to compute in uncertain environments. We made extensive use of that unconventional form of computation as it has also been identified as potentially promising for the perpetual creation of novelty, a feature that was used for our adaptation paradigm called membrane blending. Blending (or conceptual integration) is a framework of cognitive science that tries to explain how we think and deal with mental concepts. However, instead of dealing with concepts, we draw inspiration from this framework and applied it to artificial chemistries and membrane systems.

This experimental architecture represents one step towards novel and unconventional machines and is part of the general scientific challenge of seeking for further progress and new metaphors in computer science. Although a lot remains to be done and competitive real world applications for such machines are likely to be several years away, it is crucial to start developing novel mechanisms which will ultimately help us to make machines more scalable, more robust, smarter, and more 'self-everything' in general.

Future work will be focused on further refining and improving the proposed concepts and investigating their interesting properties. The gradual creation of complex, hierarchically organized systems by means of chemical blending and the investigation of large chemical reactor network dynamics are of particular interest. Ultimately, the aim is to come up with a radically new reconfigurable architecture for adaptive systems that is fully competitive with classical, regularly interconnected and arranged Field Programmable Gate Arrays (FPGAs).

There are many further research opportunities in this exciting field. As Alan Kay put it: "The computer revolution hasn't happened yet!"

Christof Teuscher is the winner of the ERCIM 2004 Cor Baayen award (see article on page 5).

Links:
BioWall: http://lslwww.epfl.ch/biowall

POEtic tissue: http://www.poetictissue.org

BLOB Computing: http://blob.lri.fr

Amorphous Computing: http://www.swiss.ai.mit.edu/projects/amorphous

Blending: http://blending.stanford.edu

Membrane Systems: http://psystems.disco.unimib.it

Please contact:
Christof Teuscher
University of California, San Diego (UCSD)
E-mail: christof@teuscher.ch
http://www.teuscher.ch/christof

A Light-Weight and Scalable Network Profiling System

by Andreas Kind, Paul Hurley and Jeroen Massar

Long-term network monitoring in high-speed networks requires new ways for collecting, storing and analyzing flow-based network traffic information. A project at the IBM Zurich Research Laboratory looked at alternatives to the conventional flow-based network profiling approach with the objective of improved scalability for high flow rates. The result is a light-weight and scalable network profiling system for NetFlow and IPFIX that is based on a novel time series aggregation database.

The continuing trend toward distribution of computing resources increases the need to tightly control the networks providing remote access to resources such as servers, storage and databases. An important means for controlling networks is network performance monitoring and, in particular, network profiling. A typical profiling system collects and analyzes information about the traffic flows passing an observation point in the network, eg, a router or traffic meter. A flow is a sequence of packets with common properties (ie, protocol and source/destination addresses/ports).

In the past, flow-based network profiling has proven to be useful for a number of applications, including network monitoring, billing and planning. To facilitate smooth operation of service access in distributed computing architectures (eg, SANs, computational Grids) the demand for profiling will continue to rise.

Other network profiling systems have a critical scalability problem regarding the storage, analysis and access of collected profiling information. Indeed, this was

the primary motivation for the development of our approach. In high-speed networks with average flow rates of 1,000 flows/s and peak flow rates of as much as 20,000 flows/s, storage for 180 MB/h (3.6 GB/h at peak times) must be provided and maintained. Over longer time periods (ie, months, years), the data accumulates, resulting in an over-loaded system with slow reporting.

Time Series Aggregation Database

Our profiling approach addresses the scalability problem by using a novel aggregation database (ADB) for time series information. ADB provides a mechanism for efficient incremental storage of primary data values which are associated with time intervals. The database stores data values in groups of circular arrays of decreasing resolution and is, therefore, able to handle large time series data sets with fast access times and limited storage. ADB automatically assures that the array resolution of older data values is lower than the resolution of newer data values. Additionally, great care was taken with the design of ADB in order to reduce memory to disk synchronization and cache the relevant arrays in memory for fast data import and export.

Array grouping in ADB is efficient for obtaining a sorted view of related parameters. This feature is of great importance in our profiling tool for efficiently displaying sorted lists of top protocols, top hosts, top flows, etc.

ADB has a built-in array allocation optimization which further reduces the storage requirements. If an array is updated with values of progressing intervals, no preceding array space is allocated because it will never be needed. Furthermore, for insertion of new values, array space is only allocated in fixed chunks in order to avoid allocation of potentially unused array space. These optimizations reduce the required storage allocations considerably for only sparsely filled time series data streams. In network profiling, these optimizations are very useful when, for instance, a dynamically assigned IP address is only observed during a certain time period (eg, a week). In this case, space in a monthly array is only allocated around the actual observation period and not for the entire month.

Implementation

The developed system is the first profiling system we know of that implements the emerging IETF IPFIX standard. As such it supports IPv6 at data and control plane and can receive flow records over SCTP.

The reports show traffic information in graphs and tables regarding domains, protocols, QoS tags, hosts/servers, individual flows, packet and flow statistics, port/host scans and other networking aspects. The reporting period can be varied over many time scales. Custom zoom reports regarding specific traffic aspects can be generated on demand using a filter mask.

A typical problem in network profiling systems is that bursts of high flow rates can be caused by observed port or host scans. In these cases, a single packet may be considered a flow since no other proceeding packet will have the same properties with regard to the source/destination addresses and ports. The rate of the flow export can, under extreme circumstances, even exceed the data rate. Unfortunately, bursts of high flow rates can not only provoke flow table overflows at the observation points but may also render the analysis and storage to be no longer able to keep up with the incoming flow information. For these cases, we added an automatic mechanism to our system to aggregate flow records of a port or host scan into a single record.

Sample reports of the traffic profiling system.

The described profiling system, including the aggregation database (ADB), has been developed over the past two years. The system has been installed at a number of IBM locations and is currently being tested at two European ISPs. Snapshots of sample reports are shown in the figure.

Link:
http://www.zurich.ibm.com/sys/storage/resource.html

Please contact:
Andreas Kind, Paul Hurley
and Jeroen Massar,
IBM Zurich Research Laboratory, Switzerland
E-mail: {ank,pah,jma}@zurich.ibm.com

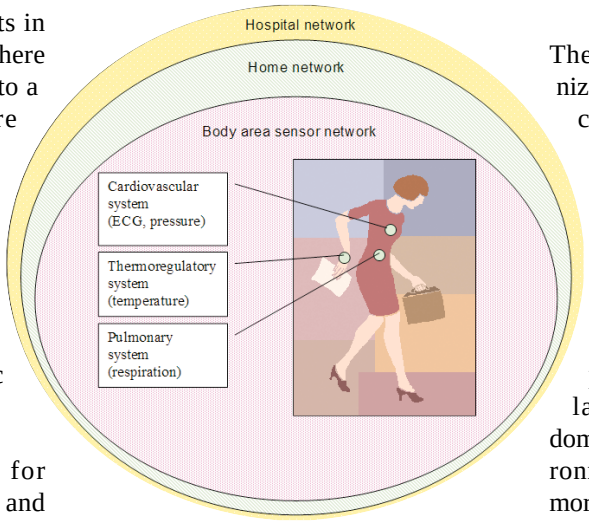
Health Care Monitoring of Mobile Patients

by Giuseppe Amato, Stefano Chessa, Fabrizio Conforti, Alberto Macerata and Carlo Marchesi

A joint activity between ISTI-CNR, IFC-CNR and the University of Florence addresses the problem of medical devices and data integration in health care. This article proposes a system for the remote monitoring of patients based on recent developments in networking and data management.

ICT has been employed in medicine and health care for many years now with great success. However, the upgrade of existing medical instruments and the design of new medical applications as a result of continuous advances in information technology must not lead to a neglect of the real needs of patients and physicians. In this respect, ICT has failed so far to fully respond to requirements related to integration and approach methodologies. Although significant progress has been made with respect to the creation of new medical instruments, the same effort has not been devoted to the integration of these instruments in operational information systems where all the devices can be integrated into a single framework of health-care resources. It should be observed that this aspect is particularly critical since different medical devices and hospital data bases using different protocols and/or data representations may be unable to interact automatically, thus failing to provide efficient diagnostic support.

In our opinion, a new scheme for cooperation between data, devices and systems is needed. As an example, let us take the scenario of a home care patient after hospitalization for cardiac infarction. Although such a patient should be guaranteed a good quality of life, he/she still needs to be in constant contact with an expert physician so that his/her cardiac activity (eg the heart rate and peripheral blood pressure), body temperature and breathing frequency can be continuously monitored. However, the health condition of a patient can only be partially evaluated through his vital signals and must be mediated and integrated by other signals and information coming both from personal characteristics (risk factors, degree of disease, age, sex, family history, psychological features, etc.) and from the environmental context (eg whether in bed or mobile, by him/herself or in company, at work or at home, the season and the temperature, etc.). The monitoring system should be able to provide a feedback to the patient as well as notifying his status to somebody else, such as a relative, the family doctor, or the hospital, depending on the degree of alert detected, and possibly adapting the level of service (ie the intensity of the monitoring activity).



Health Care Monitoring of Mobile Patients. (Source: ISTI & IFC – CNR, Pisa, Italy).

The above scenario requires the integration of different medical devices, of environmental data acquired by sensors located near the patient, of patient data available from the electronic medical records stored by the hospital, and of hospital administrative information about admission/discharge of patients, and the management of financial data and health care resources. Although the current technologies offer the necessary means to support this kind of health care, in our opinion it should be possible to access and integrate all available health care resources offering a continuous, widespread, cooperative health care system and tools for personalized patient monitoring.

To achieve this, the current view of a medical instrument as a stand-alone device needs to be rethought; it should become a node in a medical network providing results and acquiring external data in order to update its internal knowledge in order to provide customized signal/data processing and patient-oriented answers.

The medical network would be organized into layers with the patient at the centre. The inner layer which provides monitoring support is organized as a body area sensor network. This network, hosted by the patient, combines the patient's physiological data with information from the outer layers to support (basic) early diagnosis and produce (basic) alerts. The outer layer (for example the patient's domotic network) may include an environmental sensor network and one or more powerful nodes. Examples of such nodes could be an electrocardiograph offering diagnostic information or a PC receiving all the data and managing an advanced monitoring and alert detection service. This layer interacts with outermost layer (the hospital network) to exchange physiological data, alerts and patient-related data. Wireless connections should be used where possible to support mobility and adaptability at the various levels of the network.

Application scenarios of the type presented above give rise to several issues. The sensory devices constantly attached to the patient produce huge streams of physiological data which

must be collected and related to environmental conditions. These sensors should be light and portable to reduce their impact on the patient's well-being (and thus must be constrained in terms of energy capacity). Consequently, the amount of information transmitted outside the network should be minimized in order to prolong its lifetime. Data is forwarded to and analysed by the hospital network only in certain (critical) cases; otherwise data is processed locally with the support of information from the patient's case history acquired from the hospital network.

For these reasons, we have developed a network and middleware layers for sensor networks, supporting the execution of queries originated by external controllers (such as a hospital network or a single physician locally connected via his notebook). The controller issues patient-specific queries which are optimized and distributed over the sensor network. These queries specify monitoring, data collection and/or processing tasks with different levels of intensity. They also define the flux of information between sensors.

