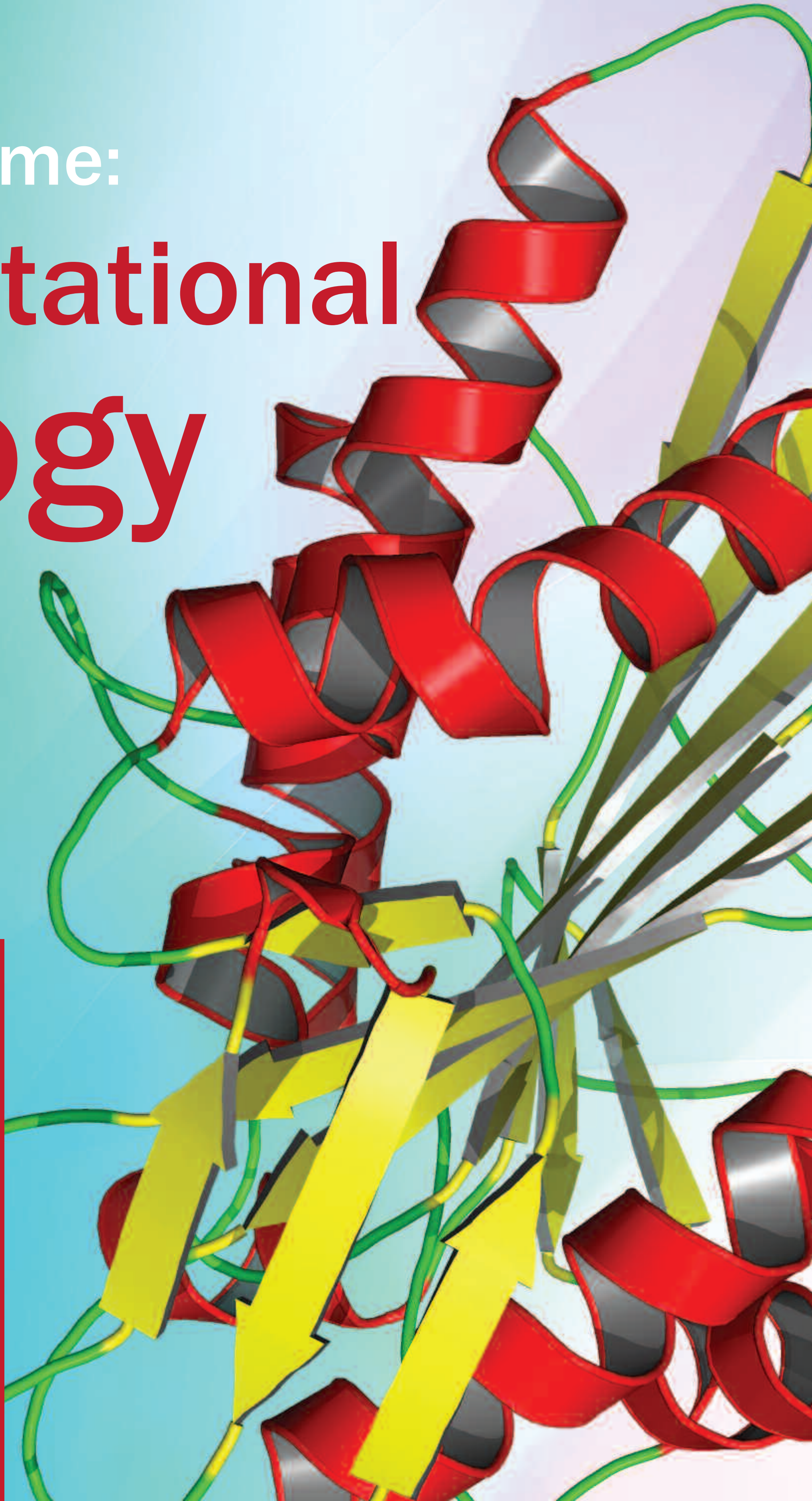
Number 82, July 2010

ERCIM NEWS

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

Special theme:

Computational Biology



Also in this issue:

Keynote:
Computational Biology – On the Verge of Widespread Impact
by Dirk Evers

European Scene:
Scientometry Leading us Astray

R&D and Technology Transfer:
Continuous Evolutionary Automated Testing for the Future Internet

ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. 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 9,000 copies. The printed version of ERCIM News has a production cost of €8 per copy. Subscription is currently available free of charge.

ERCIM News is published by ERCIM EEIG
BP 93, F-06902 Sophia Antipolis Cedex, France
Tel: +33 4 9238 5010, E-mail: contact@ercim.eu
Director: Jérôme Chailloux
ISSN 0926-4981

Contributions
Contributions must be submitted to the local editor of your country

Copyright Notice

Advertising
For current advertising rates and conditions, see
http://ercim-news.ercim.eu/ or contact peter.kunz@ercim.eu

ERCIM News online edition
The online edition is published at http://ercim-news.ercim.eu/

Subscription
Subscribe to ERCIM News by sending email to
en-subscriptions@ercim.eu or by filling out the form at the
ERCIM News website: http://ercim-news.ercim.eu/

Next issue
October 2010, Special theme: "Cloud Computing"

Cover image: Three-dimensional homology model of the human Kinesin-like protein KIF3A (See the article "Protein Homology Modelling - Providing Three-dimensional Models for Proteins where Experimental Data is Missing" by Jürgen Haas and Torsten Schwede on page 20

ERCIM
European Research Consortium
for Informatics and Mathematics

NTNU
danaim
CWI
SICS
VTT
Science & Technology Facilities Council
IUA
Fraunhofer IUK-TECHNOLOGIE
FWO
FNRS
ERCIM
INRIA
PEG
AARIT
SpaRCIM
SARIT
CNR
FORTH
Cooperating for Excellence in Research

Dr. Dirk J. Evers
Director, Computational Biology
Illumina, Inc.

Computational Biology – On the Verge of Widespread Impact

Computational Biology is defined as the science of understanding complex biological phenomena by the analysis of multi-sample and multi-variate quantitative data. With the advent of high-throughput measuring devices, particularly the new generation of DNA sequencers, the cost of generating data has dropped several orders of magnitude during the last 10 years. There is no sign of this trend slowing down. What will be the effect of cheap, fast, simple, robust, high-quality DNA sequencing? Computational Biology applications will become ubiquitous. It is well-known that Biotech and Pharma have been utilizing Computational Biology approaches successfully for some time, but now we see the technology and research spreading into industries such as Environment Management, Agrotech, Biofuels, Personalized Medicine and Consumer Genomics, to name just a few.

In preparing this keynote I perused keynote contributions in previous editions of the ERCIM news. Here are some of the recent topics: Simulating & Modelling, Digital Preservation, Towards Green ICT, Future Internet Technology, The Sensor Web, Safety-Critical Software, and Supercomputing at Work. It is very clear that the advances described therein are of fundamental importance if Computational Biology and Molecular Biology are to achieve their full impact in the coming years. We will succeed only if we can acquire samples ubiquitously, simulate and model precisely, compute efficiently and safely. We must also archive the high volumes of data efficiently while providing authorized and fast access, and present the data via the Internet in a meaningful way to all interested parties. Computational Biology is an interdisciplinary science by definition. Nonetheless, it is clear that the future complexity of this scientific area and its technology goes well beyond the core definition given above.

There is a strong European tradition of Computational Biology and Bioinformatics going back to the inception of the first Bioinformatics curriculum by Prof. Robert Giegerich (Bielefeld University, Germany) in the late 1980s which I was lucky enough to attend. Of course, research in Mathematics and Informatics for Biology and Biochemistry goes back much further than that, with too many prominent examples to single out any one specifically. EU funding organizations will need to take great care and oversight to compete successfully with North America and China in this space.

To be clear, in the next decade the impact of this science on government decisions as well as personal lifestyle and medical decisions will be profound. The EU, governments and societies will need to ensure that they are prepared for future advances in high-throughput instrumentation and Computational Biology. Research organisations and industry will have to provide safe technology and robust data interpretation. Educational institutions will need to update curricula to enable graduates, medical doctors, and scientists to harness the new technology and scientific findings at their disposal.

Computational Biology Scientists beware! Your science is leaving the safe haven of Research-Use-Only to applications of widespread public use.

Dirk J. Evers

KEYNOTE

JOINT ERCIM ACTIONS

EUROPEAN SCENE

SPECIAL THEME

Computational Biology

coordinated by: Gunnar Klau, CWI, The Netherlands
and Jacques Nicolas, INRIA, France

Introduction to the Special Theme

Invited articles

Sequences and next generation sequencing

Structural biology

Drug design and discovery

Understanding gene expression

Networks

Modelling and simulation of biological systems

Epidemiology

Text mining, semantic technologies, databases

R&D AND TECHNOLOGY TRANSFER

EVENTS

IN BRIEF

Fourth ERCIM Workshop on eMobility

by Torsten Braun

The 4th ERCIM Workshop on eMobility was held at Luleå University of Technology, Sweden on 31 May 2010, the day prior to the Wired/Wireless Internet Communications (WWIC) 2010 conference. The twenty six participants engaged in very intense discussions, making the workshop a lively event.

As in previous years, the program included both invited presentations and presentations of papers that had been selected after a thorough review process. The focus of the invited presentations was on current EU FP7 projects, in which ERCIM EMobility members are actively involved.

Marc Brogle (SAP Zürich) opened the workshop with a presentation of the ELVIRE (ELectric Vehicle communication to Infrastructure, Road services and Electricity supply) project, which aims to utilize ICT to allow smooth introduction of electric vehicles. Drivers will be supported in mobility planning and during travelling with an electric car. While ELVIRE started recently, the EU-Mesh project (Enhanced, Ubiquitous, and Dependable Broadband Access using MESH Networks), presented by Vasilios Siris (FORTH), will finish soon. Its goal is to develop, evaluate, and trial software modules for building dependable multi-radio multi-channel mesh networks that provide ubiquitous and high speed broadband access. Another project in the area of wireless networks has been SOCRATES (Self-Optimisation and self-ConfiguRATion in wirelEss networkS, presented by Hans van den Berg, TNO). This project aims to use self-optimisation, self-configuration and self-healing to decrease operational expenditure by reducing human involvement in network operation tasks, while optimising network efficiency and service quality. Two of the presented projects focussed on wireless sensor networks. The first, called GINSENG and presented by Marilia Curado (U Coimbra), investigates technological challenges of wireless sensor networks in industrial environments such as oil refineries. The second, called WISEBED (Wireless Sensor Network Testbeds) and presented by Torsten Braun (U Bern), developed a pan-European wireless sensor network, currently consisting of several hundred sensor nodes, which can be accessed by sensor network researchers via the Internet. Finally, Yevgeni Koucheryavy (Tampere UT) gave a talk on the recently established COST Action IC 0906 on “Wireless Networking for Moving Objects”, an initiative launched by ERCIM eMobility members.

In addition to the invited talks three technical sessions were held in the afternoon. Twelve out of 16 submitted long, short, and abstract papers were selected. The first session on “Mobile Networks” looked into various aspects of vehicle-to-vehicle/business communication and scheduling on LTE uplinks. Another talk presented a testbed to evaluate IMS (IP Multimedia Subsytem) implementations, whilst a separate session was devoted to IMS. One presentation investigated issues of the integration of IMS and digital IP-TV services, whilst another focussed on the integration of IMS with Web 2.0 technologies in order to support communications between family members. The final session focused on “Wireless Networks and Geocasting”. Topics included: routing and reliability in mobile ad-hoc networks, geocasting for vehicular ad-hoc networks, deployment support for wireless mesh networks, realistic traffic models for WiMAX or cellular networks as well as resource discovery in cognitive wireless networks.

Workshop participants.

A nice social event attended by nearly all participants under the bright evening sun in North Sweden concluded the event, which was perfectly organized by the local hosts under the guidance of Evgeny Osipov. Again, printed workshop proceedings could be handed out to all participants.

Link:
http://wiki.ercim.eu/wg/eMobility

Please contact:
Torsten Braun
ERCIM e-Mobility Working Group coordinator
Bern University / SARIT, Switzerland
E-mail: braun@iam.unibe.ch

SERENE 2010 - Second International Workshop on Software Engineering for Resilient Systems

by Nicolas Guelfi

The SERENE 2010 workshop, organized by the ERCIM Working Group on Software Engineering for Resilient Systems was held in London at Birbeck College on 13-16 April 2010. The workshop was co-chaired by Dr. Giovanna Di Marzo Serugendo, Computer Science and Information Systems - Birkbeck - University of London, and Dr. John Fitzgerald, School of Computing Science, Newcastle University. A two-day Spring School on resilience and self-* was organized for the first time ahead. An additional event, an evening lecture given by Robin Bloomfield of City University, London, held jointly with BCS-FACS and the UK Safety-Critical Systems Club, was added to the workshop programme and addressed modelling on infrastructure interdependencies and their impact on resilience.

Spring school

Thirty participants registered for the Spring School. The four speakers - Paola Inverardi, University of L'Aquila, Italy; Jeff Magee, Imperial College London, UK; Aad van Moorsel, University of Newcastle, UK; Mauro Pezzè, University of Lugano, Switzerland, and University of Milano Bicocca, Italy - were uniformly excellent and the level of interaction from participants was good. Several of the invited speakers remarked on the collaborative atmosphere. The event has helped to position SERENE better with the invited speakers and with student and early career scientists. This event shows a real engagement of the ERCIM Working Group with the discipline. It should be renewed for future editions.

Workshop

By category, we had twelve technical papers, one industrial/experience paper; and three project papers. No papers were submitted in the student and tools categories. We offered four reviews per paper and were met with very high compliance by the programme committee members. We accepted eleven submissions. This permitted a programme of one full day on the technical papers and a half day on the shorter papers.

Two keynotes where organized both from "non-traditional" fields - one from business information systems by Andreas Roth and one from swarm robotics by Alan F T Winfield.

SERENE 2010 had 34 registrations and was a succesfull event. For the SERENE 2011 we received two proposals: one from Professor Vyacheslav Kharchenko, National Aerospace University "Khai", department of computer systems and communications, to host the next SERENE workshop in Kirovograd (Ukraine); and one from Professor Buchs, University of Geneva, Switzerland. Eventually it was decided that SERENE 2011 will be held in Geneva in September 2011, co-chaired by Professor Didier Buchs and Dr. Elena A. Troubitsyna, Abo Akademi University, Finland.

Link:
http://serene2010.uni.lu/

Please contact:
Nicolas Guelfi
SERENE Working Group coordinator
Luxembourg University
E-mail: Nicolas.Guelfi@uni.lu

ERCIM-sponsored Conferences

Joint MFCS & CSL 2010

The 35th International Symposium on Mathematical Foundations of Computer Science is organized in parallel with the 19th EACSL Annual Conferences on Computer Science Logic (CSL 2010) 23-27 August 2010. The federated MFCS & CSL 2010 conference and its satellite workshops will be held at the Faculty of Informatics of Masaryk University which is part of the Czech ERCIM member CRCIM.
More information: http://mfcsl2010.fi.muni.cz/mfcs

ECCV 2010

The 11th European Conference on Computer Vision (ECCV 2010) will be hosted by the Greek ERCIM member FORTH, Crete, Greece, from 5-11 September 2010. ECCV is a selective single-track conference on computer vision where high quality previously unpublished research contributions on any aspect of computer vision will be presented.
More information: http://www.ics.forth.gr/eccv2010/

SAFECOMP 2010

SAFECOMP, the International Conference on Safety, Reliability and Security of Computer Systems, is held this year from 14-17 September in Vienna, at the Schönbrunn Palace Conference Center, a very attractive venue. Co-organizer is the ERCIM Working Group on Dependable Embedded Software-Intensive Systems. ERCIM and the Austrian ERCIM member AARIT are co-sponsors of the event and represented with a booth.
More information: http://www.ocg.at/safecomp2010

Opinion

Scientometry Leading us Astray

by Michal Haindl

The entire spectrum of recent scientific research is too complex to be fully and qualitatively understood and assessed by any one person, hence, scientometry may have its place, provided it be applied with common sense, being constantly aware of its many limits. However, scientometry, when applied as the sole basis for evaluating performance and funding of research institutions, is our open admission that we are unable to recognize an excellent scientist even if he is working nearby.

In recent years Czech scientists have been evaluated by the powerful Research and Development Council (RDC), an advisory body to the Government of the Czech Republic. The RDC processes regular annual analyses and assessments of research and development work and proposes financial budgets for single Czech research entities. The RDC's over-ambitious aim was to produce a single unique universal numerical evaluation criterion, a magic number used to identify scientific excellence and to determine the distribution of research funding. The resulting simplistic scientometric criterion is shown in Figure 1.

$$\vartheta = \sum_{i=1}^{10} n_i \left[c_i\,\omega_i + d_i\right]$$

i	result type	ω_i	d_i	c_i
1	impacted journal	295	10	$c_i \in < 0; 1 >$
2	EU/US/JP patent | Nature, Science, Proc.Nat.Acad.Sci.	500	0	1
3	acknowledged journal (Scopus / ERIH)	8	0	1.5 (h) | 1
4	reviewed journal	4	0	2.5 (h) | 1
5	book(world lang.) | F,G,H,N,L,R | other patent	40	0	1
6	book	20	0	2 (h) | 1
7	proceeding article	8	0	1
8	CZ/national patent	200	0	1
9	pilot plant	100	0	1
10	classified research report	50	0	1

Figure 1: RDC scientometric criterion, where ni is number of results, h humanities, F design, G prototype, H legal norm, N methodology, L map, and R software result, respectively. The coefficient ci value depends on a journal impact rank inside its discipline.

The RDC noticed difficulties in mutual comparison of humanities, natural, and technical sciences so they introduced a corrective coefficient (c_i) favouring less prestigious journals and books in humanities. Apart from the fact that it is impossible to compare a range of diverse disciplines using a single evaluation criterion, this system ignores some common assessment criteria (eg citations, grant acquisition, prestigious keynotes, comparison with the state-of-the-art, editorial board memberships, PhDs). Furthermore, single weights (ω_i) are unsubstantiated, and acquired scores cannot be independently verified.

Negative consequences have been revealed over the last three years. Significant rank changes indicate that scientific excellence can easily be simulated with a cunning tactic. For example, eight unverified software pieces (category R) will outweigh a scientific breakthrough published in the most prestigious high-impact journal within the research area. Some research entities have already opted for such an easy route to the government research pouch.

Publication cultures vary between disciplines; while some prefer top conferences (computer graphics), elsewhere prestige goes to journals (pattern recognition) or books (history and the arts). Some disciplines have one or two authors per article (mathematics) while others typically have large author teams (medicine). These factors are not taken into consideration by the current system.

Finally, this system recklessly ignores significant differences between the financing systems of different types of research entities. While universities have a major source of financing from teaching, the 54 research institutes of the Academy of Sciences of the Czech Republic (ASCR) rely solely on research. Thus last year's application of the scientometric criterion resulted in anticipated negative consequences for the Czech research community in general, but especially for its most effective contributor - the Academy of Sciences. ASCR, with only 12% of the country's research capacity, produces 38% of the country's high impact journal papers, 43% of all citations, and 30% of all patents. Nevertheless the criterion led to 45% drop in funding for ASCR in 2012. It is little wonder then, that our research community has become destabilized, that researchers, for the first time in history, have participated in public rallies, and that young researchers are losing motivation to pursue scientific careers. Over 70 international and national research bodies have issued strong protest against the government's policies (see details on http://ohrozeni.avcr.cz /podpora/), and several governmental round tables have been held in an attempt to solve this crisis.

What is the moral from the described scientometric experiment? Any numerical criterion requires humble application, careful feedback on its parameters, and wide adoption by the scientific community. We must never forget that any criterion is only a very approximate, and considerably limited, model of scientific reality and will never be an adequate substitute for peer review by true experts. In order to avoid throwing out the baby with the bath water we should consult any criterion on a discipline-specific basis, and only as an auxiliary piece of information. A criterion may be useful to compare extreme scientific performance, but is hardly suitable for assessing subtle differences or breakthrough scientific achievements. Whilst a single number has alluring simplicity, it also carries pricey long-lasting consequences. Excellent research teams are easy to damage by ill-conceived scientometry, but much harder to rebuild.

Links:
http://www.vyzkum.cz/
http://ohrozeni.avcr.cz/podpora/

Please contact:

Michal Haindl - CRCIM (UTIA)
Tel: +420-266052350
E-mail: haindl@utia.cas.cz

ERCIM
European Research Consortium
for Informatics and Mathematics

"Alain Bensoussan"
Fellowship Programme

Application deadline:
Twice per year
30 April and 30 September

Photo by courtesy of INRIA.

Who can apply?

The Fellowships are available for PhD holders from all over the world. ERCIM encourages not only researchers from academic institutions to apply, but also scientists working in industry.

Why to apply?

The Fellowship Programme enables bright young scientists from all over the world to work on a challenging problem as Fellows of leading European research centres. The programme offers the opportunity to ERCIM Fellows:

- to work with internationally recognized experts
- to improve their knowledge about European research structures and networks
- to become familiarized with working conditions in leading European research centres
- to promote cross-fertilisation and cooperation, through the Fellowships, between research groups working in similar areas in different laboratories.

What is the duration?

Fellowships are generally of 18 month duration, spent in two of the ERCIM institutes. In particular cases a Fellowship of 12 month duration spent in one institute might be offered.

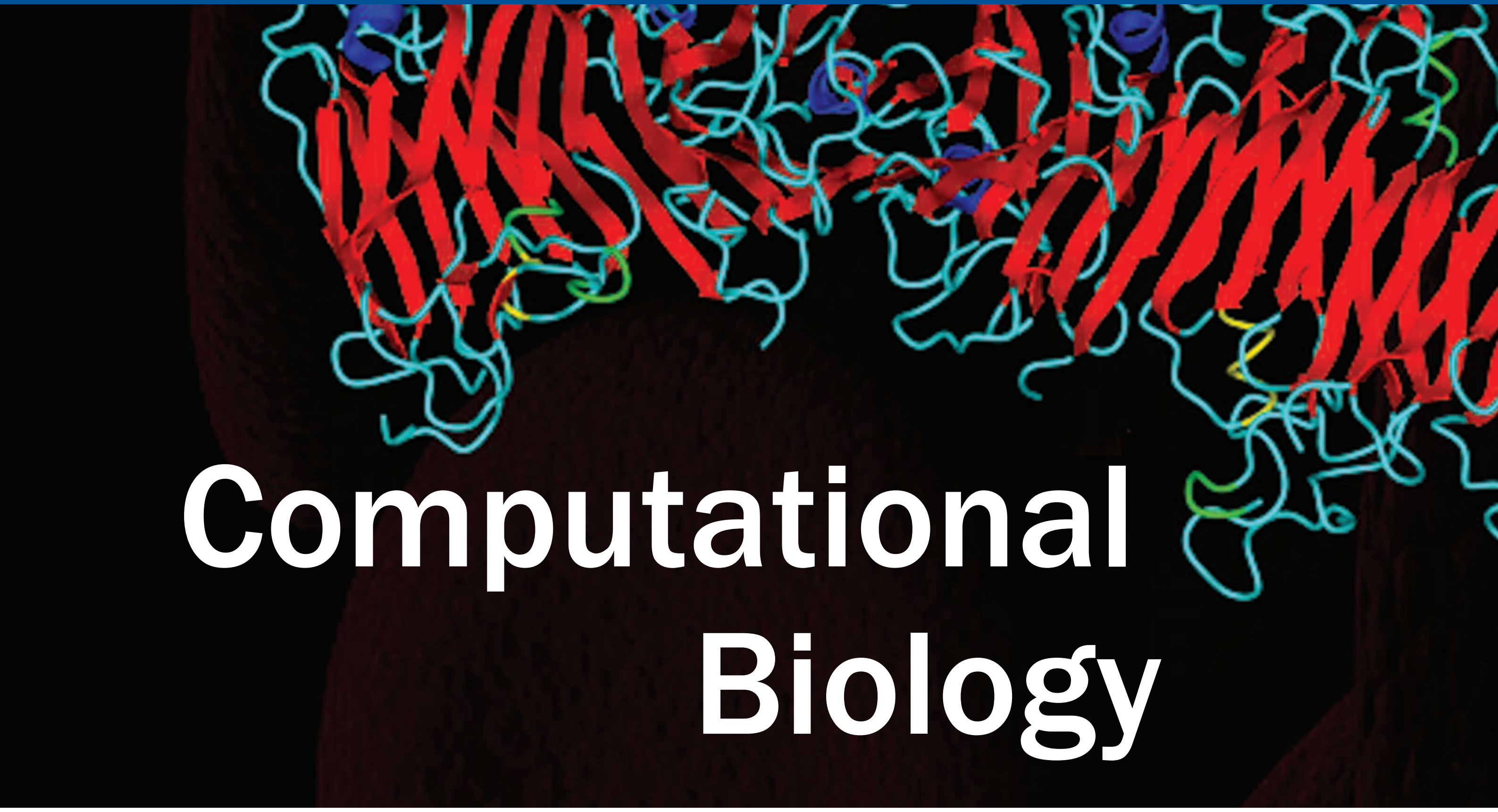
How to apply?

Applications have to be submitted online. The application form will be available one month prior to the application deadline at http://www.ercim.eu/activity/fellows/

Which topics/disciplines?

The Fellowship Programme focuses on topics defined by the ERCIM working groups and projects managed by ERCIM such as Biomedical Informatics, Computing and Statistics, Constraints Technology and Applications, Embedded Systems, Digital Libraries, Environmental Modelling, E-Mobility, Formal Methods, Grids, Security and Trust Management, and many other areas in which ERCIM institutes are active.

http://www.ercim.eu/activity/fellows/



Computational Biology

Introduction to the Special Theme

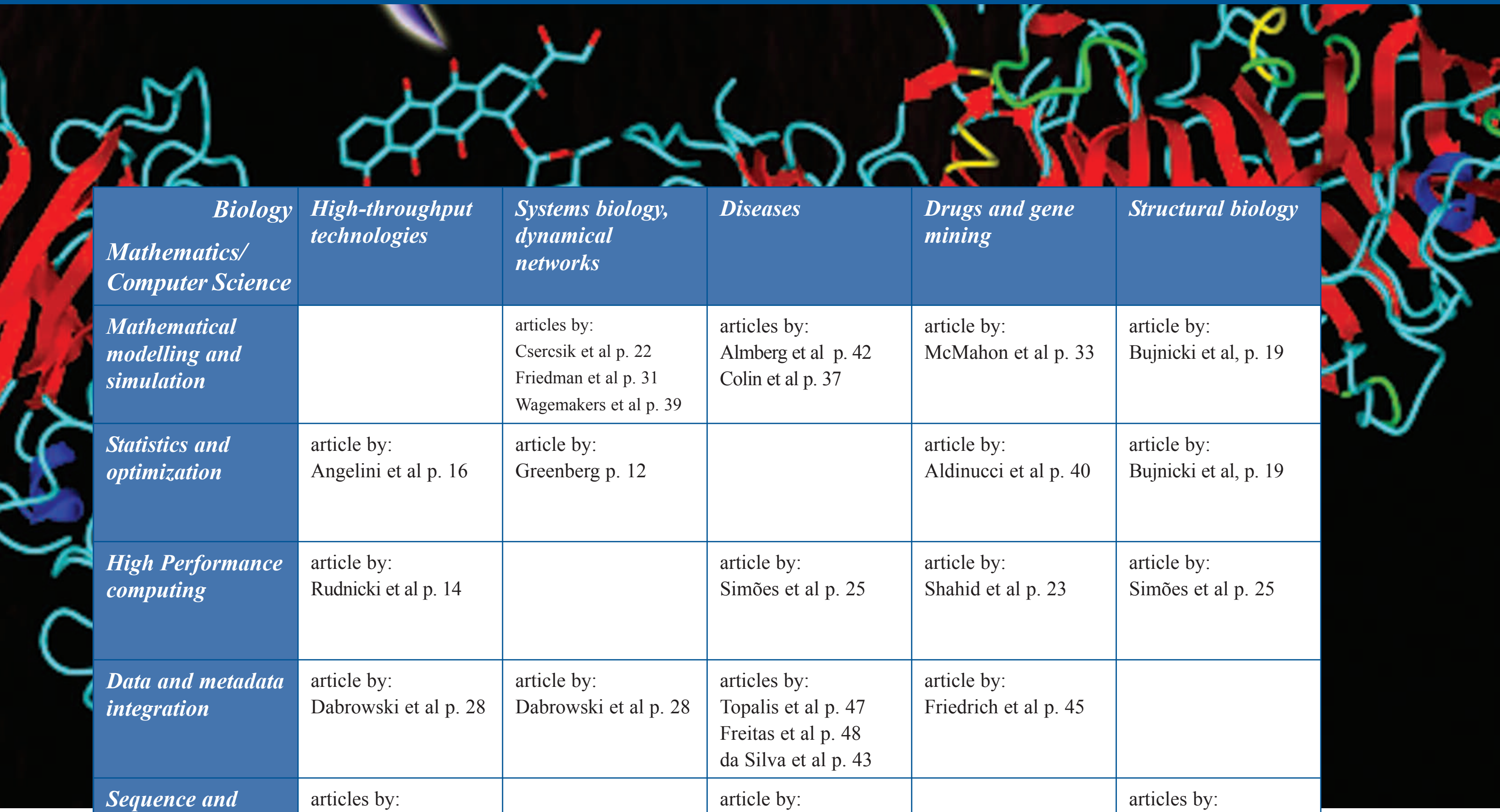
by Gunnar W. Klau and Jacques Nicolas

In the life sciences, conventional wet lab experimentation is being increasingly accompanied by mathematical and computational techniques in order to deal with the overwhelming complexity of living systems. Computational biology is an interdisciplinary field that aims to further our understanding of living systems at all scales. New technologies lead to massive amounts of data as well as to novel and challenging research questions and more and more biological processes are analyzed and modelled with the help of mathematics and can be simulated *in silico*.

Computational biology is one side of a two-sided domain and generally refers to the fundamental research field. The term bioinformatics is frequently used for the more engineering-oriented side of the domain and deals with the production of practical software enabling original analysis of biological data. This ERCIM News special theme starts with two invited articles that reflect this complementarity. Harvey Greenberg presents his view on the positive contribution of Operations Research (OR) to biology. OR is in itself a powerful, interdisciplinary domain that led, among many other important contributions to computational biology, to the recent concept of pathway signatures. Knut Reinert describes the importance of good software design practices in bioinformatics by means of library design for sequence analysis. In fact, the next European infrastructure in bioinformatics, which is currently discussed in the project ELIXIR (European Life Sciences Infrastructure for Biological Information), has put emphasis on the necessity of shared efforts for setting common standards and tools for data management.

As already mentioned in the keynote by Dirk Evers we currently observe a neat evolution of the research field that results from the maturation and large diffusion of high-throughput technologies in biological laboratories. New types of data have to be taken into account: researchers are becoming increasingly aware of the importance of the RNA world and epigenomics in explaining cellular behaviour that can not be solely understood by purely genetic aspects. Metabolomics is the first method developed that is capable of obtaining high throughput data at the phenotypic level. Examples of joint European research efforts in this direction include the European Research Council Advanced Grants “RIBOGENES” (The role of noncoding RNA in sense and antisense or orientation in epigenetic control of rRNA genes) and the FP7 project “METAcancer (Identification and validation of new breast cancer biomarkers based on integrated metabolomics)”.

These recent developments make more systemic approaches possible, where several methods and sources of data have to be combined. This has an impact on the importance of developing data integration environments, adding semantics to observations (ontologies, text mining) and offering sophisticated navigation utilities through the web. It has also fostered a number of studies in high-performance computing such as distributed or grid computing or the use of graphical processing units. These techniques are no longer restricted to demanding applications in structural modelling but have become necessary in many other domains of computational biology.