Each sensor is thus assigned a sub-query defining the set of sensors providing the data streams necessary to execute the sub-query and the rules to combine such streams. At this level, communication between sensors is implicit since sensors only use operations to open, read or write local or remote data streams. At a lower level, opening a remote data stream implies that the data produced on a remote sensor should be routed and buffered on the local sensor.

The sensors manage streams via the middleware layer (called the stream system), which provides support for the management of local and remote data streams and for stream buffering and naming, and exploits the services of the network layer for routing.

The network layer provides support to the communication models used by the stream system (unicast and multicast). It defines a virtual coordinate system which assigns a coordinate to each sensor in the network and allows for efficient geographical routing.

The virtual coordinate system is hop-based and unrelated to the physical location of the sensors. Thus it does not require sophisticated equipment to determine coordinates and has little overhead. The network layer also embeds an energy efficiency management module which turns off the wireless interface during periods of inactivity. These periods are computed taking into account the requirements of the network and the stream system layers.

Future work includes the study of dependable and secure communication protocols to connect the body area sensor network with domotic and hospital networks. These protocols should ensure confidentiality and protection against the transmission of malicious queries. They should ensure integration of the sensor network monitoring the patient with the medical devices and with the administrative and patient data available in the domotic and hospital networks.

Please contact:
Stefano Chessa, ISTI-CNR and University of Pisa, Italy
Tel: +39 050 3152887
E-mail: chessa@isti.cnr.it

Pervasive Patient Monitoring — Take Two at Bedtime...

by Dirk Husemann and Michael Nidd

A team at IBM Zurich Research Laboratory Zurich has created the IBM mobile health toolkit for gathering measurement data from a range of devices, and present it to management software via a well-defined and easily implemented interface.

About 55% of all long-term patients in the US and in Europe, it is estimated, do not take their medication (either not taking the prescribed medication at all or more than 14 hours late) Around 12% of all hospital admissions in the UK are due to this non-compliance, the damage to the US taxpayer is an estimated USD 100 billion a year. Most of the patients that do not comply are simply forgetful (about 10% deliberately do not want to take the medication). For a lot of diseases it is important that doctor and patient work together: as the visit to the doctor provides just a snapshot in time of the patients health status, the GP has to rely on observations and self-measurements by the patient herself to provide a good diagnosis and select the right therapy. To capture this information patients are often asked to keep a patient diary. Unfortunately (another piece of bad news here), over 40% of all patient diaries are false. The British Medical Journal reports of cases where patients faked their patient diaries half an hour before their doctor's appointment while sitting in the doctor's parking lot, using ten different pens to make it look credible!

Gathering current patient medical data promptly and accurately is vital to proper health care. The usefulness of electronic data capture (EDC) has been demonstrated in applications such as the home monitoring of at-risk heart patients via devices that transmit blood pressure from the home to a central database. Removing transcription effort (and associated inaccuracies) alone is worth the institution of EDC; but the side benefit

of timeliness offers the hope of identifying and responding to trends as they occur, perhaps preventing a dangerous event, instead of simply allowing its diagnosis after the danger has manifest.

The market is now at the beginning of the technology curve for EDC. A typical EDC offering is a full end-to-end solution, where a customer purchases measurement devices, data gathering and network technology, database, and visualisation/analysis software as a single package. For quality and cost to improve, work in the area must be allowed to specialise. To this end, our team has created an efficient and flexible toolkit, the IBM mobile health toolkit, for gathering measurement data from a range of devices, and present it to management software via a well defined, and easily implemented interface.

Figure 1: The IBM mobile health toolkit eco-system.

The Treatment

The IBM mobile health toolkit provides a Java-based middleware -- using J2ME MIDP 2.0 (Java Mobile Information Device Profile) and JSR 082 (Java APIs for Bluetooth) -- running on a personal (mobile) hub device to which sensors can connect wirelessly. We can perform local processing on the data, and forward the result to one or more fixed network connections. Data-handling modules can easily be added to the MIDlet suite (application suite compliant with Java Mobile Information Device Profile) on the hub, as can drivers for new sensor devices. While the data formats from these devices are usually proprietary, a popular physical medium for their transmission is Bluetooth. JSR 082, which standardised Bluetooth access from Java, means drivers for new devices are also small and easily written. A key feature of our mobile health toolkit is the use of an open, hierarchical event model segmented into \emph{domains.} Each piece of sensor information is normalized into a standardized event by the corresponding sensor driver before the information is forwarded to the toolkit; for example, a blood pressure cuff measures systolic and diastolic blood pressure values along with the current pulse --- the driver converts these sensor readings into two events: a standard blood pressure event and a standard pulse event. The use of a normalizing event model allows us to write downstream applications and adapters independent of the specific sensors used --- we can easily switch to other sensors models without having to modify existing applications.

Using a wireless link from the hub to the devices, as shown in the Figure, allows the hub to be placed in an unobtrusive location, saves the user from fiddling with cables, and saves the sensor manufacturer the trouble of finding an acceptable case location for the data connector. By requiring only Bluetooth, MIDP support, and a network connection from the hub, the range of suitable hardware choices for the hub extends from full PCs, through OSGi home gateway units, all the way to cellular phones.

The Status

By allowing different business partners to utilize one another's applications and sensors in new ways, the IBM mobile health toolkit has already enabled applications and solutions that were not previously possible.

Figure 2: Bluetooth attached blood pressure cuff, Bang & Olufssen Medicom IDAS II medication device, mobile phone as hub device.

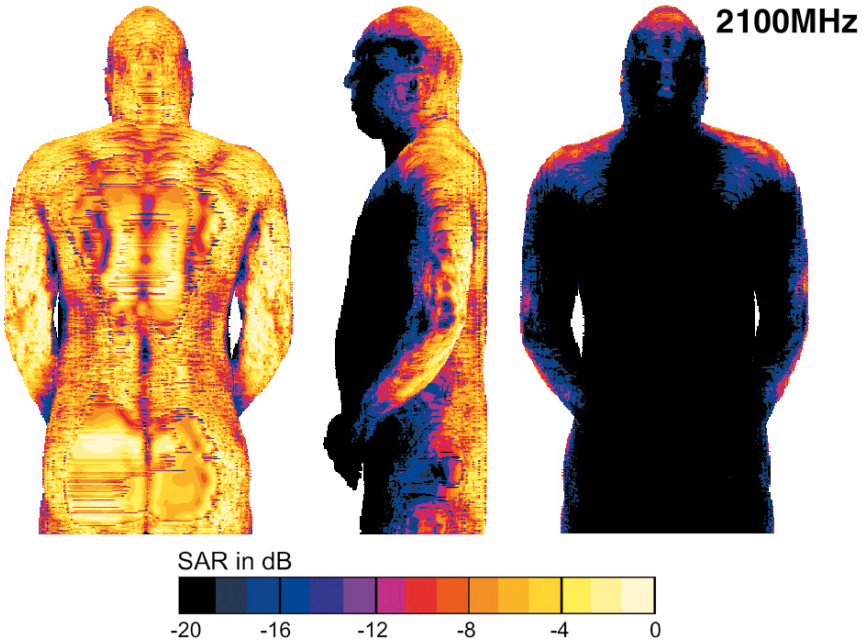
We are currently in the third iteration of the IBM mobile health toolkit and have reached a state where the software is running rather reliably and is mature enough to be deployed in first applications. The current version is implemented completely in Java and already fully supported on such restricted platforms as J2ME MIDP 2.0 (requiring the JSR 082 Bluetooth Java API for sensor attachment).

Please contact:
Dirk Husemann and Michael Nidd,
IBM Zurich Research Laboratory,
Switzerland/SARIT
E-mail: {hud,mni}@zurich.ibm.com}

New Study on Effects of UMTS Signals on Human Well-Being and Cognition

by Gregor Dürrenberger

A study carried out by TNO — Netherlands Organisation for Applied Scientific Research — on effects of UMTS-fields on human well-being, terminated in 2003, will be replicated in Switzerland. Since their communication, the TNO-findings were controversially discussed in the scientific community. An independent replication co-ordinated by the Swiss Research Foundation on Mobile Communication should clarify the robustness of the main results of the TNO-study.



Modelling of SAR-distribution at 2100MHz (UMTS). SAR = Specific Sbsorbtion Rate, power absorbed by the body tissue, in Watts per kg. (Source: IT'IS).

In September 2003, TNO published results of a study investigating the effects of radiofrequency electromagnetic fields (from 'mobile phone basestations') on subjective well-being and cognition (Zwamborn, A.P.M., Vossen, S.H.J.A., van Leersum, B.J.A.M., Ouwens, M.A., Makel, W.N. (2003) Effects of Global Communication system radio-frequency fields on Well Being and Cognitive Functions of human subjects with and without subjective complaints. Netherlands Organisation for Applied Scientific Research (TNO). FEL-03-C148) In this study, two groups of volunteers – hypersensitive and non-hypersensitive persons – were exposed in a double-blind approach to GSM, UMTS and - for control purposes - to no fields at all. 'Double-blind' means that neither the exposed person nor the researcher present during the experiment knows about the real exposure, ie, whether GSM, UMTS or no field is 'on'. Each of the 36 volunteers in the two groups attended a test session and three subsequent randomized exposure sessions. The sessions, each lasting for 30 minutes, took place within half a day. During the sessions the subjects performed a series of cognitive tests, and filled-in a well-being questionnaire after the sessions. 24 subjects were included in the (final) statistical analysis. The study found no effects of GSM on well-being. Exposure to UMTS signals, however, revealed a reduction in the subjective feeling of well-being. With regard to cognition, no consistent pattern was found.

Despite the fact that the scientific quality of the TNO-study was high, experts commented on weak points with regard to the study methodology and data analysis. In the Swiss replication some of these weak points will be eliminated and some additional aspects will be included.

Replication and Extension

The replication study focuses on the effects of UMTS signals on well-being and cognitive functions. No GSM signals will be used. The study hypothesis is that, analogous to the original study, exposure to UMTS-like radiation will decrease the feeling of well being, possibly in a dose-dependent manner, but will not affect cognitive performance of the subjects.

With improved dosimetry (organ and functional brain tissue specific calculations will be performed) two exposure levels (1V/m and 10V/m) will be used. TNO applied 1V/m only. Due to this extension, the study can detect dose-effect relationships, presuming that the effects found in the TNO-study can be reproduced and that dose/dependency exists. An exposure setup identical with the original setup will be used. Three treatments (1V/m, 10V/m and sham) will be applied on each subject in a randomized, double-blind design.

As in the original study, the Swiss study will expose both hypersensitive and non-hypersensitive persons. The overall sample size will amount to 84 persons (24 hypersensitive volunteers, 60 non-hypersensitive volunteers; a subset of the non-hypersensitive persons will be matched to the group of hypersensitive persons). This is almost double the size of the sample that entered data analysis in the original study. Hence, statistical significance of the results will be higher in the replication study. Sample sizes for both groups were calculated with a power analysis.

During exposure sessions subjects will need to perform cognitive tests on a computer. After the sessions the TNO well-being questionnaire will be handed out. Additionally, subjects will have to

fill-in an improved well-being questionnaire prior to and after exposure. Washout-periods between sessions are at least one day; if logistically manageable, one week.

The replication study will not be able to conclude about potential adverse health effects due to changes in subjective well-being. Also, no answers about specific mechanisms about the causal links between (UMTS) radiation and both well being and cognitive functions will be given. However, the detailed dosimetry carried out may generate additional insights into these topics.

Project Team, Duration and Costs

The study was designed by Dr. Peter Achermann (Institute of Pharmacology and Toxicology, University of Zurich), Prof. Niels Kuster (IT'IS and ETH Zurich) and Dr. Martin Röösli (Institute of Social and Preventive Medicine, University of Berne) in cooperation with TNO. The project is commissioned by the Swiss Research Foundation on Mobile Communication

Research work started in September 2004 and will last one year. Results will be available after publication in a peer-reviewed scientific journal, probably at the end of 2005.

The overall study costs amount to 485'000. 60% is funded by three Swiss public authorities (Swiss Federal Office of Public Health SFOPH, Federal Office of Communication OFCOM, Swiss Agency for the Environment, Forests and Landscape SAEFL) and four Dutch Ministries (Ministry of Economic Affairs EZ, Ministry of Health, Welfare, and Sport VWS, Ministry of housing, Spatial Planning and the environment VROM, Ministry of Social Affairs and Employment SZW); 40% is funded by the three Swiss mobile telephone providers: Swisscom Mobile, Orange and sunrise.

Link:
http://www.mobile-research.ethz.ch

Please contact:
Gregor Dürrenberger,
Swiss Research Foundation on Mobile Communication c/o ETH Zurich/SARIT
Tel: +41 1 632 28 15
E-mail: gregor@mobile-research.ethz.ch

Development of a Real-Time 2D and 3D Echocardiographic Diagnostic System

by Dániel Hillier, Zsolt Czeilinger, András Vobornik, Zsolt Szálka, Gergely Soós, László Kék, Viktor Binzberger, David Lopez Vilarino and Csaba Rekeczky

Within the frame of a multi-disciplinary research project (TeleSense), the Analogic and Neural Computing Laboratory of SZTAKI has developed a prototype echocardiographic diagnostic system with telepresence capabilities. The system assists clinical diagnosis through novel three-dimensional reconstruction, display and two-dimensional analysis functionalities. These features significantly improve both the planning process of cardiac surgery and the efficiency of daily cardiac diagnosis.

Our research objective in the field of echocardiography was to create a prototype system in a PC environment that can efficiently support the daily routine of cardiologists. In the case of two-dimensional (2D) echocardiography, the aim was to develop a system having the ability to estimate in real time the volume of heart chambers, as well as analysing and evaluating the wall motion of the human heart. This information can be a valuable support for cardiologists making several medical diagnoses per day. In the case of three-dimensional (3D) echocardiography, the primary aim was to model and reconstruct the structure of interatrial muscles. These can give anatomical information that is useful in the preparation of successful surgery; typically the implantation of an occluder.

To achieve this aim, an experimental system based on a cellular neural network (CNN) was designed and implemented for efficient 2D/3D echocardiography image analysis and reconstruction. The 3D view of the human heart is reconstructed from 2D projections taken at different angles by an electronically controlled transducer. The analysis of these 2D slices (filtering, segmentation, contour tracking and content/context-based recognition) is designed as an analogic CNN-UM algorithm (UM stands for Universal Machine); the 3D reconstruction from the reduced data set (contour sub-sampling-interpolation, 3D rotation-translation, and polygonal reconstruction) is designed as a DSP algorithm. The system is based on the ACE-BOX computational infrastructure hosting both types of the aforementioned microprocessors, which ideally support these computationally demanding experiments.

The main hardware-software components of the ACE-BOX-based echocardiography diagnostic system have already been developed, and the algorithm optimization is an on-going effort. These experiments incorporate hundreds of video-flows of different patients,

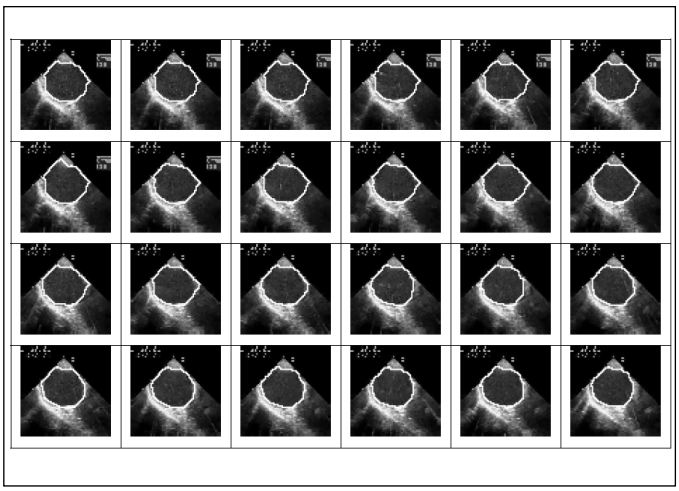


Figure 1: Single-chamber contour tracking performed by the PLS algorithm on the ACE4k platform.

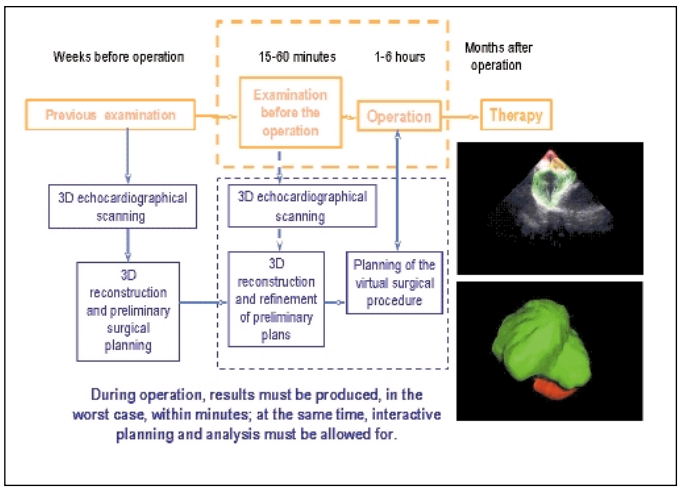


Figure 2: Aiding the surgical planning and implementation in 3D echocardiography.

which are stored in a common database. Using a specially designed software tool, the cardiology specialists in our team support the experiments by semi-automatically tracking the borders of different chambers on all 2D projections. This way a set of 'groundtruth' data is created, which validates the performance of the segmentation and tracking algorithms and is used as a reference data set for the 3D reconstruction experiments.

We have developed, implemented and tested three different versions of topographic cellular active contour techniques for heart chamber localization and tracking. These methods, called Constrained Wave Computing (CWC), Pixel Level Snakes (PLS) and Moving Patch Method (MPM), are solving the task of endocardial boundary tracking necessary for further processing. Figure 1 shows the output of the PLS algorithm in a single-chamber contour-tracking task that has been fully implemented on the ACE4k CNN-UM chip, meeting a video real-time performance with all system level overheads.

3D reconstruction examples are shown in Figure 2. In particular, two- and three-chamber analysis has been put in focus with the development of additional on-line tools that make it possible to provide efficient support to on-line surgical interventions. Figure 2 also summarizes the time criteria that must be met by engineering support in order to effectively help cardiac surgery. Screenshots illustrate the workflow when working with a 3D model of a reconstructed atrium.

One particularly valuable outcome of our work is a database that has been built with the help of the cooperating cardiologists and contains more than one thousand echocardiographic video-flows. More than half of the video-flows also contain reference contours that were traced manually by cardiologist experts. Our future work will include algorithm improvement and optimization based on this database and an experimental clinical introduction and testing of the entire system.

Our echocardiographic system also has native support for telepresence, that is, a technological environment that enables medical experts to see via wired or wireless network connection the results of a diagnosis undertaken at a remote site. At the same time the expert can consult the database and make use of the technological support developed for easy evaluation of clinical data.

Diagnosis results can also be uploaded into the database. We have ensured that it is possible to reach the ACE-BOX analogical computing platform over TCP/IP; in other words, a cardiologist can access the hardware-software devices via a simple Internet connection, and use both our algorithms and the database.

The hospitals participating in the program (Saint Francis Hospital - SzFK, György Gottsegen National Institute for Cardiology – GGy-OKI) contributed by conducting specific medical experiments, continuously building the database, manually tracing reference contours and testing the user interfaces of developed algorithms. The design and integration efforts regarding the hardware architecture and the algorithmic framework were undertaken at SZTAKI. Péter Pázmány Catholic University (PPKE) and IT Consult-Pro Ltd. contributed by developing software tools, a simulation environment and core algorithms. A hardware-accelerating platform and connected software layers were implemented by AnaLogic Ltd GE Hungary Corp. provided hardware and software tools, support and user training.

We are confident that our system, capable of the above-mentioned functions, running with hardware acceleration and providing interactive support for 3D reconstruction and surgery planning, would be welcomed in medical institutions. In the current phase of the program, important developmental steps will be made towards a final market-ready product by conducting clinical introduction and testing.

Link:
http://lab.analogic.sztaki.hu/telesense

Please contact:
Csaba Rekeczky, SZTAKI, Hungary
Tel: +36 1 2796131
E-mail: rcsaba@sztaki.hu

The Advantages of Reused Software Components

by Nancy Bazilchuk and Parastoo Mohagheghi

Software reuse in a product family approach is commonly thought to lead to fewer product problems, greater productivity and easier maintenance. However, little empirical data has been found to support this assumption – until now. Recent analysis of more than 13 000 problem reports collected by the mobile phone company Ericsson in Grimstad, Norway, has shown that software reuse does result in significantly fewer problems and better stability.

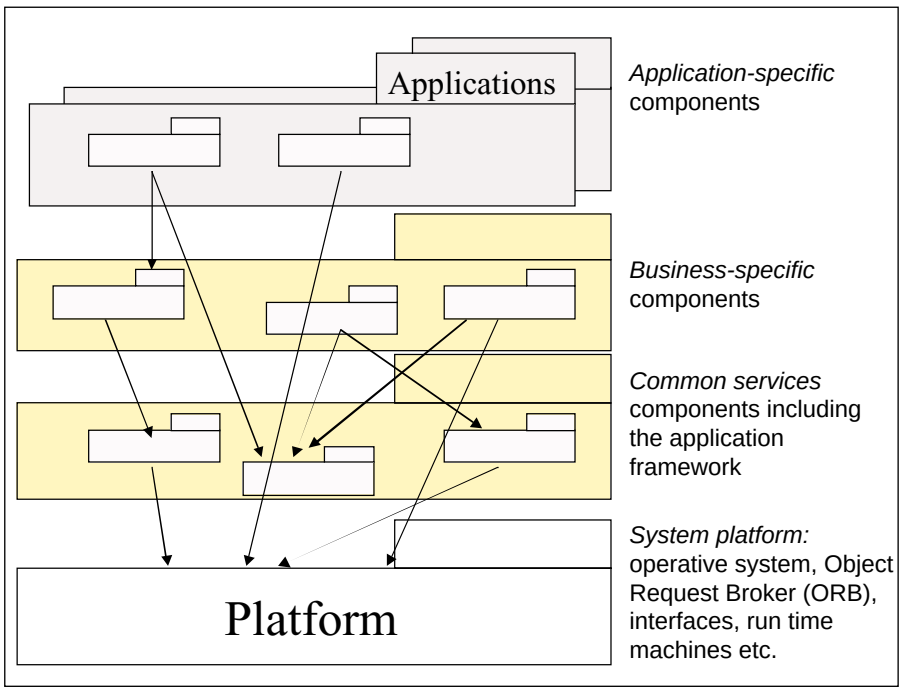
Like most companies, Ericsson-Norway collects a large amount of data relating to its software. Much of this data is awaiting analysis, and represents a mine of information on product characteristics and quality. An analysis by Parastoo Mohagheghi, a PhD graduate from the Norwegian University of Science and Technology, sifted through Ericsson's data to try to answer the question: what effect does software reuse have on product defects and stability? Her answer, discovered with co-researchers from the university, Ericsson-Norway and the Simula Research Laboratory in Lysaker, Norway, is that software reuse significantly improves quality.