Biology *Mathematics/ Computer Science*	*High-throughput technologies*	*Systems biology, dynamical networks*	*Diseases*	*Drugs and gene mining*	*Structural biology*
Mathematical modelling and simulation		articles by: Csercsik et al p. 22 Friedman et al p. 31 Wagemakers et al p. 39	articles by: Almberg et al p. 42 Colin et al p. 37	article by: McMahon et al p. 33	article by: Bujnicki et al, p. 19
Statistics and optimization	article by: Angelini et al p. 16	article by: Greenberg p. 12		article by: Aldinucci et al p. 40	article by: Bujnicki et al, p. 19
High Performance computing	article by: Rudnicki et al p. 14		article by: Simões et al p. 25	article by: Shahid et al p. 23	article by: Simões et al p. 25
Data and metadata integration	article by: Dabrowski et al p. 28	article by: Dabrowski et al p. 28	articles by: Topalis et al p. 47 Freitas et al p. 48 da Silva et al p. 43	article by: Friedrich et al p. 45	
Sequence and graph algorithms	articles by: Rivals p. 17 Reinert p. 13		article by: Simões et al p. 25		articles by: Bujnicki et al, p. 19 Haas et al p. 20 Simões et al p. 25
Artificial intelligence	article by: García-Nieto et al p. 27	articles by: Fages et al p. 36 Schaub et al p. 30	article by: Aldinucci et al p. 40	article by: Murphy et al p. 34	

Many innovative methods in the field come from health applications. A number of cohort studies are currently being carried out, for example within the 1000 genomes project or the Apo-sys (Apoptosis Systems Biology Applied to Cancer and AIDS) project. For the first time, they will give access to an in depth study of the correlations between a variety of health-related factors like human individual genomic variations, nutrition, environment and diseases. Deciphering the relationships between genotype and phenotype is a major challenge for the coming years that researchers are just starting to explore. Advances will be obtained by more connections between distant research fields, and techniques from image analysis and data mining can be expected to play increasing roles in computational biology in the future.

In addition, such global studies will more generally be beneficial for the fields of population genetics and ecology. Sets of cells in a tissue and sets of bacteria in a selected habitat can now be studied at the finest level of molecular interaction, creating pressure to develop research on multi-scale modelling and model reduction techniques. Here, examples of European projects are MetaHIT and Metaexplore for the metagenomics of the human intestinal tract and the study of the enzymatic potential in cryptic biota, respectively.

This special theme features a selection of articles that covers a number of areas of Computational Biology and provides a nice snapshot of the research variety carried out in this area in Europe. In addition to the two invited contributions, it contains 23 short articles on new approaches, frameworks and applications in the domain of computational biology. From the perspective of mathematics and computer science, the articles cover topics such as “mathematical modelling and simulation”, “statistics and optimization”, “high-performance computing”, “data and metadata integration”, “sequence and graph algorithms” and “artificial intelligence”. From the perspective of biology, the covered topics include: “high-throughput technologies”, “systems biology, dynamical networks”, “diseases”, “drugs and gene mining”, and “structural biology”. The table above gives an overview of this two-dimensional scheme.

Links:
1000 genomes: http://www.1000genomes.org
ELIXIR: http://www.elixir-europe.org
METAcancer: http://www.metacancer-fp7.eu/
Apo-sys: http://www.apo-sys.eu/
MetaHIT: http://www.metahit.eu/
MetaExplore: http://www.rug.nl/metaexplore/

Please contact:

Gunnar Klau
CWI, The Netherlands
E-mail: Gunnar.Klau@cwi.nl

Jacques Nicolas
INRIA, France
E-mail: jacques.nicolas@irisa.fr

Pathway Signatures

by Harvey J. Greenberg

How can operations research (OR), traditionally applied to management problems, help us to understand biological systems? OR emerged from WW II as not only a grab-bag of methods, but more importantly as a multi-disciplinary approach to problem-solving. Modern systems biology shares those same problem attributes, and the OR community is increasingly contributing to this frontier of medical research.

I entered computational biology when I visited Sandia National Laboratories (SNL) in 2000. Seeing the work of SNL researchers and learning about the problems showed me how operations research applies, particularly mathematical programming. This became the focus, not only of my new research, but also of my teaching and service. The following year I created the University of Colorado Center for Computational Biology and initiated a series of workshops and new courses in the departments of Mathematics, Computer Science and Engineering, and Biology. I was the main beneficiary! I continued to learn through research, teaching, and many new collaborations, particularly within the CU medical research community. In 2003 I visited Bernhard Palsson's Systems Biology Research Institute at San Diego and later visited Leroy Hood's Institute for Systems Biology. That is when I learned about systems biology and how much more OR can contribute. It is perhaps fortuitous that when I was a student, OR was almost synonymous with systems engineering (which has a different meaning in today's world of computers).

One trick I learned from OR practice is to turn questions around, "Why is this true?" We do not fully understand why certain pathways are used, and research has focused on estimating parameters of biochemical interactions among molecules. Turning this around, I ask, "Why is this particular pathway used?" This gave rise to what I call pathway signatures – the natural circumstances that trigger one particular pathway but not others.

I setup a mathematical program and identified pathways that are optimal for a particular objective (defined by parameters). An example of an objective is to maximize ATP production in a metabolic network; another is to maximize reliability (probability of successful regulation) in a cell-signaling network. A simplified version is to choose the cone of linear coefficients associated with an extreme pathway for which that pathway is uniquely optimal, hence its signature. Among the infinitely many, I favored certain properties (from biological principles) and chose signature values that minimize total similarity between pairs of extreme pathways. This is key – the dissimilarity strengthens the idea of a signature, separating one pathway from the others.

Non-extreme pathways have more than one representation as combinations of extremes, and I studied approaches to see what signature offers the most insight into natural choices. With help (and data) from Palsson's Lab, I enumerated extreme metabolic pathways (viz., red blood cell), on which I based my original studies. Another pursuit is

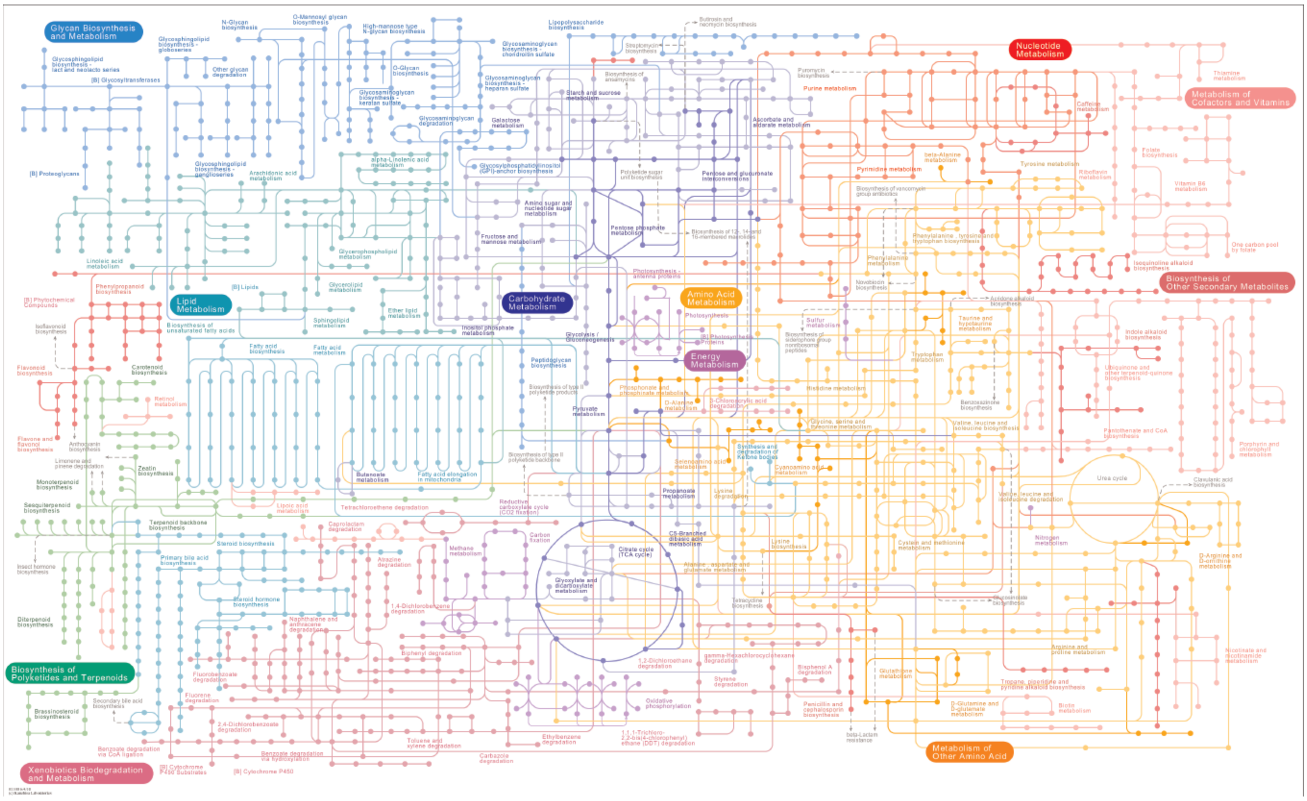

Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Biosynthesis of Other Secondary Metabolites
Biosynthesis of Polyketides and Terpenoids
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid
Citrate cycle (TCA cycle)
Purine metabolism
Pyrimidine metabolism

Figure 1: Metabolic pathways. From KEGG database http://www.genome.jp/kegg/

to consider nonlinear objectives, such as quadratic growth, for which each non-extreme pathway is optimal. Understanding why it is better than any of the extreme pathways that define it might shed light on the underlying biology. From the few experiments I ran, redundancy was a paramount signature criterion. For flux balance analysis this means that the defining input-output matrix of the reactions had metabolite generation/consumption dependent upon others. At the very least, we can use this inference as a measure of approximation error. When the linear approximation is reasonable, we can investigate how the system is affected by the imbalance that occurs without fixed system-outputs.

Another part of the story is the Pathway Inference Problem: Infer knowledge about pathways with incomplete information about their parts and interactions. In particular, there has been research into assembly – turning genes into pathways. I turn this into building a highway system and ask, “What junctions and connections provide the most efficient map for travel?” In this sense, pathway inference, or construction, is about optimal routing. Again, the idea of a signature is to find a biologically meaningful objective for which the network is designed (or revised) optimally. Much has been done in road and computer network design using OR with multiple objectives: minimize cost, minimize travel time, and maximize reliability, to name a few. There are many optimization principles at work in the design of life, so the natural laws of economics fit well into the OR descriptive motto: “the science of better.”

In conclusion, I see the future of systems biology broadening its network constructions to mixed-scale. The immune system, for example, needs to be represented by interactions among molecules, cells, tissues, and organs. A drug designed to block some pathways or enhance others can assume multiple targets where its behavior depends on the environmental context of the organism. We have seen such [re-]design problems in OR!

Links:
http://gcrg.ucsd.edu/
http://www.systemsbiology.org/

Please contact:

Harvey J. Greenberg, Professor Emeritus
University of Colorado Denver
E-mail: hjgreenberg@gmail.com

Oops I Did it Again..... Sustainability in Sequence Analysis via Software Libraries

by Knut Reinert

Maybe you like Britney Spears. Maybe even her music. Maybe you are a Britney Spears fan working as part of a group on sequence analysis algorithms in computational biology. But am I right in assuming that you don’t like to hear the above song title quoted by your coworkers or programmers when they could have been spending their time doing something productive or creative? If I am right, then you might want to read on because I will tell you how you, or your coworkers can avoid reinventing the wheel or writing a lot of inefficient scripts in sequence analysis.

Next generation sequencing (NGS) is a term coined to describe recent technological advances in DNA sequencing. Next generation sequencing allows us to sequence about 200.000.000.000 base pairs within approximately one week. That’s a two with many zeros. To explain it in different terms, whilst the human genome project spent many years and billions of dollars to sequence about 30 billion base pairs, we can now perform the equivalent amount of work within a day for a handful of dollars.

DNA as mass data has wonderful properties. While the sequencing machines churn out terabytes of data, a single byte of this data can be very important. It can decide whether you have a disease or not, whether the drugs you take do their job or not, whether you live or die. So we have to treat the data carefully. It is important to scientists working in the life sciences. At the same time, however, we need to process it fast and efficiently, after all it fills the racks of terabyte disks quickly.

In order to analyse the ever-growing volume of sequencing data it is essential that scientists in the life sciences and bioinformaticians work closely together with scientists in computer science. This can often be problematic since both communities approach the problems at hand quite differently (see Figure 1). While the the experimentalists have a holistic top-down view on what they want to achieve in a particular analysis, computer scientists usually work on a specific, small part of a larger analysis problem. On the computer science side this often results in highly efficient, but specialized algorithms that may not necessarily reflect the reality of real world data. Efforts from the life sciences side, in contrast, may result in analysis pipelines that compute a solution to the problem but are not state-of-the-art in run time or memory consumption, and hence cannot be applied to the large data volumes. The goal is obviously to use fast implementations of efficient algorithms to be able to cope with the volume of sequence data that NGS can produce. This can be achieved through algorithm libraries that collect efficient implementations of state-of-the-art algorithms, the work of algorithm designers and skilled programmers, and make them available to the data analysts and bioinformaticians. Apart from the obvious benefit of being able to efficiently compute solutions, the use of software libraries also avoids the Britney Spears effect, namely doing many things all over again, because many algorithms needed are already available in the library.

For the past seven years the Algorithmic Bioinformatics group at the FU Berlin has been working on a comprehensive algorithm library for sequence analysis.

The SeqAn project fills the gap between the experimentalists and algorithm designers. SeqAn is a C++ library and has a unique generic design based on:
- the generic programming paradigm
- a new technique for defining type hierarchies called template subclassing
- global interfaces
- metafunctions, which provide constants and dependent types at compile time.

The design of SeqAn differs from common programming practice, in particular SeqAn does not use object-oriented programming. The main consequence of this design choice is that SeqAn implements features like polymorphism at compile time thus avoiding costly run time operations like consulting a lookup table to find the appropriate virtual function, as it is necessary in object-oriented programming. This sets it also apart from other frameworks in the field like Galaxy (http://bitbucket.org/galaxy/galaxy-central/wiki/Home) or BioJava (http://biojava.org/wiki/Main_Page) and BioPerl (http://www.bioperl.org/wiki/Main_Page). SeqAn is intended to embrace high-performance algorithms from the computer science field and to cover a wide range of topics of sequence analysis. It offers a variety of practical state-of-the-art algorithmic components that provide a sound basis for the development of sequence analysis software. These include:
- data types for storing strings, segments of strings and string sets
- functions for all common string manipulation tasks
- data types for storing gapped sequences and alignments in memory and on disk
- algorithms for computing optimal sequence alignments
- algorithms for exact and approximate (multiple) pattern matching
- algorithms for finding common matches and motifs in sequences
- string index data structures
- graph types for many purposes like automata, alignment graphs, or HMMs
- standard algorithms on graphs.

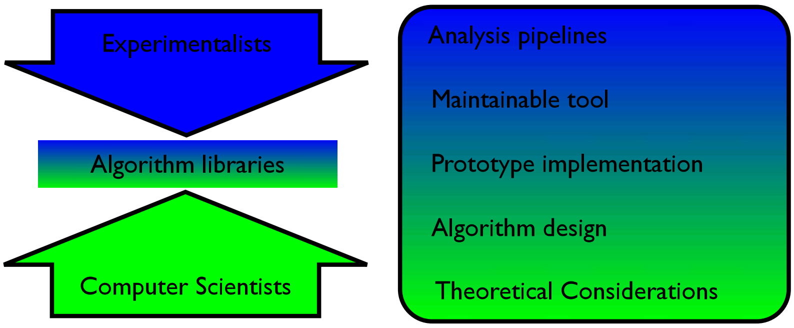


Figure 1: Top-down versus bottom-up approach.

SeqAn has growing user community throughout the EU and US and is actively developed by 4-6 scientists mainly at the FU Berlin. Potential users might be interested in the SeqAn book (see link below). SeqAn is partially supported by the DFG priority program 1307 “Algorithm Engineering”.

Links:
SeqAn project http://www.seqan.de
http://crcpress.com/product/isbn/9781420076233

Please contact:
Knut Reinert
Algorithmic Bioinformatics, Freie Universität Berlin, Germany
Tel: +49 30 838 75 222
E-mail: knut.reinert@fu-berlin.de

Application of Graphic Processors for Analysis of Biological Molecules

by Witold R. Rudnicki and £ukasz Ligowski

Graphic processors are used to improve the efficiency of a computational process in molecular biology in a project carried out by our team at the Interdisciplinary Centre for Mathematical and Computational Biology of the University of Warsaw, in cooperation with Professor Bertil Schmidt’s team from the School of Computer Engineering of the Nanyang Technological University in Singapore. The project aims to increase the efficiency of one of the most important algorithms in bioinformatics, the Smith Waterman algorithm.

The Smith Waterman (SW) algorithm, an algorithm for performing local sequence alignment, is used to establish similarity between biological macromolecules, such as proteins and nucleic acids. This is an exact algorithm giving the best possible result, but is much slower than BLAST (Basic Local Alignment Search Tool) - an alternative heuristic algorithm giving less precise answers but in a fraction of time. Availability of the fast version of SW algorithm with efficiency comparable to that of BLAST would enable researchers to take advantage of the greater sensitivity of the SW algorithm in the tasks currently performed with heuristic algorithms. This may be particularly important, for example, for gene assembly from short fragments in the new generation sequencing experiments or high sensitivity search for homologous proteins.

The evolution of graphics processing units (GPU) has resulted in transformation from chips specialised in the limited number of operations required for graphics, to massively parallel processors capable of delivering trillions of floating point operations per second (teraflops). The GPU's very high computational power is best utilised when and all threads perform identical tasks requiring many operations on the relatively small amount of data, which can be stored on on chip (in registers and a small pool of the very fast shared memory) The best practice in GPU programming is therefore to use data parallelism to formulate the problem, load the data to registers and shared memory, and perform all computations utilising registers and shared memory.

GPU has the potential to be particularly effective for applications such as scanning of large databases of sequences of biological macromolecules, because GPU works best for data-parallel algorithms, with a small but computationally intensive kernel.

Optimisation of the SW algorithm on a central processing unit (CPU) is based on vectorization and improving efficiency of the alignment of a single pair of sequences using some tricks, the efficacy of which is data-dependent. In particular these optimisations work poorly on short sequences, as well as on sets of sequences which are similar to one other. A GPU version of the algorithm, in contrast, performs a more efficient data base scan by performing thousands of individual scans in parallel. No data-dependent optimisations are employed, hence the performance of the algorithm is stable both for similar and short sequences.

Our algorithm scans several thousand sequences in parallel. The smallest chunk of sequences which are processed together comprises 256 sequences. These are processed on a single multiprocessor. The chips of current high-end GPUs contain at least 16 identical multiprocessors, therefore a GPU performs alignment of at least 4096 sequences in parallel.

Protein sequences are represented as a string of 20 characters corresponding to 20 types of amino acid. Homology between proteins is inferred from the similarity of their sequences - if sequence similarity is higher than expected by chance then the proteins are homologous. The alignments can contain gaps – these represent mutations, which correspond to deletions or insertions in the proteins. The score of the alignment is the sum of the similarity scores for amino acid pairs along alignment after subtracting the penalties for introducing gaps. The similarity score for a pair of amino acids is proportional to the logarithm of probability of mutation between these amino acids in proteins.

The Smith Waterman algorithm finds best local alignment by identifying all possible alignments. From among these it selects the optimal alignment. This is achieved using a dynamic programming approach. First a rectangular matrix is built, then each cell with index [i,j] is filled with a similarity score for the i-th residue of the query sequence and j-th residue of the target sequence. Then, starting from the upper left corner corresponding to the start of both sequences, one determines the best possible alignment ending in each cell. An appropriate score is then assigned to the cell. To process the cell one needs to know information on score and auxiliary variable in three adjacent cells - upper, left and upper-left.

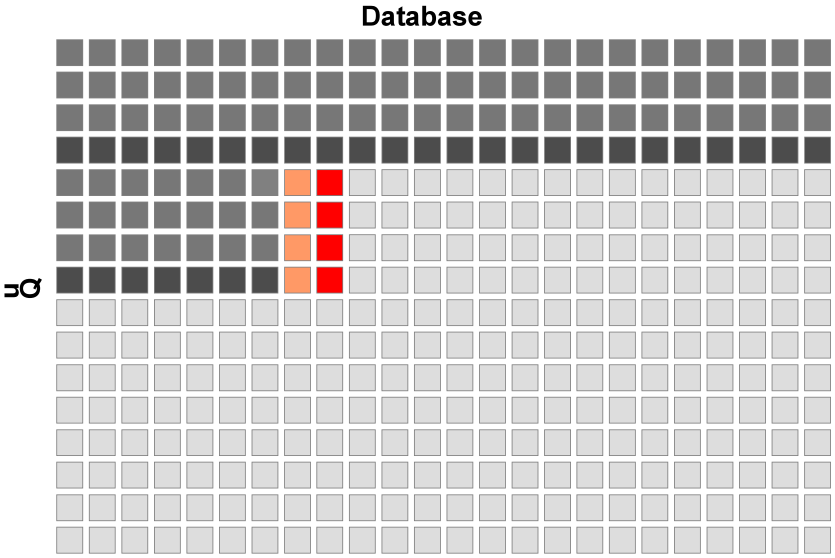


Figure 1: Processing of the dynamic programming matrix in horizontal bands. Cells, which are already processed, active cells and cells waiting for processing are displayed in dark gray, red and light gray, respectively. Data in orange cells to the left of active cells is stored in shared memory. Cells containing data accessed on beginning of the loop are highlighted in deep dark gray.

In our algorithm the SW matrix is processed in 12-cell high bands, as shown in Figure 1. Each band is processed column by column. The global memory access is only required twice, once at the entrance to the main loop and once at the exit. The score and auxiliary variable from the preceding column are stored in the shared memory. The algorithm achieves an approximately 100-fold increase in speed compared withthe single core of CPU, and its overall efficiency is on par with BLAST - the heuristic algorithm executed on CPU.

The sw-gpu algorithm is used as an engine of our similarity search server accessible at http://bioinfo.icm.edu.pl/services/sw-gpu/

We are currently working on using SW algorithm in an iterative way, analogous to the PSI-BLAST extension of the BLAST algorithm. In this approach the results of the search are used to develop an improved position-specific similarity function, which is then used for a refined search. The procedure is repeated several times, until no new sequences are found. Another application field is assembling of the DNA/RNA sequences from short fragments which are generated in the new sequencing methods (next generation sequencing).

Link:
http://bioinfo.icm.edu.pl/ services/sw-gpu/

Please contact:

Witold R. Rudnicki,
University of Warsaw, Poland
Tel: +48 22 5540 817
E-mail: W.Rudnicki@icm.edu.pl

Analyzing the Whole Transcriptome by RNA-Seq Data: The Tip of the Iceberg

by Claudia Angelini, Alfredo Ciccodicola, Valerio Costa and Italia De Feis

The recent introduction of Next-Generation Sequencing (NGS) platforms, able to simultaneously sequence hundreds of thousands of DNA fragments, has dramatically changed the landscape of genetics and genomic studies. In particular, RNA-Seq data provides interesting challenges both from the laboratory and the computational perspectives.

Gene transcription represents a key step in the biology of living organisms. Several recent studies have shown that, at least in eukaryotes, virtually the entire length of non-repeat regions of a genome is transcribed. The discovery of the pervasive nature of eukaryotic transcription and its unexpected level of complexity - particularly in humans – is helping to shed new light on the molecular mechanisms underlying inherited disorders, both mendelian and multifactorial, in humans.

Prior to 2004, hybridization and tag-based technologies, such as microarray and Serial/Cap Analysis of Gene Expression, offered researchers intriguing insights into human genetics. Microarray techniques, however, suffered from background cross-hybridization issues and a narrow detection range, whilst tag-based approaches required laborious time- and cost-effective steps for the cloning of fragments prior to sequencing. Hence, the recent introduction of massively parallel sequencing on NGS platforms has completely revolutionized molecular biology.

RNA-Seq is probably one of the most complex of the various “Seq” protocols developed so far. Quantifying gene expression levels within a sample or detecting differential expressions among samples, with the possibility of simultaneously analysing alternative splicing, allele-specific expression, RNA editing, fusion transcripts and expressed single nucleotide polymorphisms, is crucial in order to study human disease-related traits.

To handle this novel sequencing technology, molecular biology expertise must be combined with a strong multidisciplinary background. In addition, since the output of an RNA-Seq experiment consists of a huge number of short sequence reads - up to one billion per sequencing run - together with their base-call quality values, terabytes of storage and at least a cluster of computers are required to manage the computational bottleneck.

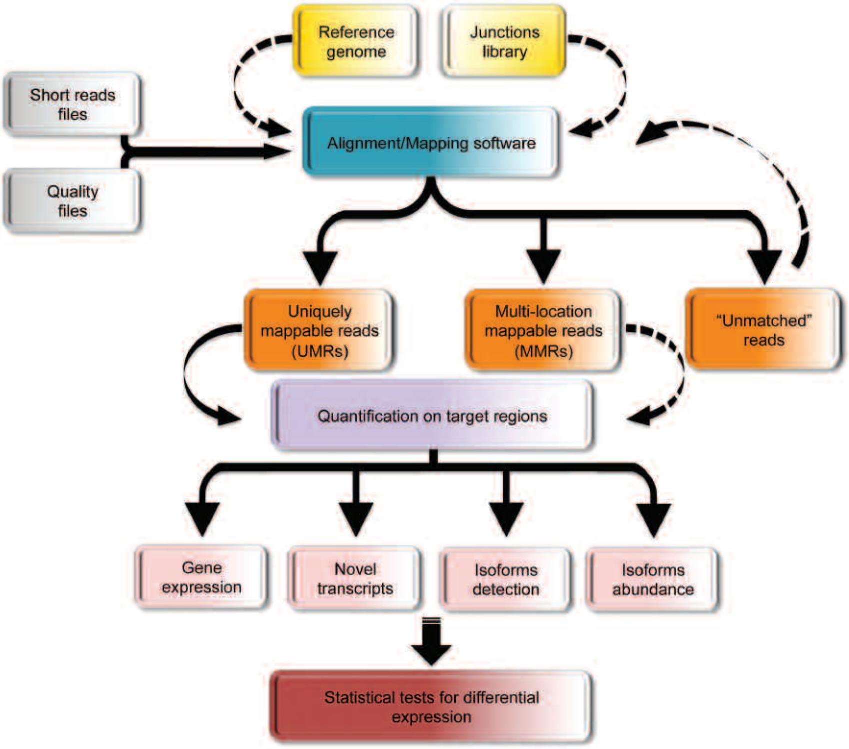

Reference genome
Junctions library
Short reads files
Quality files
Alignment/Mapping software
Uniquely mappable reads (UMRs)
Multi-location mappable reads (MMRs)
“Unmatched” reads
Quantification on target regions
Gene expression
Novel transcripts
Isoforms detection
Isoforms abundance
Statistical tests for differential expression

Figure 1: An example of the computational pipeline for the analysis of RNA-Seq data.

Recently, the Institute of Genetics and Biophysics (IGB) and the Istituto per le Applicazioni del Calcolo (IAC), both located in the Biotechnological Campus of the Italian National Research Council in Naples, have started a close collaboration on RNA-Seq data which aims to fill the gap between data acquisition and statistical analysis. In 2009 IGB, a former participant in the Human Genome project, acquired the SOLiD system 3.0, one of the first and most innovative platforms for massively parallel sequencing installed in Italy. IAC has great experience in developing statistical and computational methods in bioinformatics and is equipped with two powerful clusters of workstations capable of handling massive computational tasks.

The collaboration started with two pilot whole-transcriptome studies on human cells via massively parallel sequencing aimed at providing a better molecular picture of the biological system under study and at setting up an efficient computational open-source pipeline to downstream data analysis.

The computational effort focuses on the use of efficient software, the implementation of novel algorithms and the development of innovative statistical techniques for the following tasks:

a) alignment of the short reads to the corresponding reference genome
b) quantifying gene expressions
c) procedures of normalization to compare samples obtained in

different runs and assessment of the quality of the experiments

d) identification of novel transcribed regions and refinement of previously annotated ones

e) identification of alternative spliced isoforms and assessment of their abundance;

f) detection of differential genes/isoform expression under two or more experimental conditions

g) implementation of user-friendly interfaces for data visualization and analysis.

Each of these tasks requires the integration of currently available tools with the development of new methodologies and computational tools. Despite the unprecedented level of sensitivity and the large amount of data available to provide a better understanding of the human transcriptional landscape , the useful genetic information generated in a single experiment clearly represents only “the tip of the iceberg”. Much more research will be needed to complete the picture.

For steps a) and f), we are integrating some open source software into our pipeline. These are well-consolidated phases and there are several methods available in the literature. Points b) – e) are far more difficult as statistical methodologies are still lacking. The mathematical translation of the concept of “gene expression” and its modelling needs to be reassessed since we are now faced with discrete variables; we are now applying innovative methods.

The results we obtain from the computational analysis of the two pilot projects will be validated by quantitative real-time polymerase chain reaction (PCR) and, where deemed crucial for the analysis, the related protein products will be assessed by Western Blot. Biological validation will provide fundamental feed-back for optimizing the parameters of the computational analysis.

Please contact:
Claudia Angelini or Italia De Feis
IAC-CNR, Italy
E-mail: c.angelini| i.defeis@iac.cnr.it

Alfredo Ciccodicola or Valerio Costa
IGB-CNR, Italy
E-mail: ciccodic|costav@igb.cnr.it

Reliable Bacterial Genome Comparison Tools

by Eric Rivals, Alban Mancheron and Raluca Uricaru

Some bacterial species live within the human body and its immediate environment, and have various effects on humans, such as causing illness or assisting with digestion, while others colonize a range of other ecological niches. The availability of whole genome sequences for many bacteria opens the way to a better understanding of their metabolism and interactions, provided these genome sequences can be compared. In collaboration with biologists and mathematicians, a bioinformatics team at Montpellier Laboratory of Informatics, Robotics, and Microelectronics (LIRMM) has been developing novel tools to improve bacterial genome comparisons.

With the successful sequencing of the entire genome sequence of some species, the 1990s heralded a post-genomic era for biology: for the first time, the complete gene set of a species was amenable to - mostly in silico - investigations. Through comparative genomics approaches, seminal works have, for instance, determined the core gene set across nine bacteria, ie those genes indispensable to the life of any of these species.The availability of whole genome sequences has consequently opened up new research avenues. The current revolution of sequencing techniques provides such an improvement in yield that it brings within reach the sequencing of thousands of bacterial, or even longer, genomes within a few years. However, the exploitation of the information encoded in these sequences requires an increased capacity to compare whole genome sequences. Indeed, even finding the genes within the genome starts by comparing it to that of the evolutionary most closely related species, if available. Today, genome comparisons are necessary in order to survey the biodiversity of genotypes within populations, and to design vaccines .

Given that genomes are DNA sequences evolving under both point mutations, which alter a base at a time, and block rearrangements, ie duplication, translocation, inversion, or deletion of a DNA segment, comparing genomes is formulated as a string algorithm problem.

Two lines of approach have been followed: (i) extend gene sequence alignment algorithms by making them more efficient and coupling them with a procedure to deal with some types of rearrangements, (ii) consider only block rearrangements on a sequence of ordered markers (typically genes), but disregard the DNA sequence. Type two approaches adequately model changes in the gene order, but require well annotated genes and the knowledge of their evolutionary relationships across species, and are unable to report functionally relevant conservation or evolution in non-gene coding DNA regions. Note that handling multiple comparisons makes most of these formulations NP-hard, and, as with most evolutionary questions, benchmarks are missing since evolution occurs once and cannot be replayed. Thus, currently available solutions are time consuming, require fine parameter tuning, deliver results that are hard to visualise and interpret, and furthermore, do not fit the whole spectrum of applications.

In the framework of the project "Comparisons of Complete Genomes (CoCoGen)" supported by the French National Research Agency, we designed novel algorithms for comparing complete bacterial genomes.