Companies like Ericsson are increasingly moving toward component-based software engineering (CBSE), where related products and systems can be assembled from pre-built components. These reusable components can take a variety of forms, from existing software libraries, to free-standing commercial, off-the-shelf products (COTS) or open-source software (OSS), to entire software architectures and their components. CBSE promises many advantages, such as a shortened product development time, reductions in total costs, and — since new software components can be purchased instead of developed in-house — fast access to new technology.

Ericsson engineering teams in Sweden and Norway built two large-scale, distributed telecom systems containing several reused components. One of these systems was evaluated as part of Mohagheghi's doctoral research. The system's platform was developed by a sister Ericsson organisation, and was considered to be a COTS component, although in reality it too could have been evaluated as a reused component. The middleware, component framework, and the business-specific software were all reused components. Data from several releases of the system were collected and analysed, with the results of the analysis of one release presented in an award-winning paper at the 26th International

Conference on Software Engineering in Edinburgh, Scotland (ICSE '04). Mohagheghi's paper examined one release that comprised 470,000 lines of non-commented Code (KLOC), of which 64 percent was in Erlang, 26 percent was in C, and the remainder was in other programming languages like Java or Perl.

If a defect in a component is detected during integration or system testing, or later during maintenance, Ericsson programmers write a Trouble Report that describes the defect in detail. The report typically contains information such as the severity of the defect, the assumed origin of the problem, the amount of time needed to correct the defect, and a defect code that assigns the problem to a trouble category, such as coding, wrong design rule, or documentation problem. Mohagheghi and her co-researchers examined 13,000 of these Trouble



An overview of Ericsson's GPRS software architecture that has been designed to support software reuse. The reused components are found in the business-specific and common services packages, and are shared between two GPRS solutions for different networks.

Reports to compare the quality of the reused and non-reused components.

The analysis showed that reused components had a lower defect density than non-reused ones, with defect density calculated by dividing the number of defects by the number of lines of code. At the same time, however, the reused components had more defects with the highest level of severity than the total distribution, but less of these defects after delivery. Mohagheghi and her co-workers concluded this was because defects in reused components were assigned a higher priority to be fixed. Other researchers have assumed that reused components would change more often than non-reused components because the reused components had to meet the requirements of several different systems. However, Mohagheghi's analysis also showed that reused components were far less likely to be modified between successive releases, as measured by the percentage of code that was modified. This meant the reused components were far more stable than their non-reused counterparts. Mohagheghi has also studied how prone components were to change, as measured by the number of changes in requirements and deliveries for reused and non-reused components in this release, but did not find any significant difference between the two types.

This study was financed by the INCO project (INcremental and COmponent-based Software Development), a Norwegian R&D project in 2001-2004.

Link:
The entire study is available in Mohagheghi's dissertation, available at http://www.idi.ntnu.no/grupper/su/publ/phd/mohagheghi-thesis-10jul04-final.pdf.

Please contact:
Parastoo Mohagheghi
NTNU, Norway
E-mail: Parastoo.Mohagheghi@idi.ntnu.no

POAC: Optical Computer for Large Data Sets

by Ahmed Ayoub, László Orzó and Szabolcs Tőkés

A new Programmable Optical Analogic Array Computer (POAC) has been developed at the Analogic and Neural Computing Laboratory of SZTAKI. It optically identifies, clusters, discretizes, and classifies features of large data sets at high speed and with high parallelism. It has proven to be an excellent device for tracking moving objects.

The ultimate speed is the speed of light. There is no such thing as a short circuit with light, so beams can cross without any problem. This and other features of light interaction motivated the development of the Programmable Optical Analogic Array Computer (POAC), which implements through optical means the Cellular Nonlinear/Neural Computer Universal Machine (CNN-UM). The work having commenced in 2000, POAC has gone through several developmental phases in the last five years, passing from early prototypes to the first real optical computer. This new supercomputing machine is based on multidisciplinary research that has merged cellular neural network (CNN) theories, optical principles, and concepts of biological and holographic memories. Its basic optical processor is a new type of holographic correlator that uses bacteriorhodopsin as a dynamic holographic material.

The most recent version of POAC can process in a flash an input array size of up to 250 000 elements (pixel format). Compared to digital computers, its processing power is estimated to be about 300 giga-operations per second, and it takes only a single millisecond to execute a complete flow-operation. VLSI (Very Large Scale Integration) technology is still not able to provide devices that can cope with such extreme speed and parallelism. It is also cascadable, meaning more complex operations can be carried out. Eighteen optical micro-programs (B-templates) and one optical macro-program (algorithm) have so far been developed. With a height of eighteen centimeters, the physical dimensions of POAC are comparable to the size of this page.

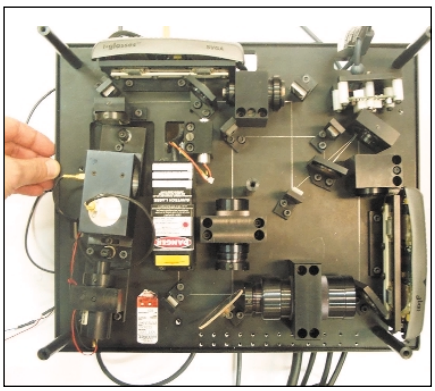


Figure 1: The most recent laptop hardware version of the programmable optical analogic array computer (POAC) provides high computing speed and full parallelism, having a large input array and template array size.

Among the other applications for which POAC can prove useful, feature identification and classification of large data sets are straightforward tasks. In the former case, data sets of size 500x500 can be processed with featuring templates, resulting in a detection rate of up to 1000 different features per second (single-input multi-template). In the latter case, the input can be changed fifteen times per second (multi-input multi-template). Other examples include target recognition and tracking. Not only can the exact target be recognized and tracked within a video frame rate, but a parallel operation can also be run to classify the similarity to the target of other objects in the scene. The degree of similarity between the classified set of objects and the original target is a subproduct of the whole process and is easy to obtain.

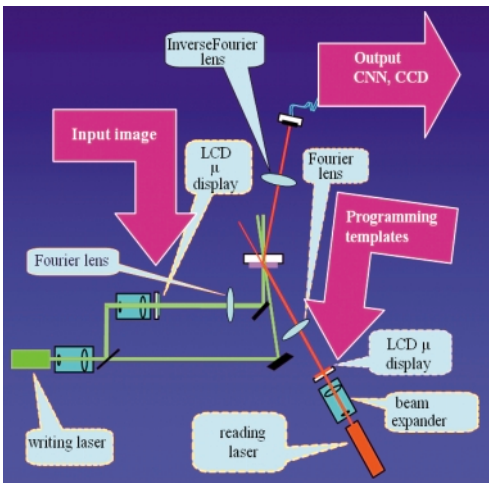


Figure 2: The internal optical architecture of POAC. The correlator uses a green laser for writing the hologram of the input array and a red laser for reading-processing.

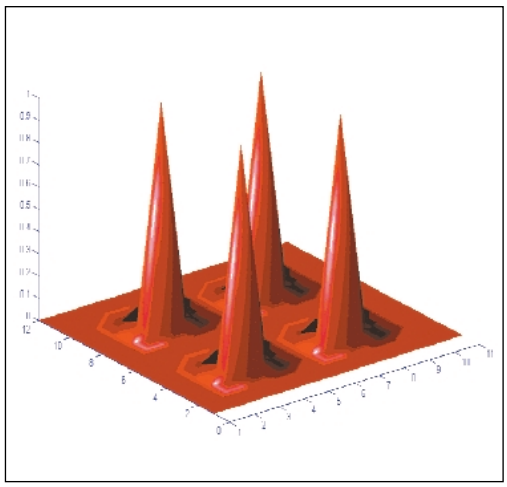


Figure 3: The optical software promises to exploit the programmability feature of POAC to attain its high computing capabilities. Here the result of the corner detection template operation as applied to a square is shown.

Future activities will centre on the enhancement of the most recent version of POAC. The hardware is being improved through the upgrading of key elements and the addition of new optical computing modules, such as coherent optical post-processing of the correlogram, A-template feedback and post-processing by a CNN-UM visual microprocessor. The software will include a greater variety of optical operations, including feature detectors, geometric operations, image analysis, operations based on mathematics and mathematical morphology, and region-based processing. In addition, POAC will be able to manage 800x600 input elements with an equivalent speed of 7.1 tera-operations per second. A proposal to embed the state-of-the-art visual CNN chip (Xenon) within the POAC architecture will enhance its ability to handle large data sets and images at extremely high speed.

Link:
http://lab.analogic.sztaki.hu/research

Please contact:
Ahmed Ayoub, Szabolcs T kés, SZTAKI
Tel: +36 1 279 6135
E-mail: ayoub@sztaki.hu, tokes@sztaki.hu

CASCOM — Context-Aware Health-Care Service Co-ordination in Mobile Computing Environments

by César Cáceres, Alberto Fernández, and Sascha Ossowski

The field of telemedicine is one of the fastest growing application areas for intelligent mobile services. Scientists at University Rey Juan Carlos, Madrid, are developing an open infrastructure for business application services across mobile and fixed networks. This project, known as CASCOM (Context-Aware Business Application Service Coordination in Mobile Computing Environments), is funded by the European Commission's FP6 IST programme. CASCOM will deliver a demonstrator for medical emergency assistance that accounts for the on-the-fly coordination of pervasive health care services.

Since its creation, the Internet has revolutionized many aspects of our life. Like many other fields, telemedicine has benefited from the ubiquitous access to (medical) knowledge granted by the Internet. Nowadays, a patient can be monitored without needing to be moved to a health-care centre. In addition, there is potential for improved diagnosis and treatment, through the transfer of patient records from one hospital to another, for instance. Two requirements are crucial to the field: the interoperability of medical information systems and electronic health records on the one hand, and the security and privacy of personal information on the other.

Until recently however, it was necessary to have at least a personal computer physically linked to a fixed wired network, a fact that severely restricted the mobility of users of telemedicine applications. Still, the latest advances in wireless and mobile technologies (eg Bluetooth, WiFi, GPRS and UMTS) have overcome this limitation. Furthermore, the availability and popularity of small mobile devices (eg PDAs, mobile phones, GPS and medical devices) have created a plenty of opportunities for increased user mobility in the field.

It is therefore not surprising that telemedicine is one of the fastest-

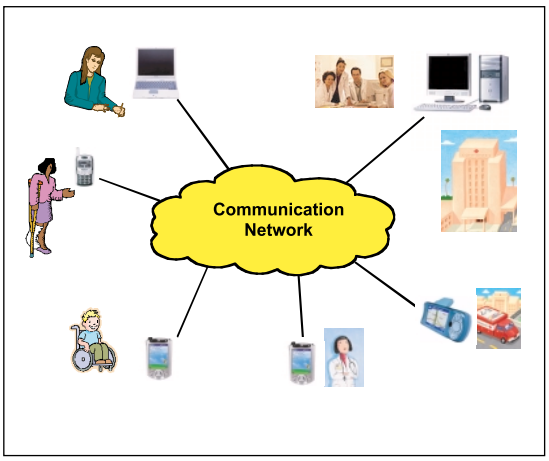


Figure 1: Telemedicine scenario.