Genome alignment algorithms first search for anchors, which are pairs of matching genomic regions, and chain them to maximize the coverage on both genomes. A chain is an ordered subset of non-overlapping collinear anchors,

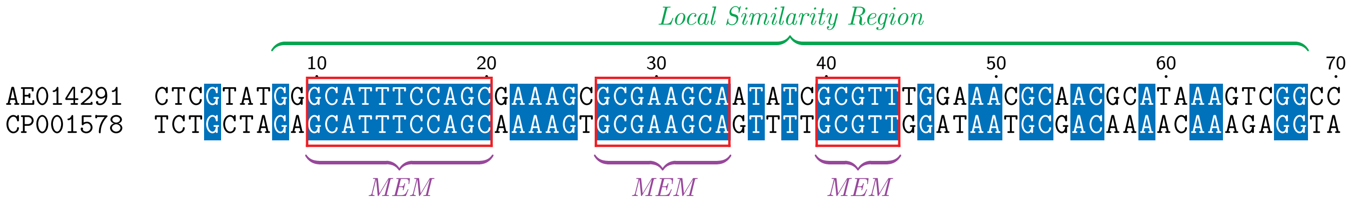
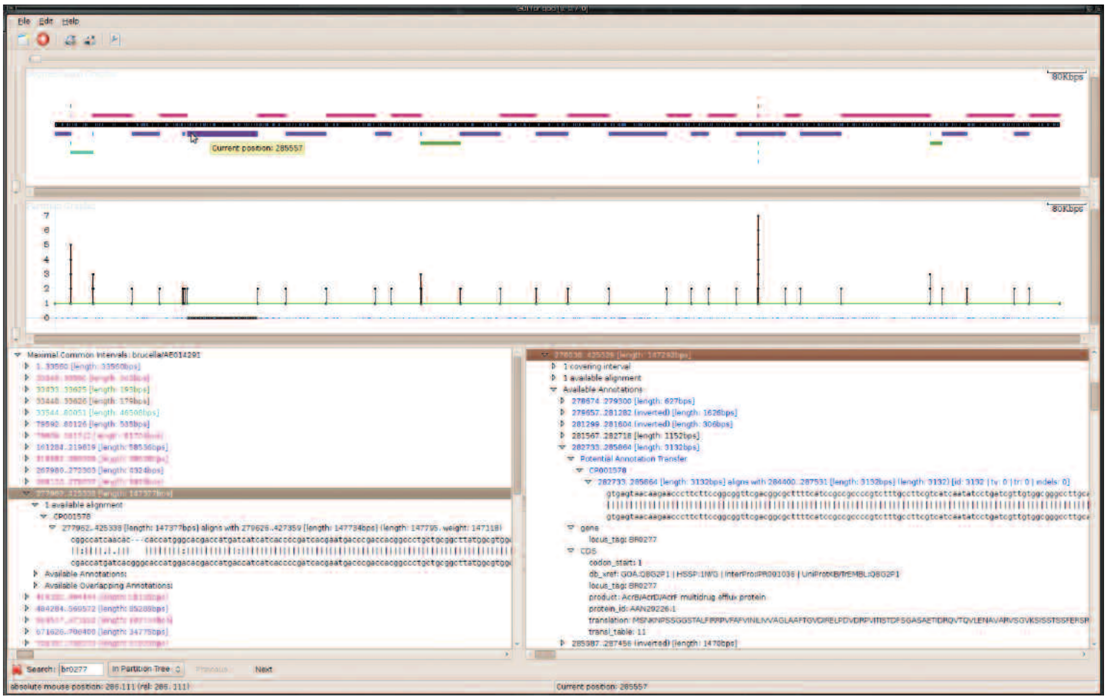

Local Similarity Region
10
20
30
40
50
60
70
AE014291 CTCGTATGGGCATTTCCAGCGAAAGCGCGAAGCAATATCGCGTTTGGAAACGCAACGCATAAAGTCGGCC
CP001578 TCTGCTAGAGCATTTCCAGCAAAAGTGCGAAGCAGTTTTGCGTTGGATAATGCGACAAAACAAAGAGGTA
MEM
MEM
MEM

Figure 1: Different anchors for whole genome alignment. Alignment of fragments of Brucella suis and B. microtii genomes showing on one hand Maximal Exact Matches (MEM) of length > 5 and a (much longer) Local Alignment. Identical bases are shown in blue background color.



```
▼ Maximal Common Intervals: brucella/AE014291
  ▶ 1..33560 [length: 33560bps]
  ▶ [illegible]
  ▶ 33433..33625 [length: 193bps]
  ▶ 33448..33626 [length: 179bps]
  ▶ 33544..80051 [length: 46508bps]
  ▶ 79592..80126 [length: 535bps]
  ▶ [illegible]
  ▶ 161284..219619 [length: 58536bps]
  ▶ [illegible]
  ▶ 267980..272303 [length: 4324bps]
  ▶ [illegible]
▼ 277962..425338 [length: 147377bps]
  ▼ 1 available alignment
    ▼ CP001578
      ▼ 277962..425338 [length: 147377bps] aligns with 279626..427359 [length: 147734bps] (length: 147795, weight: 147118)
          cggccatcaacac---caccatgggcacgaccatgatcatcatcaccccgatcacgaatgacccgaccacggccctgctgcggcttatggcgtgg
          ||:||||.|.|||   |||||||:|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
          cgaccatgatcacgggcaccatggacacgaccatgaccatcatcaccccgatcacgaatgacccgaccacggccctgctgcggcttatggcgtgg
  ▶ Available Annotations:
  ▶ Available Overlapping Annotations:
  ▶ [illegible]
  ▶ 484284..569572 [length: 85289bps]
  ▶ [illegible]
  ▶ 671626..706400 [length: 34775bps]
  ▶ [illegible]
Search: br0277   In Partition Tree
absolute mouse position: 286.111 (rel: 286. 111)
```

```
▼ 278036..425329 [length: 147292bps]
  ▶ 1 covering interval
  ▶ 1 available alignment
  ▼ Available Annotations
    ▶ 278674..279300 [length: 627bps]
    ▶ 279657..281282 (inverted) [length: 1626bps]
    ▶ 281299..281604 (inverted) [length: 306bps]
    ▶ 281567..282718 [length: 1152bps]
    ▼ 282733..285864 [length: 3132bps]
      ▼ Potential Annotation Transfer
        ▼ CP001578
          ▼ 282733..285864 [length: 3132bps] aligns with 284400..287531 [length: 3132bps] (length: 3132) (id: 3132 | tv: 0 | tr: 0 | indels: 0)
              gtgagtaacaagaacccttcttccggcggttcgacggcgcttttcatccgcgcccgtctttgccttcgtcatcaatatcctgatcgttgtggcgggccttgca
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
              gtgagtaacaagaacccttcttccggcggttcgacggcgcttttcatccgcgcccgtctttgccttcgtcatcaatatcctgatcgttgtggcgggccttgca
      ▼ gene
          locus_tag: BR0277
      ▼ CDS
          codon_start: 1
          db_xref: GOA:Q8G2P1 | HSSP:1IWG | InterPro:IPR001036 | UniProtKB/TrEMBL:Q8G2P1
          locus_tag: BR0277
          product: AcrB/AcrD/AcrF multidrug efflux protein
          protein_id: AAN29226.1
          translation: MSNKNPSSGGSTALFIRRPVFAFVINLIVVAGLAAFTGVDIRELPDVDRPVITISTDFSGASAETIDRQVTQVLENAVARVSGVKSISSTSSFERSR
          transl_table: 11
    ▶ 285987..287456 (inverted) [length: 1470bps]
Current position: 285557
```

Figure 2: Output of QOD, when used to compare pairwise Brucella suis and B. microtii genomes. The top diagram represents B. suis genome as a central horizontal black and blue bar, and each other colored horizontal bar depicts a maximal region of similarity with B. microtii genome. The high genome coverage by such bars indicates the proximity of the two species. Lower parts of the window allows to browse the segmentation together with intersecting annotations of the target genome.

meaning that the succession of regions appears in the same order on both genomes. Current tools use pairs of short exact or approximate matches as anchors (Mumer, Mauve, MGA), to be sure to find secure anchors in quite divergent regions. Instead, we focus on local alignments (LA) found using sensitive spaced seeds filtration. The length of local alignments adapts to the divergence level of compared regions, and moreover, LA are selected and ranked according to their statistical significance (see Figure 1). However, an important disadvantage is that the regions of distinct anchors may overlap within a genome. By adapting a Maximum Independent Set algorithm on trapezoid graphs, we exhibited a chaining algorithm that allows overlaps between anchors that depend proportionally on the anchor lengths. This improved significantly the coverage obtained with LA, and overcomes the overlap created by prevalent genomic structures like tandem repeats.

Unravelling the pathogenicity mechanisms of some bacterial strains requires knowledge of which genomic regions encode the corresponding genes compared to less virulent or non-pathogenic strains. These genes may be relevant in the process of diagnosis, prognosis, and treatment of disease. For instance, a recent approach to vaccine design relies strongly on this type of genome comparison. As many applications of comparative genomics aim at finding such distinguishing regions between strains or species, we conceived methods that can directly pinpoint such regions. In existing algorithms, genome comparison is formulated as a maximization problem and used heuristics tend to incomporate unreliably aligned region pairs in the final output. Consequently, distinguishing regions may be missed due to misalignment, since an anchor can be found by chance in or around such regions. We propose another genome segmentation algorithm that automatically partitions the target genome into regions that are shared or not shared with *k* reference genomes (see software QOD in Figure 2). Performing several comparisons with a distinct set of reference genomes allows us to rapidly determine which genomic regions distinguish a subset of pathogenic strains/species.

Application of our methods to the comparison of multiple bacterial strains that are pathogenic for ruminants has enabled discovery of the new genes that were overlooked by several previous genome annotation projects. It has also facilitated identification of a gene whose sequence varies among strains, and was known to code for a protein that generates important immunological reactions in related bacterial species (the human and dog pendant pathogens).

Link:
http://www.lirmm.fr/~rivals/CoCoGEN/

Please contact:

Eric Rivals
The Montpellier Laboratory of Informatics, Robotics, and Microelectronics (LIRMM), France
E-mail: rivals@lirmm.fr

ModeRNA builds RNA 3D Models from Template Structures

by Magdalena Musielak, Kristian Rother, Tomasz Puton and Janusz M. Bujnicki

Biological functions of many ribonucleic acid (RNA) molecules depend on their three-dimensional (3D) structure, which in turn is encoded in the RNA sequence. We have developed ModeRNA, a program that constructs 3D models of RNAs based on experimentally determined "template" structures of other, related RNAs. This approach is less time and cost intensive than experimental methods.

RNAs are linear polymers comprising tens to thousands of nucleotide residues that function as building blocks. There are four basic building blocks: adenosine, guanosine, cytidine and uridine, but they are often enzymatically modified in the cell to form more than one hundred derivatives with different chemical structures. RNA molecules and their complexes are main players in the production of proteins and other processes in cells. One prominent example is transfer RNA (tRNA), a molecule with a complex 3D structure, used to decipher the genetic code and translate the genetic information in messenger RNA (mRNA) into protein sequences. Determining 3D structures by experimental methods is time-consuming and expensive, compared to methods used for obtaining linear RNA sequences. Thus far only 150 tRNA structures have been solved experimentally, as opposed to more than 300,000 known nucleotide sequences.

Because all tRNAs are evolutionarily related, their structures are similar to one other. We can use experimental 3D data of one tRNA as a template to create a model of another related tRNA with a known, different sequence. In addition to the template structure, information about correspondencies of nucleotide residues between the target sequence and the template sequence (sequence alignment) is required. The alignment is interpreted as a set of instructions regarding which nucleotide residues of the template are to be replaced by which residues of the target. One can think of this concept as creating different pictures from a set of jigsaw puzzle pieces by editing an existing puzzle – for instance replacing a zebra by a lion or adding a mountain in the background in a savannah landscape.

This technique, called homology modelling or comparative modelling, has been implemented in the ModeRNA program. ModeRNA builds RNA structures starting with the easiest part: nucleotides identical between the target and the template are placed in exactly the same position. Then, single nucleotide substitutions are introduced, for example an adenine can be exchanged for a guanine. Finally, parts of the RNA structure that are not present in the template are modelled. For that purpose, ModeRNA uses a library of over 100,000 structural fragments, from which the program chooses the one that geometrically fits best into the insertion site in the model. Because even a well-suited fragment may cause small distortions of bond lengths and angles, ModeRNA can optimize atom coordinates to achieve a stereochemically reasonable conformation. Referring to the puzzle analogy, this would be inserting a set of coherent pieces and applying a rasp to make their edges fit to the remainder of the puzzle.
To manipulate 3D coordinates of atoms, ModeRNA employs the Kabsch algorithm for superposition, the Full Cyclic Coordinate Descent algorithm for closing gaps, and the NeRF algorithm to construct atom coordinates. To identify nucleotides, a subgraph matching procedure has been implemented. The program also contains a multitude of functions, eg, to analyse the geometry of existing structures, to find interatomic clashes, or simply remove unwanted nucleotide residues. These functions are available via a scripting interface. ModeRNA has been written in the Python language, using the BioPython library for basic tasks like parsing structural data from the PDB format. ModeRNA is available under the GPL Open Source license. The program is being used by several lab members and collaborators, who provide constant feedback and suggestions for improvement.

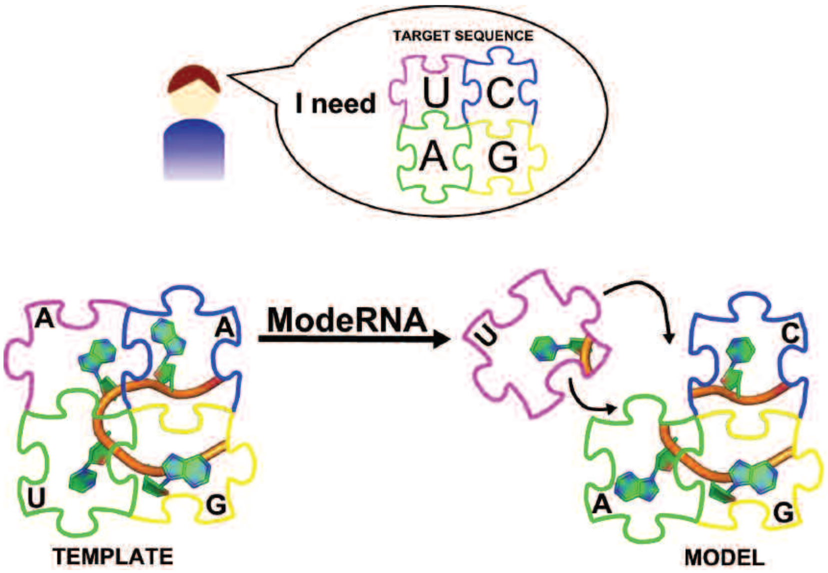


Figure 1: ModeRNA builds a 3D model of an RNA molecule based on a template structure and an alignment of two sequences.

To test ModeRNA, we constructed a series of 9801 tRNA models for 100 structures determined by X-ray crystallography. We calculated the root mean square deviation (RMSD) of the atomic coordinates of modelled versus experimental structures. The results showed that the RMSD between the models and the original structures correlates well with the RMSD among experimentally solved structures. The best models reached an RMSD up to 1 Å, with a majority around 4-5 Å. Obviously, the quality of the model depends on how similar the template is to the target, which highlights the importance of the choice of the right template. However, it must be remembered that the RNA structure can change depending on the functional state. For example the anticodon loop and acceptor stem regions of tRNA can adopt different conformations depending on whether the molecule is bound to a protein, or to the ribosome, or if it is free from interactions. We can model these subtle differences by choosing a template structure that is in the desired state.

As the template has such an important influence on model quality, a striking question is: "Do we really need a template, or could we just connect nucleotides from scratch?" In fact, methods like the MC-Fold/MC-Sym pipeline have been used successfully to model small RNA structures (12-50 nucleotides) from nothing more than a nucleotide sequence. However, when the sequence is longer, building a structure without further knowledge becomes computationally unfeasible. For many larger RNA families known 3D structures are available, among them tRNA having around 75 nucleotides, and ribosomes with more than 1000 nucleotides. Thus comparative modeling is a method of choice for 3D structure prediction of large structured RNAs.

In summary, ModeRNA is a tool for construction of 3D models for RNA sequences using structures of another, related RNA as building blocks (templates). ModeRNA also facilitates RNA 3D structure analysis.