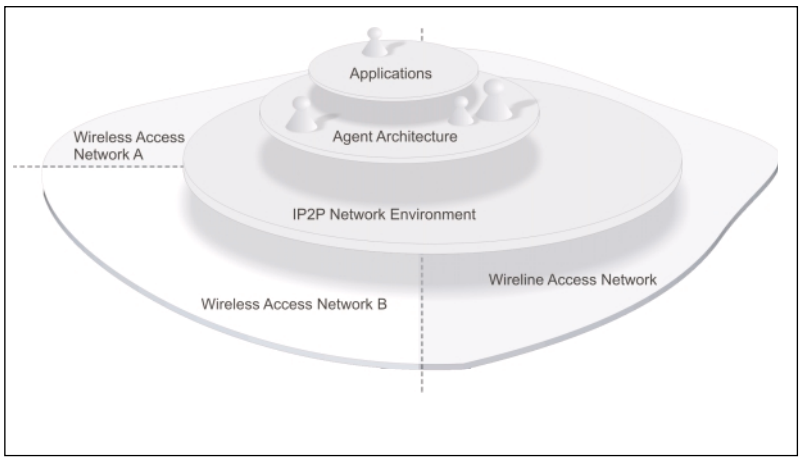


Figure 2: CASCOM layered architecture.

growing application fields for intelligent mobile services. Nevertheless, a significant number of new challenges have come up. In particular, certain technological and domain constraints (computational and bandwidth limitation, security and privacy concerns etc) must be taken into account in order to successfully build open, large-scale, pervasive applications for the health-care domain.

These challenges are now being addressed by CASCOM. To this end, we are developing value-added supportive infrastructures for business application services to be employed by mobile workers and users across mobile and fixed networks. The CASCOM approach is the innovative combination of intelligent agent technology, semantic Web services, peer-to-peer and mobile computing for intelligent peer-to-peer (IP2P) mobile service environments. IP2P represents an extension to conventional P2P architectures, with components for mobile and ad hoc computing, wireless communications, and a range of pervasive devices.

The CASCOM architecture is organized into three layers:

Network Layer: this layer provides a generic, secure and open IP2P network infrastructure, taking into account the varying quality of service of wireless communication paths, the limitations of resource-poor mobile devices, and the contextual variability of nomadic environments.

Service Coordination Layer: this layer uses agent technology for flexible semantic Web service discovery, dynamic context-aware semantic Web service composition, fault-tolerant interleaving of planning and service execution, and secure service execution and monitoring providing service data consistency.

Application Layer: this layer gives CASCOM applications functionality in a variety of business service scenarios. In particular, a demonstration will be developed for the healthcare domain that includes:

- the integration of business process models of medical experts across Europe
- the provision and coherent integration of distributed patient records
- the development of methods for maintaining application semantics along the business process models that in turn are based on service agents and semantic Web service descriptions.

In the application layer, particularly for healthcare scenarios, we plan to use W3C standard OWL and OWL-S when describing semantic Web application services. These standards will be also used in the service coordination layer and IP2P network layer wherever applicable. In implementing the agent-based service coordination layer and IP2P network layer service agents, FIPA standards - in particular those related to agent communication and management - will be followed. FIPA application standards (eg Nomadic Application Support) will be used in the IP2P network layer, while the IP2P transport layer will be based on IETF standards such as HTTP and TCP. Further, throughout the CASCOM architecture, UML and AUML will be used in the design of architecture components.

Although the CASCOM architecture is not expressly designed for a specific application, its primary field of validation is the telemedicine domain and the on-the-fly coordination of pervasive healthcare services. In particular, a prototype for medical emergency assistance will be designed, implemented, and evaluated. It is worthwhile pointing out, however, thatin addition to the health-care domain, we will ensure that CASCOM's innovative project technology is applicable to a variety of other business application scenarios, including personal assistance, virtual enterprises and entertainment (eg distributed games and other interactive settings).

The CASCOM consortium consists of DFKI (Germany), TeliaSonera AB (Sweden), EPFL (Switzerland), ADETTI (Portugal), URJC (Spain), EMA (Finland), UMIT (Austria) and FrameTech (Italy). This article reports on joint work being undertaken by the consortium. The authors would like to thank all partners for their contributions.

Link:
CASCOM: http://www.ist-cascom.org/

Please contact:
César Cáceres, Alberto Fernández
and Sascha Ossowski
University Rey Juan Carlos, Madrid, Spain
E-mail: {ccaceres, al.fernandez,
s.ossowski}@escet.urjc.es

Virtual Reconstruction of an Egyptian Beaker

by Marco Callieri and Flora Silvano

The Visual Computing Laboratory of ISTI-CNR, Pisa, frequently collaborates with museums and conservators-restorers in the development of new instruments that can be used in cultural heritage preservation, restoration and display. We present a small but interesting case study in which a broken artifact is virtually reconstructed and can be displayed with a user-friendly visualization tool.

The most immediate use for 3D models generated by scanning real objects is the preservation of their form for measuring purposes and documentation. But what can be done if the object is fragmented or incomplete? Thanks to the Department of Ancient World of the University of Pisa, we had the chance to work on an Egyptian artifact with such characteristics. Working with the digital model of the fragments, we were able to reconstruct the entire form of the object and enable users to interact with it via an ad-hoc display procedure. This kind of reconstruction can be obtained without altering the actual state of the artifact, unlike hands-on interventions using resins or other filling materials, and can be used to experiment and evaluate different alternatives.

The most complete piece of engraved glass found in the excavations at the site of Medinet Madi (Egypt) is a beaker of colorless glass (H. 15 cm.; D. rim 13.1), dated 2nd or 3rd century AD. The beaker has a rounded rim, straight sides and on the exterior presents a wheel-cut decoration. On the concave base there is a star-like motif consisting of a twice-outlined pentagon with slightly curved sides. There are five circular depressions in each semicircular space between the points and a single larger one inside the star. On the exterior surface there is a row of narrow oval facets running horizontally just below the rim, and below these a herring-bone pattern. The main register, with two horizontal grooves above and below, is composed of four panels divided by four vertical pillars of rice-shaped facets. Two of the panels show a diamond pattern with a floral design inside. The edge of the lozenge consists of a double row of slit-like facets, and four curving stalks with four rice-shaped facets. The other two panels are very fragmented and are characterized by a cross-hatched pattern with a double incised border. Unfortunately, no fragment of the upper part of the panel has been found and no similar pattern has been seen on related objects, so it is not possible to completely reconstruct the decoration.

The 3D scanning of a glass object is not an easy task as the laser light of the scanner passes through the transparent surface of the object. For this reason, the object was first coated with a white dulling powder and then acquired using a Minolta Vivid 900 3D scanner. The powder was then removed without any harm to the artifact. Different shots were taken of each fragment to ensure a complete (and precise) coverage of the surface. The shots were then aligned carefully using our software and merged. The result was four high-resolution 3D models describing the geometry of the single fragments. The figures show how the fine details of the artifact have been preserved in the 3D model; the engrav-

Figure 1: Reconstructing the original shape. Top left: the three upper fragments; bottom left: two views of the base fragment; right: the pieces are aligned with the geometric template of the whole object.

Figure 2: In close-up examination, light can be moved to give a 'glazing light' effect.

ings are mainly1.5mm wide and 0.4 mm deep.

Since most of the object is missing, it is not possible to reconstruct the beaker entirely without introducing artificial data. For this reason, we decided to use a 'real world' museum technique. A geometric template representing the beaker's external surface was built in order to position the fragments correctly. The program combined data extracted from the 3D model and sketches produced by museum restorers.

The fragments were aligned on the template with the help of our aligner tool. They were first positioned manually using the sketches as a guide; an automatic refinement process then aligned them precisely on the template.

Our visualization tool is a simple application, based on the OpenGL graphics library. For the purposes of this particular artifact, it has been implemented with a user-friendly interface: the object can be rotated and particular areas zoomed with simple movements of the mouse. The user can change the direction of the light to have a 'glazing light' visualization if he/she wants to enhance the geometric 'readability' of the object. It is possible to view just the original fragments or both the original parts and the whole beaker template. To guarantee that the tool can operate on older (less powerful) systems, different levels of geometric detail has been generated; the user can choose the level of complexity that best fits his/her requirements (fast object browsing or slow close-up examination).

Although not particularly innovative, we believe that this is a good example of a small, easy-to-use application targeted to the specific needs of inexperienced users, and implemented with state-of-the-art technology. It also represents a further step towards closer collaborations between the research community and the world of cultural heritage restoration.

Please contact:
Marco Callieri, ISTI-CNR, Italy
E-mail: marco.callieri@isti.cnr.it
or Flora Silvano, Pisa University, Italy
E-mail: silvano@sta.unipi.it

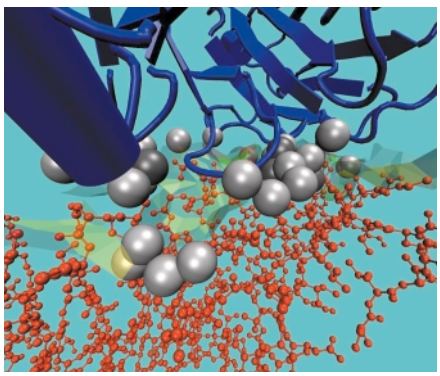
The GeometryFactory and CGAL — The Computational Geometry Algorithm Library

by Andreas Fabri

The mission of the GeometryFactory is to make the large body of geometric computing accessible to industrial developers in the form of easy to integrate C++ software components. The GeometryFactory is a start-up of the of the academic CGAL project (www.cgal.org) which was founded by seven European research groups in 1996, and which develops CGAL, the Computational Geometry Algorithm Library.

An Ubiquituous Need for Geometric Computing

The CGAL users are in domains as different as computer aided design (ECL), structural geology (Midland Valley, Agip, IFP), geographical information systems (Leica Geosystems, BAeSystems), finite elements for weather forcast (WeatherNews), antenna placement (France Telecom, British Telecom), location based services (TruePosition), VLSI (Toshiba) biochemistry (see figure), etc. The above mentioned companies switched from pure inhouse software development, based on scientific publications and on non supported freely accessible prototypes, to 'commercial off the shelf software', which allows them to focus on the application layer where their core competence is.



Modeling protein-protein interfaces using the Voronoi diagram of balls (that is the weighted alpha shape of CGAL). The yellow/green/red facets model interactions between the two (red and blue) proteins, and the interface of the complex consists of the collection of such facets. The interface is discontinued due to crystallographic water molecules trapped between the proteins. (Courtesy of Frederic Cazals, INRIA).

Technology

The library contains data structures and algorithms like Delaunay triangulations in 2D, 3D and dD, Voronoi diagrams for points, segments, and circles, Boolean operations on 2D polygons and 3D polyhedra, arrangements of segments and circular arcs, nearest neighbor search, convex hull, and much more.

The geometric software components of CGAL are a unique combination of extreme reliability, speed, ease of integration, interoperability, and in many cases unique functionality. The value for the users are a reduction of the risk of delays, a reduction of development costs, and of time to market.