Link:
http://iimcb.genesilico.pl/moderna/

Please contact:

Janusz M. Bujnicki
International Institute of Molecular and Cell Biology, Poland
Tel: +48 22 597 07 50
E-mail: iamb@genesilico.pl

Protein Homology Modelling - Providing Three-dimensional Models for Proteins where Experimental Data is Missing

by Jürgen Haas and Torsten Schwede

A linear sequence of amino acid letters or a three-dimensional arrangement of atoms in a polypeptide chain? Most biologists and bioinformaticians will have their preferred view when imagining a "protein". Although these views represent two sides of the same coin, they are often difficult to reconcile. There are two main reasons for this: a lack of experimental structural information for the majority of proteins, and a different "culture" in handling data between the two communities, which results in a number of technical hurdles for somebody trying to bridge the gap between the two paradigms. Protein homology modelling resources establish a natural interface between sequence-based and structure-based approaches within the life science research community.

Natural proteins are exciting molecules which - in contrast to the regular helical DNA molecule - have characteristic, well-defined three-dimensional structures. It is the individual structure of a protein, with its intricate network of atomic interactions formed by the backbone and side chains atoms of the amino acids in a polypeptide chain, which allow it to perform highly specific functions in a living organism. These include mechanical force generated by motor proteins, enzymes degrading nutrients, antibodies recognizing foreign substances such as a pathogen, or olfactory receptors sensing the smell of perfume in one's nose. But the most exciting property of these molecular machines is that they don't need specific tools to be assembled: The linear sequence of amino acids alone contains sufficient information to define the three-dimensional structure when a protein is synthesized by a cell.

Mind the gap

Crucial to the understanding of molecular functions of proteins are insights gained from their three-dimensional structures. However, while thousands of new DNA sequences coding for proteins are sequenced every day, experimental elucidation of a protein structure is still an expensive and laborious process, typically taking several weeks. Not surprisingly, direct experimental structural information is, to date, only available for a small fraction of all proteins, and this gap is widening rapidly.

Whilst theoretically a nearly unlimited number of amino acid sequence combinations are possible, the number of different three-dimensional protein structures ("folds") actually observed in nature is limited. This allows for homology (or comparative) modelling methods to build computational models for proteins ("targets") based on evolutionary related proteins for which experimental structures are known ("templates"). Hence experimental structure determination efforts and homology modelling complement each other in the exploration of the protein structure universe.

The SWISS-MODEL Expert System

For an experimental scientist, being interested in obtaining a three-dimensional structural model of a protein to study a biological question should not require becoming an expert in molecular modelling or programming the necessary software tools. The aim of the SWISS-MODEL expert system is to provide a user-friendly Web-based system for protein structure homology modelling which is usable from any PC equipped with an Internet connection – without the need for installing complex software packages or huge databases. Sixteen years ago, SWISS-MODEL was the first fully automated protein modelling service on the Internet, aiming to make protein structure modelling easily accessible to life science research. Today SWISS-MODEL is one of the most widely used Web-based modelling services world wide. SWISS-MODEL hides the complexity of a stack of specialized modelling software, mirrors public databases of protein sequences and structures, automates procedures for data updates, and has tools for result visualization. Users interact with a set of partially or fully automated workflows in a personalized Web-based workspace. SWISS-MODEL is developed by the Computational Structural Biology Group at the Swiss Institute of Bioinformatics (SIB) and the Biozentrum of the University of Basel, Switzerland.

The Protein Model Portal

Diversity is essential to the success of any biological entity. However, there is reason to doubt that the same is true for the technical diversity observed among typical bioinformatics resources, especially in the field of molecular modelling. A simple question like “What is known about the three-dimensional structure of a given protein?” can easily end in an hour-long odyssey through the Internet for proteins which are not fully experimentally characterized. While all experimental data is collected centrally in the wwPDB – currently approximately 60,000 experimental structures are deposited in this database - computational structure models are generated by many specialized computational modelling groups spread out in the academic community. The heterogeneous user interfaces and formats used at different sites using various incompatible accession code systems are an additional challenge in accessing structure model information.

In order to provide a single interface to access both experimental structures and computational models for a protein we have thus developed the Protein Model Portal (PMP) by federating the available distributed resources. PMP is part of the Nature – PSI Structural Biology Knowledgebase (PSI-SBKB) project, and is developed by the Computational Structural Biology Group at the Swiss Institute of Bioinformatics (SIB) and the Biozentrum of the University of Basel, Switzerland in collaboration with the PSI SGKB partner sites, specifically Rutgers, The State University of New Jersey. The PSI-SGKB informs about advances in structural biology and structural genomics in an integral manner, including not only newly determined three-dimensional structures, but also the latest protocols, novel materials and technologies. The current release of the portal (May 2010) consists of 12.7 million model structures provided by different partner resources for 3.4 million distinct protein sequences. PMP is available at http://www.proteinmodelportal.org and from the PSI Structural Biology Knowledgebase (http://kb.psi-structuralgenomics.org).

In order to integrate the meta-information on protein models available at the different partner sites – using heterogeneous data structures and incompatible naming conventions and accession code systems – we created a common independent reference system based on cryptographic md5 hashes for all currently known, naturally occurring amino acid sequences. This database is continuously updated while new sequences are being discovered, and allows us to dynamically federate information from different providers, without the need for the distributed resources to change their mode of operation. Using portal technologies, queries are transparently mapped to various database accession code systems, providing dynamic functional annotation for the target sequences.

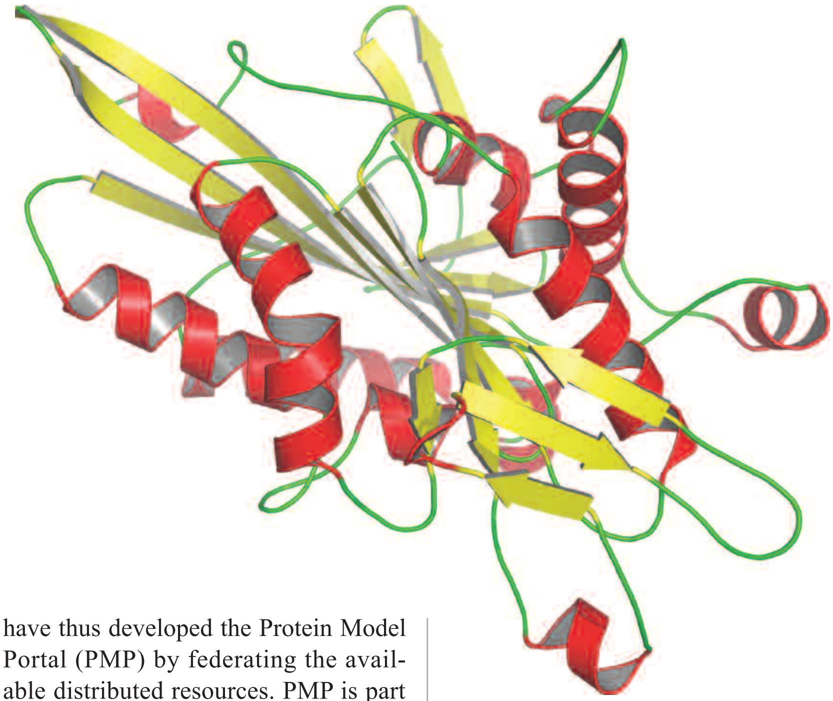


Figure 1: Three-dimensional homology model of the human Kinesin-like protein KIF3A, a protein involved in the microtubule-based translocation. The model was generated using the experimental structure of a related protein, the motor domain of the human Kinesin-like protein KIF3B sharing 66% sequence identity, as template.

For the first time it is now possible to query all participating federated structure resources - both experimental and computational models - simultaneously and compare the available structural information in a single interface. The dynamic mapping of the growing universe of all protein sequences even allows searching for features which were not implemented in the original data sources. Sometimes, the whole is greater than the sum of its parts.

Links:
http://www.proteinmodelportal.org
http://kb.psi-structuralgenomics.org
http://www.wwpdb.org

Please contact:

Jürgen Haas and Torsten Schwede
Biozentrum University of Basel
SIB Swiss Institute of Bioinformatics
Tel: +41 61 2671581
E-mail: juergen.haas@unibas.ch,
torsten.schwede@unibas.ch

Modelling of Rapid and Slow Transmission Using the Theory of Reaction Kinetic Networks

by Dávid Csercsik, Gábor Szederkényi and Katalin M. Hangos

With their interdisciplinary background and interest in nonlinear process systems, the Process Control Research Group at Computer and Information Research Institute Hungarian Academy of Sciences carries out research in modelling, analysis, representations and control of reaction kinetic systems, and their application in systems biology.

The modelling and analysis of signalling pathways in living organisms is one of the most challenging problems in systems biology that can be handled using the theory of chemical reaction networks (CRNs) obeying the mass-action law. In addition to being used to describe pure chemical reactions, such networks are also widely used to model the dynamics of intracellular processes, metabolic or cell signalling pathways. This model class enables the use of the deficiency-based, multi-stability-related results of Martin Feinberg et al., which provide very strong theorems about qualitative behaviour of the modelled system, based only on the structure of the reaction network, independently of its parameters.

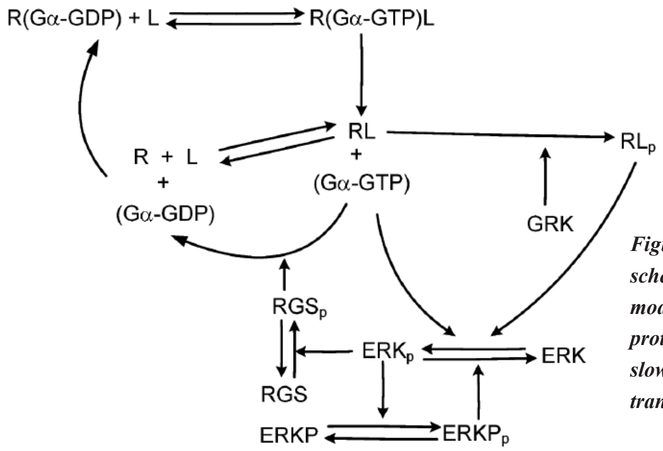


Figure 1: The reaction scheme of the kinetic model describing fast (G protein coupled) and slow (β-arrestin coupled) transmission).

In the case of reaction kinetic models, we consider a system of n chemical species participating in an r reversible steps reaction network. For graphical representation of the kinetic system, reaction schemes can be used, which describe the structure of the enzymatic and non-enzymatic reactions in a compressed way (not depicting every single reaction). Reaction schemes can be depicted in mathematical terms using hypergraphs, where the edges may be adjacent to more than two vertices. The vertices of a reaction scheme correspond to the non enzymatic complex type components, while the hyper-edges describe chemical reactions (not necessarily reaction steps). An enzyme-catalytic reaction corresponds to a pair of hyper-edges with different directions, both adjacent to three components S, P and E (substrate, product and enzyme, respectively). An example of a reaction scheme is shown in Figure 1.

One recent project in which CNRs have been applied to biological systems is the modelling of rapid (G protein coupled) and slow (β-arrestin coupled) transmission. Until the 2000s the most widely accepted classic paradigm of signalling, related to G protein coupled receptors, hypothesized that the most important elements which contribute to informa-

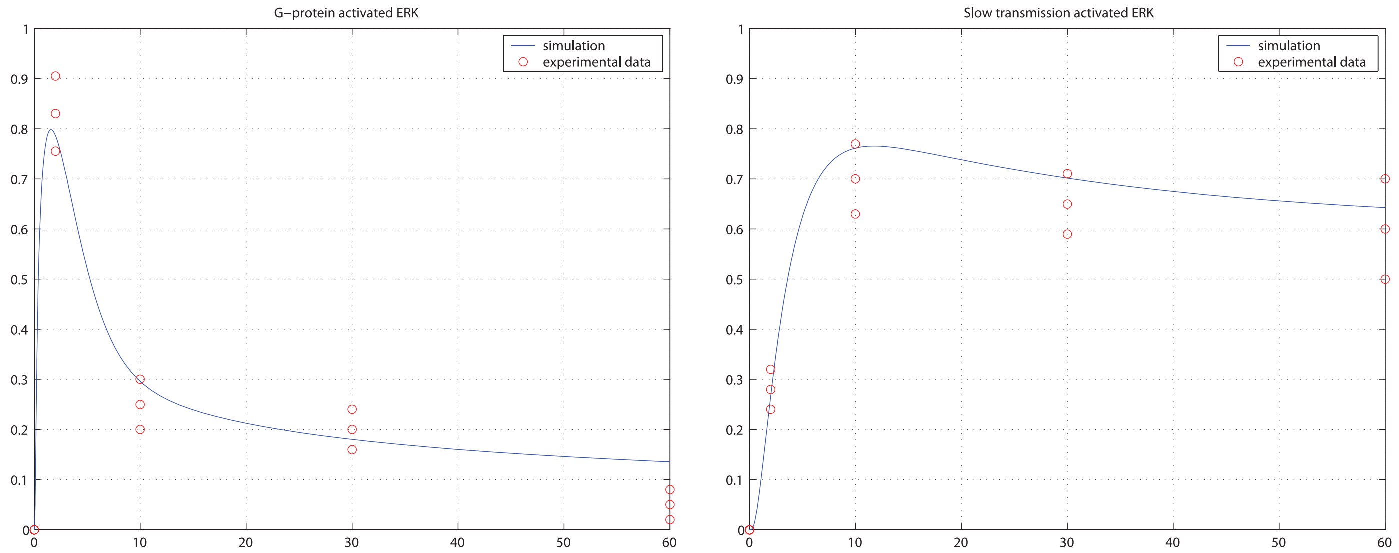


Figure 2: The simulation results of the model describing fast (G protein coupled) and slow (β-arrestin coupled) transmission.

tion transfer to the internal system of the cell are the α and βγ subunits of G proteins. In recent years, it has been shown that, in addition to taking part in receptor desensitization and attenuation of G protein coupled signalling, β-Arrestins also form an endocyctic protein complex. This complex initiates a G protein independent transmission and regulation of extracellular regulated kinase (ERK), an important kinase thatplays a central role in the intracellular signalling network (ERK is also activated by G protein coupled pathways). The recognition that a single receptor acts as a multiple source of signalling pathways, and various drugs binding to this receptor, might differentially influence each pathway (in contrast to pathway-specific drugs), led to the reassessment of the efficacy concept. These recent biological findings led to the concept of a dynamical model, capable of describing the interactions of the two convergent, but qualitatively different signalling mechanisms.

In cooperation with the Neuromorphological and Neuroendocrine Research Laboratory of the Department of Human Morphology and Developmental Biology (Hungarian Academy of Sciences and Semmelweis University), we have developed a simplified dynamic model that describes the dynamic behaviour of G protein signalling. The model takes into account the effect of slow transmission, RGS mediated feedback regulation and ERK-phosphatase mediated feedback regulation. The parameters of the model have been determined via numerical optimization.

The proposed reaction kinetic model, depicted in Figure 1 as a reaction scheme, gives rise to an acceptable qualitative approximation of the G protein dependent and independent ERK activation dynamics that is in good agreement with the experimentally observed behaviour (see Figure 2).

The developed and validated model could potentially be applied to disorders of the reproductive neuroendocrine system. In cases of polycystic ovary syndrome for instance, treatment may include administration of the key hormone GnRH (which acts via G protein coupled receptors) or its analogues. The importance of slow transmission is also becoming evident in other areas of physiology and medicine, for example recent studies suggest that ?-arrestins may play a central role in diabetes mellitus and insulin resistance.

Link:
http://daedalus.scl.sztaki.hu/PCRG/

Please contact:

Dávid Csercsik
SZTAKI, Hungary
Tel: +36 1 279 6163
E-mail: csercsik@scl.sztaki.hu

Gábor Szederkényi
SZTAKI, Hungary
Tel: +36 1 279 6163
E-mail: szeder@scl.sztaki.hu

Katalin M. Hangos
SZTAKI, Hungary
Tel: +36 1 279 6101
E-mail: hangos@scl.sztaki.hu

SCAI-VHTS - A Fully Automated Virtual High Throughput Screening Framework

by Mohammad Shahid, Torbjoern Klatt, Hassan Rasheed, Oliver Wäldrich and Wolfgang Ziegler

One of the major challenges of in-silico virtual screening pipelines is dealing with increasing complexity of large scale data management along with efficiently fulfilling the high throughput computing demands. Despite the new workflow tools and technologies available in the area of computational chemistry, efforts in effective data management and efficient post-processing strategies are still ongoing. SCAI-VHTS fully automates virtual screening tasks on distributed computing resources to achieve maximum efficiency and to reduce the complexities involved in pre- and post-processing of large volumes of virtual screening data.