The CGAL library represents cutting-edge technology. This holds for the geometric algorithms, which are directly transfered from academia, as well as for

the software design, where we did not reinvent the wheel, but follow best practice.

Exact Computation Paradigm

Correct geometric algorithms need reliable numerical computations as foundation. We follow the exact computation paradigm that is well established in the computational geometry research community. It requires the exact evaluation of geometric predicates, ie, decisions derived from geometric computations have to be correct. Results obtained by researchers of the CGAL project make that this exact evaluation of predicates can be done efficiently. In a nutshell, this is achieved by combining interval arithmetic on floating point numbers with arbitrary precision numbers as they are provided by the GMP (GNU Multiple Precision) arithmetic library, http://www.swox.com/gmp) or the CORE library (http://www.cs.nyu.edu/exact/core/).

Our approach is in contrast to current industry practice that uses a small value below which two numbers are considered equal. Such an approach leads to unreliable solutions that are hard to debug and almost impossible to prove to be correct.

Generic Programming Paradigm

Generic programming gives a tremendous flexibility at development time, and efficient code at run time of a program. Its power became apparent with the STL, the Standard Template Library, which is shipped with every C++ compiler.

As the 'std::set' containers of the STL is parameterised with the type of the objects it contains, and an order predicate for comparing the elements in the set, CGAL data structures are parameterised by, for example, the point type and orientation predicates on this point type.

Back in 1995 we took the risk to adopt the generic programming paradigm for the design of CGAL. It was a risk, because neither the compilers nor the potential customers were mature. Now, in 2004, all C++ compilers comply with the C++ ISO standard. Furthermore, today most C++ developers are familiar with generic programming. We are hence beyond the phase where only early adopters of technology use CGAL.

License Issues

CGAL is available under the {\sc Qpl} Open Source license which does not allow commercial usage. For the latter users need a commercial license from the GeometryFactory. The Industry Research License gives access to all CGAL components for a low annual rental fee. The Industial Development License for individual software components allows integration in commercial products.

Links:
http://www.cgal.org
http://www.geometryfactory.com

Please contact:
Andreas Fabri, Geometryfactory, France
E-mail: Andreas.Fabri@geometryfactory.com

SUGGEST: An Online Recommender System for Large Web Sites

by Ranieri Baraglia and Fabrizio Silvestri

The continual increase in Web usage has led to the need for automatic Web-mining tools able to accurately extract, filter and select information of interest from the huge quantitiesof data available. In particular, Web Usage Mining (WUM) tools typically extract knowledge by analyzing the data logged during user navigation, and can be used to develop personalization or recommender systems whose main goal is to improve web site usability.

SUGGEST is a recommender system that has been designed to dynamically generate personalized content of potential interest for users of large web sites. The system has been developed at ISTI-CNR, Pisa, in the context of a project of the Italian Ministry of Education and Research called 'Services for Enhanced Contents Delivery'. It is implemented as a module of the Apache Web server, and its usage does not require any modification to the site being examined. Personalization is achieved by means of a set of suggestions (page links) dynamically generated on the basis of the active user session, which are used to personalize the HTML page requested on-the-fly.

Typically, the WUM personalization process is structured according to two components, performed off-line and online with respect to the web server activity. By analyzing the historical data (ie server access log files), the off-line component builds a knowledge base which is used in the online phase to generate the personalized content. This content can be expressed in several forms, such as links to pages or advertisements considered of interest for the current user.

The main limitation of this two-tier approach is the loosely coupled integration of the WUM system with the web server activity. This comports the periodic running of the off-line component to update the knowledge base; the frequency with which this updating operation should be performed is case-sensitive. The merging of the two

components also raises other problems in terms of system efficiency.The integration must have little impact on user response times, and the knowledge mined by a single component must be comparable or better than that obtained using two separate components.

The solution introduced by SUGGEST eliminates drawbacks and satisfies the criteria mentioned above. By exploiting a single component working completely online with respect to the web server functionalities, the system can update the knowledge base incrementally and automatically and can generate a list of suggestions.

Example of suggestions generated by SUGGEST.

SUGGEST is structured in the following three steps:

User Session Identification

User sessions are identified by means of cookies stored on the client side. Cookies contain the keys to identify the client sessions. On each page request, SUGGEST identifies the URL requested and the URL from which the request originates. The knowledge base is updated according to the characteristics of the current session, and suggestions are then generated. To extract information about navigational patterns, SUGGEST models the web page of a site as an undirected graph whose nodes are associated with the identifiers of the accessed pages, and edges are associated with weights representing the degree of correlation existing between pages. Presuming that interest in a page depends on its content and not on the order in which a page is visited during a session, the edge weight is computed as $W=N_{ij}/\max\{N_i,N_j\}$, where N_{ij} is the number of sessions containing both pages i and j, and N_i and N_j are the number of sessions containing only page i or j, respectively. Dividing by the maximum number between single occurrences of the two pages has the effect of discriminating internal pages from the so-called index pages (eg home pages) that are of little interest as potential suggestions.

In order to manage web sites with an a priori unknown number of pages, eg sites that use dynamic pages intensively, SUGGEST indexes pages only when they are required. This solution can lead to a large increase in the adjacency matrix M used to store the weights related to each pair of pages. To avoid M assuming an unmanageable size, a 'Least-Recently-Used' algorithm is applied. According to this algorithm, information about a page less recently accessed is replaced with that for a currently accessed page. The smaller the matrix size, the poorer the system performance due to frequent page replacements. Parameters such as web site size, available resources and performance level required can help to define the size of M.

Page Clustering

To find groups of strongly correlated pages, the graph is partitioned according to its connected component. Starting from the current page identifier u, a Depth First Search is applied on the graph induced by M and the component reachable from u is connected. To reduce the contributions of poorly represented links, the computation of the connected components is driven by the predefined threshold values *Minfreq* and *MinClusterSize*. Edges with a weight below *Minfreq* identify poorly correlated elements which are not considered by the connected components algorithm. Components of size smaller than MinClusterSize are considered not sufficiently significant and are discarded. Pages in the same cluster are ranked according to their co-occurrence frequency.

Suggestion Building

In order to build suggestions, the current user session must be classified. This is done in a straightforward manner by finding the cluster that includes the largest number of pages in that session. Suggestions are composed by the most relevant pages in the cluster, according to the order determined by the clustering phase. An example of how suggestions are presented to the user is given in the figure.

More details about SUGGEST can be found in the paper: R. Baraglia, F. Silvestri 'An Online Recommender System for Large Web Sites', IEEE/WIC/ACM International Conference on Web Intelligence, Beijing, China, September 20-24, 2004 (best paper award).

Link:
http://www.miles.cnuce.cnr.it/

Please contact:
Ranieri Baraglia, ISTI-CNR, Italy
Tel: +39 050 315 2994
E-mail: ranieri.baraglia@isti.cnr.it

Fabrizio Silvestri, ISTI-CNR, Italy
Tel: +39 050 315 3011
E-mail: fabrizio.silvestri@isti.cnr.it

Ambulant Player: A Universal Multimedia Player

by Annette Kik and Dick Bulterman

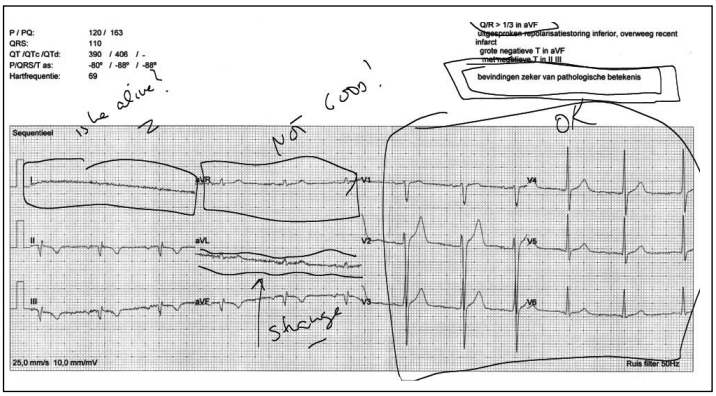
Putting your CD collection on computer, sending pictures with your cell phone: Multimedia technology appears more and more in daily life. Shortly it also will be possible to take fragments of TV programs and share them with your chat partners. Researchers at CWI are developing a software infrastructure to create multimedia presentations for parallel and distributed systems. This has applications in health care and at home. As a first step they developed the Ambulant Player, an open source multimedia player, as a test platform for researchers all over the world.

Multimedia researchers use many different kinds of multimedia players. This makes comparing results difficult. With the Ambulant Player CWI has created an environment everybody can use. At the moment it is one of the fastest and most complete SMIL-standard-based media players in the world. In addition the Ambulant Player is quite portable: It can be used on PCs and Macs as well as on PDAs. The player was released in August 2004 as open-source software and was funded by NLnet. It can be downloaded from the CWI website.

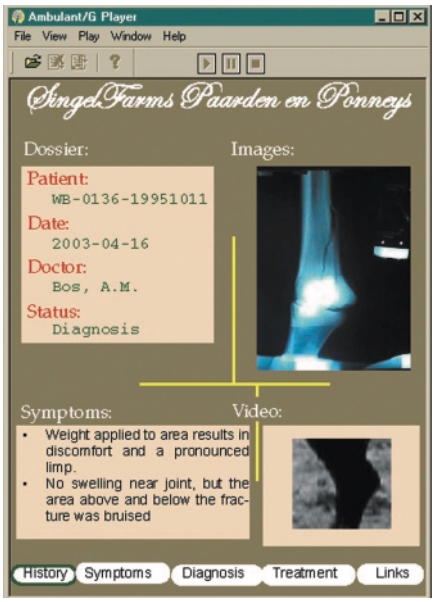
The Ambulant Player is the first step towards a software infrastructure to create multimedia presentations for parallel and distributed systems. SMIL, the W3C's Synchronized Multimedia Integration Language, enables users to integrate sound, images, film and text into one multimedia presentation that can be played on a wide variety of devices and shared via the Web. Currently, image and sound are played from one server. In the future this should be done from more servers simultaneously. This requires exact tuning of the different data streams.

One of the possible applications of this research is to compile, maintain and enrich medical files. A pilot project was conducted with electronic medical files of horses containing text, sound, images, and video. Multimedia is especially useful in this case. Since horses do not talk, vets must rely on observations of the animal's behaviour and on medical data. When a vet wants to consult a colleague, it is easier to transport a detailed file than the horse itself. Furthermore, vets are more inclined to cooperate since most horses are not insured. Physicians are under pressure from government and insurance companies, leaving less time for experiments.

Building up an electronic medical file takes several steps. First it requires an administrative entry. Just as with humans a basic number of administrative data are needed that can be consulted both locally and from far away. Next media data must be collected. During diagnosis text, image, sound and video images can be stored. Subsequently all irrelevant material is removed. Finally, a report is generated using a multimedia editor based on templates.



Ambulant Annotator allows physicians to annotate electro cardiograms in an electronic medical file.



The possibilities of the Ambulant Player were demonstrated in a pilot project with electronic medical files for horses. All information relevant for the diagnosis and treatment including text, video, images and sounds can be stored in these files.

For physicians it is also important to annotate data. They can for example make notes for remarkable parts in an electro cardiogram, or they can circle a fracture on an X-ray. This way other physicians can quickly see what is wrong. They can also record comments with video recordings and consequently want to accelerate or slow down the images while the comments stay the same. This creates enormous challenges. For this purpose CWI develops the Ambulant Annotator.