Virtual screening is an important and complementary step in the modern drug discovery process. Small molecules in virtual compound databases are computationally screened against specific biological protein targets by computing the binding energy of these molecules inside the protein active site. Scoring and ranking is performed to filter and select drug-like molecules, which are active against the biological targets of interest. The three dimensional structures of both the protein targets and small molecules are required to perform virtual screening in the structure based drug discovery process. There are more than 60,000 protein 3D structures available in the Protein Data Bank (PDB) and millions of small molecules in compound databases, which are publicly available. Furthermore, there are billions of virtual compounds that could be obtained from combinatorial chemistry space. Even simulating a few million of such large datasets increases the demand for computing resources, as well as the effort involved with management of large datasets.

Without a framework that fully automates the virtual screening workflow, manual execution and management of the workflow is very cumbersome. First, there is the subtle issue of handling huge amounts of data in the form of large numbers of input/output files. Data management during pre- and post-processing stages is the most tedious, laborious and time-consuming work. Other important issues include the manual distribution of the workload on the available computing resources, and tracking and monitoring of the running tasks. Furthermore, fault tolerance is an important issue that requires great effort in failure management, including identification and resubmission of tasks that are failed or lost for various reasons.

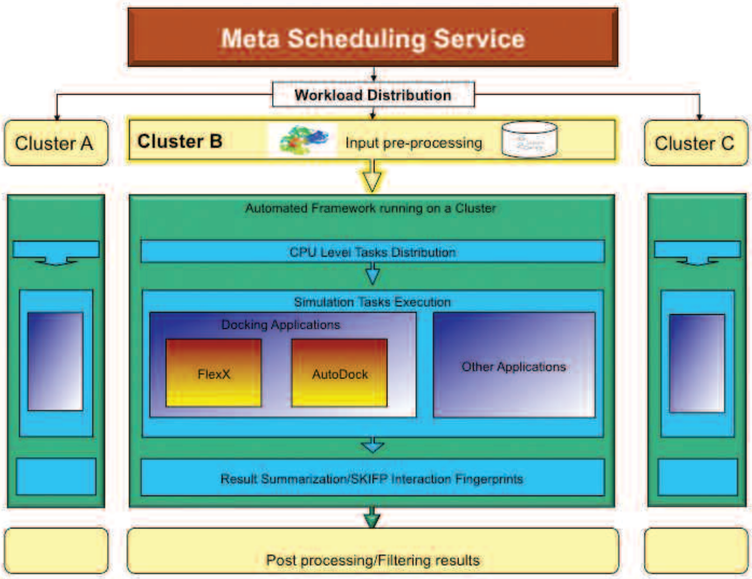

Meta Scheduling Service
Workload Distribution
Cluster A
Cluster B
Input pre-processing
Cluster C
Automated Framework running on a Cluster
CPU Level Tasks Distribution
Simulation Tasks Execution
Docking Applications
FlexX
AutoDock
Other Applications
Result Summarization/SKIFP Interaction Fingerprints
Post processing/Filtering results

Figure 1: Architecture of the automated virtual screening framework.

Many recent research projects have been demonstrating the role and relevance of using Grid technologies for virtual screening. Grid technologies can help accelerate the screening process as well as providing required resources such as computing power and large scale data storage that meet the demands of CPU- and data-intensive biomedical applications. In research environments Grids are established as a useful technology that allows sharing and on demand utilisation of geographically dispersed heterogeneous resources including high performance computing resources. Grid technologies are applied to perform high throughput computation in the computational chemistry domain. The use of this technology enables researchers to collaborate in a virtual laboratory that provides an integrated resource platform. A framework for integrating tools and methodologies for efficient deployment of virtual screening experiments forms the basis of such a virtual screening laboratory.

SCAI-VHTS, a fully automated virtual high throughput screening framework, enables the researchers to perform large scale virtual screening experiments that include complex workflows with great ease of use. The framework is based on the functionalities provided by the UNICORE Grid middleware, together with SCAI's Meta Scheduling Service extended in the PHOSPHORUS project (see links below). A virtual screening workflow typically consists of pre-processing the input data, the distribution of the data to the computing systems, execution of the screening applications that perform simulations, post-processing and collection of the results. The architecture of the automated virtual screening framework (see Figure 1) provides a single point of interaction to distributed computing resources while hiding the complexity of the underlying infrastructure.

Within this framework two compute and data intensive applications FlexX and AutoDock have been integrated with the components for performing pre-processing of the input data as well post-processing of the results. Pre-processing includes formatting the input data for the respective application while post-processing includes filtering of virtual screening results by applying post-docking strategies. Post-processing is mainly comprised of ranking by the built-in scoring function of the application as well as re-ranking by a modified protein-ligand interaction fingerprints approach.

The researcher simply has to submit a single virtual screening job on the Grid through a graphical UNICORE client plug-in, after which the Meta Scheduling Service performs work load distribution, manages execution and monitoring on the computing resources available. Wrapper programs configured locally on the compute clusters further facilitate maximum utilisation of the compute resources as well as the management of the distributed virtual screening data. The modular workflow is fully extensible to allow easy integration of other applications such as molecular dynamics simulations.

The SCAI-VHTS framework relieves the researcher from dealing with the challenges of deploying large scale virtual screening experiments and the tedious and laborious work involved in management of large volumes of input/output data. As a result, the researcher can concentrate on his primary goals: finding and evaluating new drugs. SCAI-VHTS can assist with data management, analysis, and automated knowledge discovery by facilitating remote collaborations of distributed and heterogeneous Grid resources, thus creating a virtual laboratory for biological research and development.

Links:
UNICORE:
http://www.unicore.eu
PHOSPHORUS:
http://www.ist-phosphorus.eu/

Please contact:
Mohammad Shahid,
Fraunhofer SCAI, Germany,
Tel: +49 2241 14 2777
E-mail: shahid@scai.fraunhofer.de

Wolfgang Ziegler,
Fraunhofer SCAI, Germany,
Tel: +49 2241 14 2248
E-mail:
Wolfgang.Ziegler@scai.fraunhofer.de

Searching for Anti-Amyloid Drugs with the Help of Citizens: the “AMILOIDE” Project on the IBERCIVIS Platform

by Carlos J. V. Simões, Alejandro Rivero, Rui M. M. Brito

The current “target-rich and lead-poor” scenario in drug discovery, together with the massive financial resources required to develop a new drug, mean that new approaches and renewed efforts by researchers in academia are required, in particular in the area of neglected and rare diseases. Virtual screening and volunteer computing are extremely useful tools in this fight, and together have the potential to play a crucial role in the early stages of drug development. From a social and economic point of view, amyloid neurodegenerative diseases, including Alzheimer´s, Parkinson´s, familial amyloid polyneuropathy and several others, currently represent important targets for drug discovery. Here, we provide a short account of our current efforts to develop new compounds with anti-amyloid potential using a volunteer computing network.

Transthyretin Amyloid

Despite recent advances in our understanding of biological systems, the discovery and development of a new therapeutic agent is still a slow, difficult and very expensive process. This is mainly due to the low success rates associated with finding the “magic bullet” for a given biological target. Computational methodologies play an increasing role in drug discovery but, due to limitations of the approximations used, it is clear that a “one recipe fits all” approach is far from suitable.

At the Structural and Computational Biology group of the Center for Neuroscience and Cell Biology, University of Coimbra, we devised a research program to find new lead compounds for the treatment of amyloid diseases, in particular, familial amyloid polyneuropathy (FAP), a genetic neurodegenerative disease that is prevalent in Portugal and with other *foci* around the world. Known as a highly impairing disease, it is characterized by loss of sensation to temperature and pain in the lower limbs during its early stages, later evolving to muscular weakness and general autonomic dysfunction. The only effective treatment known to date is liver transplantation, since the liver is the main site of synthesis of the protein transthyretin, the causative agent of FAP.

Transthyretin (TTR) is a protein of the blood plasma known to bind and transport the hormone thyroxine. The cytotoxic role of TTR on the peripheral nerves involves the formation of amyloid aggregates and amyloid fibrils through a series of molecular events, whereby the protein dissociates, partially unfolds and aggregates. These events can be modulated by the binding of thyroxine-like molecules to the two binding pockets of TTR. Several of these small molecules have been shown to stabilize the protein and thereby decrease or even prevent amyloid formation, but in general they are associated with undesirable side effects, low solubility, and/or inability to strongly bind TTR in the plasma. However, a Phase III clinical trial has recently been concluded with one of these molecules, with very promising results.

Virtual Screening and Volunteer Computing: the AMILOIDE-Ibercivis project

In order to search for new lead compounds capable of stabilizing the native form of TTR, we envisaged a rigorous computationally-driven search to minimize the occurrence of false positives. Multiple combinations of docking algorithms and scoring functions were validated by assessing the software’s ability to reproduce the binding mode of known TTR binders in experimental X-ray structures. Docking algorithms such as AutoDock4, FRED, eHiTS and GOLD were tested, along with their inbuilt and additional scoring functions. For TTR, the combination of AutoDock4 with a scoring function called DrugScore-CSD produced the most reliable results.

In parallel, restricting “chemical space” to a biologically relevant and synthetically accessible set of molecules is an essential step in a Virtual Screening (VS) campaign. We have generated a tailored library of 2,259,573 compounds by filtering an initial collection of approximately 11 million molecules deposited in the ZINC database. The filtering process involved combining rules for bioavailability with our knowledge of the physico-chemical properties of the known TTR stabilizers. However, the docking of approximately 2.3 million compounds to TTR on a single desktop computer would take about 43 years of CPU time.

The AMILOIDE project takes advantage of the expansion of the Ibercivis network to Portugal, implementing AutoDock4 on a large-scale volunteer computing platform. Powered by the BOINC middleware, Ibercivis was initially developed at the Institute for Biocomputation and Physics of Complex Systems, University of Zaragoza, and it is guided by two main objectives: i) to provide an easily operable distributed computing platform available to multiple scientific projects; ii) to provide a science communication forum where citizens and scientists may interact.

In AMILOIDE-Ibercivis, each BOINC Work Unit is comprised of one TTR receptor structure with its respective 3D maps of atomic affinity, one ligand structure, and docking parameter files. The Work Units are sent to the network clients where AutoDock4 carries out the calculations. When these are concluded, the results are sent back to one of the Ibercivis servers and later transferred to the researchers’ machines for analysis. Here, the predicted complexes of TTR are submitted to rescoring with DrugScore-CSD. The most promising molecules are pur-

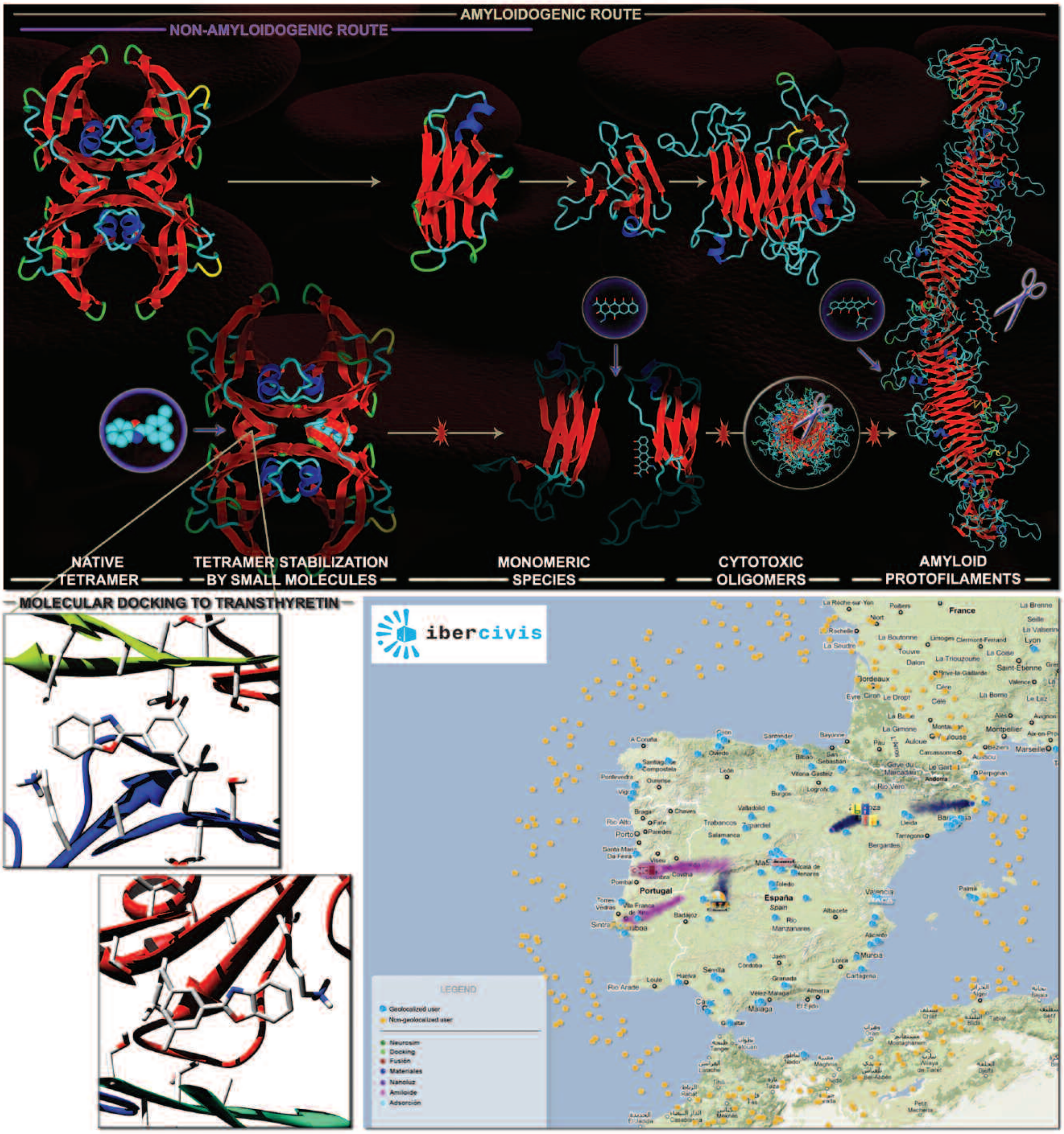


Figure : Physiopathological model for amyloid fibril formation by Transthyretin (TTR) (top panel, top row) and possible points of therapeutic intervention (top panel, bottom row): stabilization of the native tetramer; inhibition of aggregation; solubilization of aggregates and fibrils. Stabilization of the native tetramer is currently the strategy of choice, due to the large amount of experimental data available. Virtual high-throughput docking of small organic molecules to Transthyretin (bottom left panel) is being conducted on the large volunteer computing network Ibercivis (bottom right panel).

chased and tested experimentally for activity.

During the period between October 2009, when the first production Work Unit of the AMILOIDE-Ibercivis project was launched, and May 2010, all 2,259,573 Work Units were sent out to BOINC clients on Ibercivis. Of these, approximately 80% have been successfully returned. This represents an average validation rate of 8,600 Work Units per day, or 258,000 Work Units per month. Client errors are mostly due to aborted jobs (74%), followed by computation errors (23%). Download errors account for only 3% of the failures. The computing performance of this experiment is staggering: the Ibercivis platform provided the AMILOIDE project with over 350,000 hours of CPU time during the first seven months of execution.

The returned docking results are very promising, indicating that the top-ranked compounds hold distinctive features from the known TTR binders, with better solubility, lower fraction of potentially toxic halogen atoms and better combination of binding pocket affinities.

Link:
http://www.ibercivis.net

Please contact:

Rui M. M. Brito
Center for Neuroscience and Cell Biology, University of Coimbra, Portugal
E-mail: brito@ci.uc.pt

Swarm Intelligence Approach for Accurate Gene Selection in DNA Microarrays

by José García-Nieto and Enrique Alba

DNA microarrays have emerged as powerful tools in current genomic projects since they allow scientists to simultaneously analyse thousands of genes, providing important insights into the functioning of cells. Owing to the large volume of information that can be generated by a microarray experiment, the challenge of extracting the specific genes responsible for a given illness can only be solved by using automatic means. This challenge has driven our research group at the University of Málaga to design swarm intelligence approaches with the aim of performing accurate biological selection from gene expression datasets (AML-ALL leukemia, colon tumour, lung cancer, etc.).