The Ambulant project builds on CWI's long history in the area of multimedia. CWI researchers developed the CMIF language for synchronizing over networks. Around 1993 this language was the basis for SMIL, now the world-wide standard thanks to the World Wide Web Consortium. The institute is one of the main suppliers for W3C's SMIL working group.

Links:
http://www.ambulantplayer.org
http://www.cwi.nl/sen5

Please contact:
Dick Bulterman, CWI, The Netherlands
Tel: +31 20 592 4300
E-mail: Dick.Bulterman@cwi.nl

FMICS 2004 — Ninth International Workshop on Formal Methods for Industrial Critical Systems

by Juan Bicarregui, Andrew Butterfield and Alvao Arenas

The Ninth International Workshop on Formal Methods for Industrial Critical Systems (FMICS 04) was held in Linz, Austria, during September 20-21, as a co-located event of the 19th IEEE Conference on Automated Software Engineering. The workshop series promotes the use of formal methods for industrial applications by supporting research in this area and by serving as a forum for the exchange of ideas between researchers and practitioners, in both industry and academia.

This workshop, organised by the ERCIM Working Group on Formal Methods for Industrial Critical Systems and CCLRC Rutherford Appleton Laboratory, was attended by 35 participants from academia and industry from 16 countries. The two keynote speakers gave interesting and stimulating presentations. Jeremy Dick from Telelogic spoke on linking formal methods to formal requirements describing how existing tools for supporting requirements traceability could be adapted to work with formal specification and refinement documents. Cedric Fournet of Microsoft Research spoke on the verification of the security of XML-based web-services and described how the 'applied' pi-calculus was used to analyse the safety of security policies, work which has contributed to recent revisions of Microsoft code. FMICS would like to thank both invited speakers for their relevant and highly informative contributions to the success of the workshop.

Seventeen submitted papers were presented with authors from 17 countries spanning formal methodologies as diverse as state-charts, model-checking, mixed intuitionistic logic and Boolean equation systems; and applications ranging from operating systems, network services, communications protocols and middleware behaviour, to flight guidance.

The best paper award supported by the European Association of Software Science and Technology (EASST) was awarded to Martin Fränzle and Christian Herde for their paper on proof engines for bounded model checking of hybrid systems.

Other papers presented included Object Oriented concepts identification from formal B specifications; an Abstract Interpretation Toolkit for _CRL; Early Verification and Validation of Critical Systems; and Model Checking Flight Guidance Systems: from Synchrony to Asynchrony, among others.

The proceedings of the workshop are published as a technical reports of the Johannes Kepler University and will appear in Electronic Notes in Theoretical Computer Science. Selected papers will be invited for publication in an special issue of Formal Methods in System Design.

The organisers wish to thank FME and the i-Trust Working Group for sponsorship for the invited speakers. Participants enjoyed a good Austrian dinner courtesy of ERCIM.

Links:
Ninth International Workshop on Formal Methods for Industrial Critical Systems:
http://www.fmics04.clrc.ac.uk/

ERCIM Working Group on Formal Methods for Industrial Critical Systems:
http://www.inrialpes.fr/vasy/fmics/

Please contact:
Juan Bicarregui, CCLRC, UK
Tel: + 44 1235 445710
E-mail: J.C.Bicarregui@rl.ac.uk

CALL FOR PAPERS

CIMED 2005 — Computational Intelligence in Medicine and Healthcare

29 June - 1 July 2005, Costa da Caparica, Lisbon, Portugal

The second International Conference on Computational Intelligence in Medicine and Healthcare (CIMED) evolved from the successful series of International Conferences on Neural Networks and Expert Systems in Medicine and Healthcare (NNESMED).

CIMED'2005 is organised by the IST Project, BIOPATTERN, and is co-sponsored by UNINOVA, IEE, IEEE and IPEM. The focus of CIMED'2005 is on the development and application of novel and robust intelligent computational methods and systems to support key areas of biomedical, clinical and healthcare practice, making it a strongly interdisciplinary conference, bringing together healthcare specialists, clinicians, biomedical engineers, bioinformaticians, computer scientists, communications and computer network engineers and medical/biostatisticians.

Topics

- bioinformatics
- computational intelligence for biodata analysis - trends and progress in biomedical and healthcare applications
- evaluation and benchmarking
- medical data acquisition and standards
- e-delivery technologies for e-healthcare (including Grid, multi-agent, mobile and wireless technologies; IP and broadband networks; quality of service and security issues in ehealthcare)
- clinical applications of computational intelligence for biomedical informatics.

Important Dates

- Submission of full papers: 31 January 2005
- Notification of provisional acceptance: 28 February 2005
- Submission of camera ready papers: 18 April 2005
- Early registration deadline: 30 April 2005.

More information:
http://www.uninova.pt/cimed2005/
http://www.biopattern.org (then select 'Events').

CALL FOR PAPERS

Hypertext 2005 - The 16th International Conference Hypertext and Hypermedia

Salzburg, Austria, 6-9 September 2005

The Sixteenth International ACM Conference on Hypertext and Hypermedia: Concepts and Tools for Supporting Knowledge Workers will focus on concepts, methodologies and tools for supporting knowledge workers. This year's program focuses on:

- hypermedia in digital libraries
- hypermedia in the humanities
- hypermedia and information retrieval.

Deadlines (preliminary)

- full papers and hypertexts: 17 March 2005
- other categories: 9 June 2005.

More information:
http://www.ht05.org/

CALL FOR PARTICIPATION

INTEROP-ESA'2005, the First International Conference on Interoperability of Enterprise Software and Applications

Geneva, Switzerland, 23 - 25 February 2005

The INTEROP-ESA conference aims at bringing together researchers, users, and practitioners dealing with different issues of the Interoperability of Enterprise Applications and Software. The conference will focus on interoperability-related research areas ranging from enterprise modeling to architecture and platforms to ontologies defining interoperability semantics in the enterprise. The interoperability in enterprise applications can be defined as the ability of a system or a product to work with other systems or products without special effort from the customer or user.

More information:
http://interop-esa05.unige.ch

CALL FOR PARTICIPATION

WWV 2005, The First International Workshop on Automated Specification and Verification of Web Sites

Valencia, Spain, 14-15 March 2005

The increased complexity of Web sites and the explosive growth of Web-based applications has turned their design and construction into a challenging problem. WWV'05 provides a forum for researchers from the communities of Rule-based programming, Automated Software Engineering, and Web-oriented research to facilitate the cross-fertilization and the advancement of hybrid methods that combine the three areas. The program includes two invited talks and presentations in the categories 'work in progress', 'overviews of more extensive work', 'position papers' and 'reports of practical experiences'.

More information:
http://www.dsic.upv.es/workshops/wwv05

ANNOUNCEMENT OF A COMPETITIVE CALL FOR AN ADDITIONAL PROJECT PARTNER

The following project currently active in the *Sixth Framework programme of the European Community for research, technological development and demonstration activities contributing to the creation of the European research area and to innovation (2002-2006)* requires the participation of a new project partner to carry out certain tasks within the project.

Project contract number: 511598; Project acronym: COGAIN; Project full name: Communication by Gaze Interaction
Strategic objective: IST / eInclusion; Instrument type: Network of Excellence
Expected duration of participation in project: from September 2005 to August 2009
Language in which proposal should be submitted: English
Opening date and time of the call: February 23rd, 2005, 17h00 Brussels time
Closing date and time of the call: March 30th, 2005, 17h00 Brussels time
Address for further information: http://www.cogain.org/call (web), call@cogain.org (mail)

COGAIN is a Network of Excellence, aiming at a durable integration of activities in its field (using eye-gaze to help in the communication of special user groups, especially those suffering from motor neuron diseases). The project started in September 2004, and the goal is that it will during its five-year funding period create structures that remain active and influential through other means of funding after August 2009. COGAIN aims to achieve this goal by forming a network of (a) developers of eye communication applications, (b) developers of eye-tracking technology, and (c) representatives of user organisations and communication centres. Partners in the network should be willing to share their visions and to commit to using jointly developed platforms and application interfaces. The degree of integration is a key element when the progress of the project is evaluated.

Summary of task(s) requested: With this call, we look for (1-4) partners who develop eye communication applications or are able to provide new tracking solutions, interface approaches, or language models used for text entry; (1-4) partners who promote exploitation of the technology and applications developed in the project; and (1) partner to expand our contacts with user communities; in total, at most 5 new partners will be selected.

Total Commission funding available for the tasks by the new partners is maximum EUR 141.600 (about EUR 7.000 per partner per year on the average). The project budget is revised yearly; actual funding depends on how actively the COGAIN partners are involved in the activities of the project. The bulk of the funding goes to integration activities.

ERCIM NEWS

ERCIM News is the magazine of ERCIM. Published quarterly, the newsletter reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of 11,000 copies.

Advertising
For current advertising rates and conditions, see http://www.ercim.org/publication/ERCIM_News/ads.html

Copyright Notice
All authors, as identified in each article, retain copyright of their work.

ERCIM News online edition is available at http://www.ercim.org/publication/ERCIM_News/

ERCIM News is published by ERCIM EEIG, BP 93, F-06902 Sophia-Antipolis Cedex
Tel: +33 4 9238 5010, E-mail: office@ercim.org
ISSN 0926-4981

Director: Michel Cosnard, ERCIM Manager, INRIA

Central Editor:
Peter Kunz, ERCIM office
peter.kunz@ercim.org

Local Editors:
AARIT: n.a.
CCLRC: Martin Prime
M.J.Prime@rl.ac.uk
CRCIM: Michal Haindl
haindl@utia.cas.cz
CWI: Annette Kik
Annette.Kik@cwi.nl
CNR: Carol Peters
carol.peters@isti.cnr.it
FORTH: Eleftheria Katsouli
ekat@admin.forth.gr
Fraunhofer ICT Group:
Michael Krapp
michael.krapp@scai.fraunhofer.de
FNR: Patrik Hitzelberger
hitzelbe@crpgl.lu
FWO/FNRS: Benoît Michel
michel@tele.ucl.ac.be
INRIA: Bernard Hidoine
bernard.hidoine@inria.fr
Irish Universities Consortium:
Ray Walshe
ray@computing.dcu.ie
NTNU: Truls Gjestland
truls.gjestland@ime.ntnu.no
SARIT: Harry Rudin
hrudin@smile.ch
SICS: Kersti Hedman
kersti@sics.se
SpaRCIM: Salvador Lucas
slucas@dsic.upv.es
SRCIM: Gabriela Andrejkova
andrejk@kosice.upjs.sk
SZTAKI: Erzsébet Csuhaj-Varjú
csuhaj@sztaki.hu
VTT: Pia-Maria Linden-Linna
pia-maria.linden-linna@vtt.fi
W3C: Marie-Claire Forgue
mcf@w3.org

Subscription
Subscribe to ERCIM News free of charge by:
- sending e-mail to your local editor
- contacting the ERCIM office (see address above)
- filling out the form at the ERCIM website at http://www.ercim.org/

News about legal information relating to Information Technology from European directives, and pan-European legal requirements and regulations.

Update on Content Rating

As the Internet has become an important source of information for society, issues that are controversial are also widely distributed on the internet.

With the protection of minors in mind, it raises the question if access to this kind of information should be restricted. After all minors surf the web frequently and can be confronted with websites containing harmful or sometimes even illegal content, such as racism, pornography and bullying. For many years now this discussion has been focussing on technological measures to tackle this problem.

One of the starting points for technological solutions however is that the content should be labelled in an uniform way. The system used most often, ICRA (Internet Content Rating Association), is based on rating of the content by its publisher or website owner, which enables filtering on used keywords in the html header code. In Europe we find this system being rejected by ISPs (Internet Service Providers), publishers and other organisations in several countries.

Besides the fact that experts warn us that children are more familiar with the Internet than most parents and teachers (software installed on their computers will trigger them to find ways to get around the filters). There are other important reasons why filtering is rejected:

- content rating will only work if all major parties will cooperate with such system
- it's widely accepted that ISPs only perform a 'mere conduit' activity. Generally speaking an ISP is not liable for 3rd parties content (for example the webpage of a subscriber)
- as long as there's no serious consideration to do so (like a legal injunction) an ISP will be held responsible by the website owners if their website hosted by the ISP concerned is hindered. For ISPs this threat has consequences for their willingness to participate in rating policies on harmful, but not illegal, content
- as long as the information involved is not illegal within the country itself, most European countries and ISPs will consider blocking access to harmful information in other countries as highly undesirable.

Two reports on the subject may be helpful in finding solutions: Recently MEP Mrs. Edith Mastenbroek (Committee on Civil Liberties, Justice and Home Affairs) told the European Parliament that more attention should be paid to creating awareness and education. The same kind of advice can be found in the report on the SAFT project. (http://www.saftonline.org/). SAFT can be considered as the first large scale, European study on the subject of awareness for and by youngster.

As a result of these two reports the expert debates on safer internet are now changing at least those in the Netherlands - from proposals on legislation, forcing providers and publishers to work together on content rating, to a discussion on the role of education for all internet users.

The EC has announced to support initiatives for a safer internet for the next years with 45 million Euro. Although much of the money will be spent on technological solutions, there is also a considerable amount for educational purposes.

by Rashid Niamat, NLIP (Dutch Internet Service Provider Association)
Editor: Heather Weaver, CCLRC, UK

CWI— **Lex Schrijver** wins Lanchester Prize. The Institute for Operations Research and the Management Sciences (INFORMS) has awarded the 2004 Frederick W. Lanchester Prize to Lex Schrijver. Schrijver, leader of CWI's Probability, Networks, and Algorithms cluster, received the award on October 25, 2004 at the INFORMS Annual Meeting in Denver. The Lanchester prize is awarded for the best contribution to operations research and management sciences published in English. It includes a commemorative medallion and a cash award. Schrijver earned the prize for his book 'Combinatorial Optimization -Polyhedra and Efficiency', published in 2003 by Springer Verlag. It is the second time Schrijver wins the Lanchester Prize. In 1987 he received the award for his book Theory of Linear and Integer Programming. More information can be found on PNA1 (http://www.cwi.nl/pna1), Springer online (http://www.springeronline.com/sgw/cda/frontpage/0,11855,4-40109-22-2250233-0,00.html) or INFORMS (http://www.informs.org/Prizes/LanchesterDetails.html#).


Photo: CWI

Fraunhofer ICT Group — **Ina Schieferdecker** has been awarded the Alfried Krupp Prize. The award honors outstanding young university teachers in natural sciences and engineering sciences, and is endowed with 500,000. Ina Schieferdecker is head of the Competence Center for Testing, Interoperability and Performance (TIP) at the Fraunhofer Institute for Open Communication Systems FOKUS in Berlin, and an associate of its spin-off Testing Technologies IST. In November 2003 she also became Fraunhofer Professor at the Institute for Telecommunication Systems of the Technical University of Berlin where she heads the department of design and testing. Ina Schieferdecker is a member of the ERCIM Working Group on Formal Methods for Industrial Critical Systems.


© FOKUS

VTT Information Technology — **Jussi Tuovinen** has been appointed research director for VTT Infor-mation Technology from 1st November. Until now he has been director for MilliLab-Milli-metre Wave Laboratory of Finland, which is a joint venture between VTT and Helsinki University of Technology and an ESA external laboratory. Pekka Silvennoinen continues as executive director.

SpaRCIM — **Software Engineering and Production Technologies Society founded**. For the first time in Spain, a scientific society about Software Engineering and Software Production Technologies (SE&PT) has been founded with the aim of contributing to scientific and technological advances in these subjects. The society (the Spanish name is SISTEDES) is a non-profit scientific organisation. SISTEDES will become the scientific voice vis-a-vis the public sector and governmental institutions. The aim is to promote research, innovation and technological transfer between the different sectors and to establish relationships and agreements with other international societies. To accomplish these goals an annual multiconference (called CEDI) will be organized. 22 different conferences will run at the same time. Many of the SISTEDES society founders are ERCIM members, among them the SISTEDES president and vice-president.


Photo: CWI

CWI — **Ronald Cramer** joins Dutch Young Academy of Arts and Sciences. The Royal Netherlands Academy of Arts and Sciences KNAW has selected CWI researcher Ronald Cramer as member of their Young Academy. The Young Academy, which will be installed in February 2005, consists of 40 young, distinguished researchers. Its goal is to stimulate contact between scientists from different disciplines. Ronald Cramer joined CWI in June this year as leader of the new Cryptology and Information Security theme. In June he was also appointed part-time professor of Cryptology at Leiden University.

NTNU — **Norwegian University of Science and Technology students win regional programming contest.** What's faster at computer programming than a three-headed monkey? The NTNU's winning computer programming student team /dev/duff/, which took first place at the 2004 Northwestern Europe Programming Contest by beating Sweden's Royal Institute of Technology team 'Three Headed Monkey'. The contest was held 14 November 2004 at the Lund Institute of Technology in Lund, Sweden.

The winners, Børge Mikkelsen, Øyvind Grotmol and Erling Ellingsen, along with coach Nils Grimsmo, will travel to Shanghai, China for the 2005 finals in April 2005. The Association for Computer Machinery's (ACM) International Collegiate Programming Contest is the oldest, largest, computer programming contest in the world. This year's annual contest, the 29th, has drawn more than 3,000 student teams from 1,400 universities worldwide. Students compete in regional contests like the one in Sweden to narrow the field; the 75 best teams – including NTNU's victors – will compete in the 2005 finals. The victory is a good example of the strength that can come from interdisciplinary cooperation, because the team members come from three different departments at NTNU. Mikkelsen is a student from the Department of Mathematical Sciences, Grotmol is from the Department of Electronics and Telecommunications, and Ellingsen is from the Department of Computer and Information Science.

ERCIM – The European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development, in information technology and applied mathematics. Its national member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.

ERCIM is the European Host of the World Wide Web Consortium.

Austrian Association for Research in IT
c/o Österreichische Computer Gesellschaft
Wollzeile 1-3, A-1010 Wien, Austria
Tel: +43 1 512 02 35 0, Fax: +43 1 512 02 35 9
http://www.aarit.at/

Council for the Central Laboratory of the Research Councils, Rutherford Appleton Laboratory
Chilton, Didcot, Oxfordshire OX11 0QX, United Kingdom
Tel: +44 1235 82 1900, Fax: +44 1235 44 5385
http://www.cclrc.ac.uk/

Consiglio Nazionale delle Ricerche, ISTI-CNR
Area della Ricerca CNR di Pisa,
Via G. Moruzzi 1, 56124 Pisa, Italy
Tel: +39 050 315 2878, Fax: +39 050 315 2810
http://www.isti.cnr.it/

Czech Research Consortium
for Informatics and Mathematics
FI MU, Botanicka 68a, CZ-602 00 Brno, Czech Republic
Tel: +420 2 688 4669, Fax: +420 2 688 4903
http://www.utia.cas.cz/CRCIM/home.html

Centrum voor Wiskunde en Informatica
Kruislaan 413, NL-1098 SJ Amsterdam,
The Netherlands
Tel: +31 20 592 9333, Fax: +31 20 592 4199
http://www.cwi.nl/

Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
Tel. +352 26 19 25-1, Fax +352 26 1925 35
http:// www.fnr.lu

FWO
Egmontstraat 5
B-1000 Brussels, Belgium
Tel: +32 2 512.9110
http://www.fwo.be/

FNRS
rue d'Egmont 5
B-1000 Brussels, Belgium
Tel: +32 2 504 92 11
http://www.fnrs.be/

Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
Tel: +30 2810 39 16 00, Fax: +30 2810 39 16 01
http://www.ics.forth.gr/

Fraunhofer ICT Group
Friedrichstr. 60
10117 Berlin, Germany
Tel: +49 30 726 15 66 0, Fax: ++49 30 726 15 66 19
http://www.iuk.fraunhofer.de

Institut National de Recherche en Informatique
et en Automatique
B.P. 105, F-78153 Le Chesnay, France
Tel: +33 1 3963 5511, Fax: +33 1 3963 5330
http://www.inria.fr/

Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and
Electrical Engineering, N 7491 Trondheim, Norway
Tel: +47 73 59 80 35, Fax: +47 73 59 36 28
http://www.ntnu.no/

Spanish Research Consortium for Informatics
and Mathematics c/o Esperanza Marcos, Rey Juan Carlos
University, C/ Tulipan s/n, 28933-Móstoles, Madrid, Spain,
Tel: +34 91 664 74 91, Fax: 34 91 664 74 90
http://www.sparcim.org

Swedish Institute of Computer Science
Box 1263,
SE-164 29 Kista, Sweden
Tel: +46 8 633 1500, Fax: +46 8 751 72 30
http://www.sics.se/

Swiss Association for Research in Information Technology
c/o Prof. Dr Alfred Strohmeier, EPFL-IC-LGL,
CH-1015 Lausanne, Switzerland
Tel +41 21 693 4231, Fax +41 21 693 5079
http://www.sarit.ch/

Slovak Research Consortium for Informatics and
Mathematics, Comenius University, Dept.of Computer
Science, Mlynska Dolina M, SK-84248 Bratislava, Slovakia
Tel: +421 2 654 266 35, Fax: 421 2 654 270 41
http://www.srcim.sk

Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
Tel: +36 1 279 6000, Fax: + 36 1 466 7503
http://www.sztaki.hu/

Irish Universities Consortium
c/o School of Computing, Dublin City University
Glasnevin, Dublin 9, Ireland
Tel: +3531 7005636, Fax: +3531 7005442
http://ercim.computing.dcu.ie/

Technical Research Centre of Finland
P.O. Box 1200
FIN-02044 VTT, Finland
Tel:+358 9 456 6041, Fax :+358 9 456 6027
http://www.vtt.fi/tte

Order Form

If you wish to subscribe to ERCIM News ***free of charge*** *or if you know of a colleague who would like to receive regular copies of ERCIM News, please fill in this form and we will add you/them to the mailing list.*

Send, fax or email this form to:

ERCIM NEWS
2004 route des Lucioles
BP 93
F-06902 Sophia Antipolis Cedex
Fax: +33 4 9238 5011
E-mail: office@ercim.org

Data from this form will be held on a computer database. By giving your email address, you allow ERCIM to send you email

I wish to subscribe to the

❑ ***printed edtion*** ❑ ***online edition (email required)***

Name:

Organisation/Company:

Address:

Post Code:

City:

Country

E-mail:

You can also subscribe to ERCIM News and order back copies by filling out the form at the ERCIM website at http://www.ercim.org/publication/Ercim_News/