A DNA microarray consists of a series of thousands of DNA molecules located in different positions within a matrix structure. These DNA molecules are mixed with cellular cDNA molecules during a hybridization process, after which, abundant sequences generate strong signals while rare sequences generate weak signals. Microarrays are normally used to compare gene expression intensity within a sample, and to look at differences in the expression of specific genes among samples. A sample is a test focussing on one disease or on healthy-unhealthy tissues. This is especially appropriate in cancer analysis, since it allows discrimination between tumour tissue and normal tissue. Several gene expression profiles obtained from tumours such as leukaemia, colon, and lung cancers are ready for research in computational biology, and we use them in our research.

The vast amount of data involved in a typical microarray experiment usually requires complex statistical analyses, with the goal of performing a supervised division of the dataset into correct classes. The key issue in this classification is to identify representative gene subsets that may be later used to predict class membership for new external samples. These subsets should be as small as possible in order to develop fast processes for the future class prediction done in an actual laboratory. The main difficulty in microarray classification versus classification of other datasets (found in other domains) is the availability of a relatively small number of samples in comparison to the huge number of genes in each sample (a typical microarray can have a few tens of tested sampled for several thousands of genes). In addition, expression data are highly redundant and noisy, and most genes are believed to be uninformative, as only a small proportion of genes may present distinct profiles for different classes of samples.

In this context, machine learning techniques have been applied to handle large datasets, since they are capable of isolating useful information by rejecting redundancies. In practice, computational feature selection (gene selection, in biology) is often considered as a necessary pre-process step prior to analysing large datasets, in order to reduce the size of the dataset. Feature selection for gene expression analysis often uses supervised classification methods such as K-Nearest Neighbour (K-NN), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) to discriminate a type of tumour. Nevertheless, optimal feature selection is a complex problem that has proved to be NP-hard, and therefore, efficient, automated and intelligent approaches are needed to tackle it.

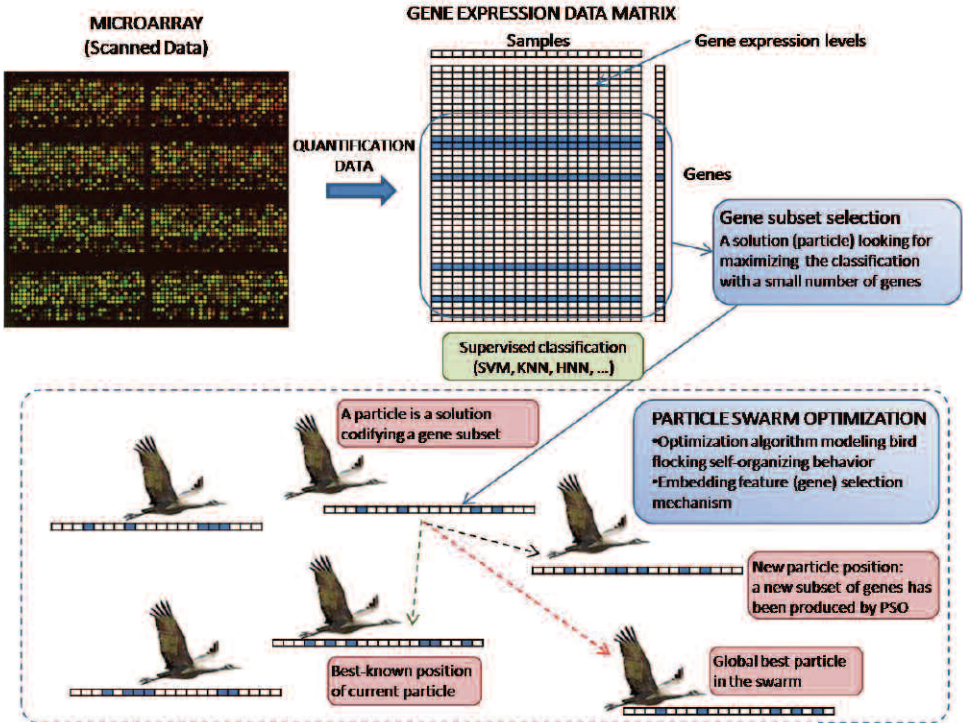

MICROARRAY
(Scanned Data)
QUANTIFICATION DATA
GENE EXPRESSION DATA MATRIX
Samples
Gene expression levels
Genes
Gene subset selection
A solution (particle) looking for maximizing the classification with a small number of genes
Supervised classification (SVM, KNN, HNN, ...)
A particle is a solution codifying a gene subset
PARTICLE SWARM OPTIMIZATION
•Optimization algorithm modeling bird flocking self-organizing behavior
•Embedding feature (gene) selection mechanism
New particle position: a new subset of genes has been produced by PSO
Best-known position of current particle
Global best particle in the swarm

Figure 1: Complete process for optimal gene selection. Once the expression data are generated, our swarm intelligence algorithm obtains optimal subsets of representative genes that are offered to the human specialist as the genes responsible for the illness.

Our research addresses all these issues. We use swarm intelligence algorithms in order to perform an efficient gene selection from large microarray datasets. Swarm intelligence approaches are computational procedures that model the collective behaviour of decentralized and self-organized systems found in nature (ant colonies, bee swarms, bird flocking, etc.) to solve an optimization or search problem. Using this approach, it is possible to reach optimal or quasi-optimal solutions to a given problem. Our goal is to minimise classification error whilst using the smallest possible set of genes to explain the results provided by a microarray. Swarm intelligence can result in savings of both time and resources in comparison to exhaustive and traditional search techniques. Essentially, our model consists of a particle swarm optimization (PSO) algorithm, in which a feature selection mechanism facilitates identification of small samples of informative genes among thousands of genes.

As shown in Figure 1, reported solutions by PSO (codifying gene subsets) are evaluated by means of their classification accuracy using SVM and cross-validation. The contribution of our swarm intelligence approach is notable, since it offers an improvement on existing state of the art algorithms in terms of computational effort and classification accuracy (see the links section). Furthermore, the gene ensembles found by this technique can be successfully interpreted in the light of independent existing results from biology, ie they are biologically (not just computationally) meaningful.

Link:
NEO Research Group:
http://neo.lcc.uma.es

Please contact:

José García-Nieto
University of Málaga/SpaRCIM, Spain
Tel: +34 952133303
E-mail: jnieto@lcc.uma.es

Enrique Alba
University of Málaga/SpaRCIM, Spain
Tel: +34 952132803
E-mail: eat@lcc.uma.es

From Global Expression Patterns to Gene Co-regulation in Brain Pathologies

by Michal Dabrowski, Jakub Mieczkowski, Jakub Lenart and Bozena Kaminska

Understanding how multiple genes change expression in a highly ordered and specific manner in healthy and diseased brains may lead to new insights into brain dysfunction and identification of promising therapeutic targets, for example key signalling molecules and transcriptional regulators. To this end, we have been performing integrated analyses of high-throughput datasets, genomic sequence-derived data and functional annotations.

The computational work of our group (Laboratory of Transcription Regulation, The Nencki Institute) focuses on developing tools that permit a systemic view of global transcriptional/epigenetic changes and applying these tools to the data from experimental models studied in our laboratory, in particular: animal models of stroke, brain tumours and in vitro experimental models of brain tumour-host interactions. In this way computational results can guide experimental effort, which in turn verifies the computational predictions.

Computational/bioinformatics activities of our group focus on the following areas:

- building a comprehensive database of predicted or experimentally identified cis-regulatory regions, transcription factor binding sites (TFBS) and chromatin modifications in vertebrates (human, mouse, rat)
- functional interpretation of global gene expression or other high-throughput data in the context of databases of functional annotations (Gene Ontology) or biological pathways (KEGG)
- identification of transcriptionally co-regulated genes from global gene expression, genomic sequence-derived and chromatin modification data.

We have previously established a local Transcription regulatory regions and motifs (TRAM) database; of putative transcription regulatory regions (AVID-VISTA conserved non-coding regions) and TFBS motif instances (Genomatix) for all pairs of corresponding (orthologous) genes in human and rat. This database is now being updated, extended, and prepared for future automated actualization and public release, as a part of the Nencki Regulatory Genomics Portal (NRGP) project, started in early 2010 and scheduled for four years. The data layer of the NRGP will incorporate additional vertebrate species; and alternative sources of putative regulatory regions; both computationally predicted (BlastZNet, cisRED promoters) and established experimentally (Enhancer VISTA, Ensembl funcgen). These regulatory regions will be scored with each of the three major TFBS motif libraries, both commercial (Transfac, Genomatix) and public (JASPAR). We will implement an innovative way of mapping regulatory regions to the genes, taking into account positions of insulators. The data layer of NRGP is illustrated in Figure 1A.

Nencki Regulatory Genomics Portal will be an internet service providing access to the its data layer and tools for analysis of user-supplied gene expression data, aiming at identification of functions and/or cis-regulatory mechanisms common to many genes. The service will be available by web browser for the users want to analyze the data with the help of provided algorithms, and by a database client for the ones who need direct access to the stored information. DAS technology will be used to make our data layer accessible from the Ensembl genome browser.

Architecture of NRGP, based on MVC (Model-View-Control) model, will consist of the following layers:

- *Data layer* – responsible for all data storage and actualization corresponding to changes made in its sources like Ensembl or motif libraries
- *Application layer* – divided into the front-end responsible for implementation of main algorithms associated with the refinement of

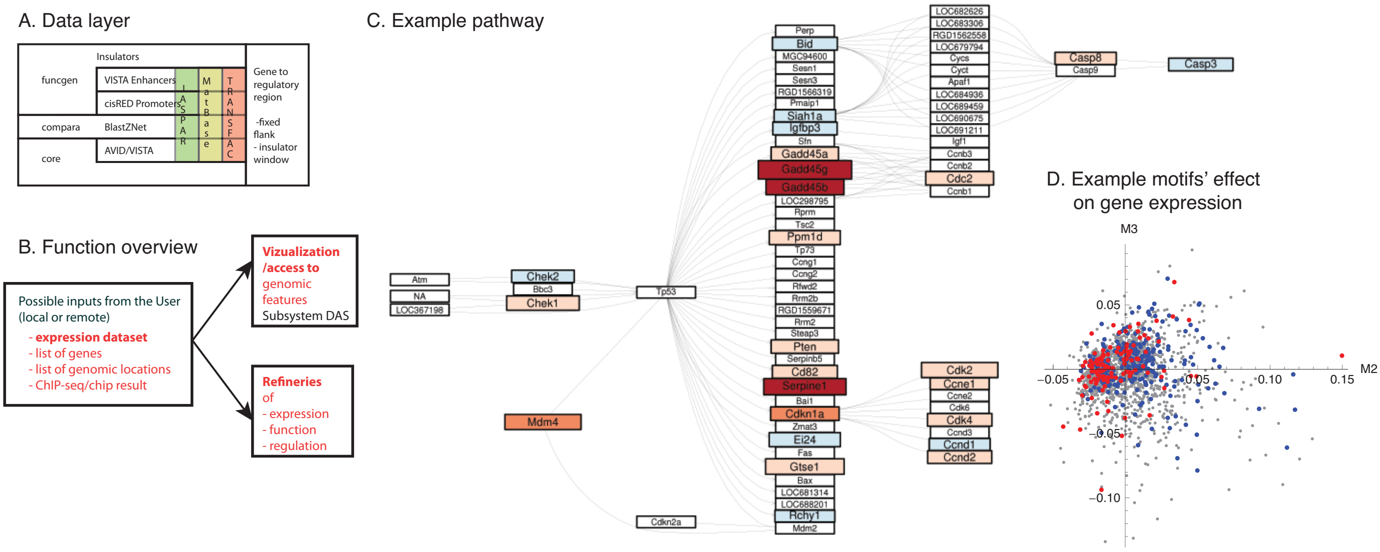


Figure 1: Functionality of the future Nencki Regulatory Genomics Portal.

gene expression, regulation and function; and the back-end – providing services related to the actualization, like searching for the TFBS motifs

- *Presentation layer* – responsible for GUI (Graphical User Interface), adapted to run in web browser.

In order to achieve high level of quality of supported analysis NRGP will integrate worldwide known and used components like Mathematica, R statistical programming language, and commercial libraries. Moreover, the team developing the service consists of experts in scientific areas like: biology of gene expression, and informatics of software engineering, and mathematics of machine learning and data mining.

NRGP will provide the following functionalities (see Figure 1B):

- *Refinement of expression* – starting from an expression dataset uploaded by the user, this module permits identification and visualization of common patterns of gene expression, with standard techniques of analysis microarray gene expression data (analysis of variance, several clustering algorithms, SVD/PCA). In a recent paper, linked below, we show that probe set filtering increased correlations between microarray and qRT-PCR results in two types of studies: detection of differential expression computed by p-values of t-test and estimation of fold change between analyzed groups.

- *Refinement of function* – permits identification of functional annotations statistically associated with a set of genes with a particular pattern of expression, with standard techniques (Fisher Exact Test, Wilcoxon signed rank test). We are currently working on a novel analysis of changed signal (ACS) algorithm for identification of signalling pathways affected by changes in gene expression, which takes into account established topology of signalling pathways, is threshold-free, and does not require the assumption of independence between genes. Figure 1C shows visualization of changes in expression of the genes in a particular signalling pathway.

- *Refinement of regulation* – aims at identification of TFBS motifs and/or chromatin modifications associated with a particular pattern of gene expression, by finding motifs over-represented in putative regulatory regions (Fisher Exact Test), followed by linear or logistic regression to study the effect of motif multiplicity on the given pattern of expression. We previously reported analysis of cis-regulation in subspaces of conserved eigensystems (bi-orthogonal components, also called SVD modes). In a recent paper, linked below, we demonstrate how bi-orthogonality of gene expression data can emerge as a result of biological processes occurring in different cell types, with signal passing between them. In collaboration with Dr. Norbert Dojer (Institute of Informatics, University of Warsaw) we used our method, in the framework of Bayesian Networks learning, to dissect transcriptional regulation in brain following stroke and seizures (Figure 1D). Moreover, we show that effects of motif multiplicity on gene expression analyzed in subspace follow the predictions of the linear response model of gene regulation.

In our research we collaborate with the Computational Biology Group led by Prof. Jerzy Tiuryn (Institute of Bioinformatics, Uniwersity of Warsaw) and with the Linneus Center for Bioinformatics (Uppsala, Sweden) directed by Prof. Jan Komorowski.

Links:
http://www.nencki.gov.pl/pl/struktura/biologia_komorki/lab_02.html
http://www.biomedcentral.com/1471-2105/7/367
http://www.biomedcentral.com/1471-2105/11/104

Please contact:
Michal Dabrowski
The Nencki Institute, Poland
Tel: 4822 58 92 232
E-mail: m.dabrowski@nencki.gov.pl

Testing, Diagnosing, Repairing, and Predicting from Regulatory Networks and Datasets

by Torsten Schaub and Anne Siegel

We use expressive and highly efficient tools from the area of Knowledge Representation for dealing with contradictions occurring when confronting observations in large-scale (omic) datasets with information carried by regulatory networks.

The availability of high-throughput methods in molecular biology has led to a tremendous increase of measurable data along with resulting knowledge repositories, gathered on the web usually within biological networks. However, both measurements and biological networks are prone to considerable incompleteness, heterogeneity, and mutual inconsistency, making it difficult to draw biologically meaningful conclusions in an automated way.

Further probabilistic and heuristic methods exploit disjunctive causal rules to derive regulatory networks from high-throughput -static- experimental data. For instance, disjunctive causal rules on influence graphs were originally introduced in random dynamical frameworks to study global properties of large-scale networks, using a probabilistic approach. These were demonstrated mainly on the transcriptional network of yeast. However, these methods are mostly data driven, and they lack the ability to perform corrections in a fast and global way. In contrast, efficient model-driven approaches based on model checkers - such as multi-valuated logical formalisms - are available to confront networks and measured data. These however, make use of time-series observations and can only be applied to small-scale parametered systems, since they need to consider the full dynamics of the system.

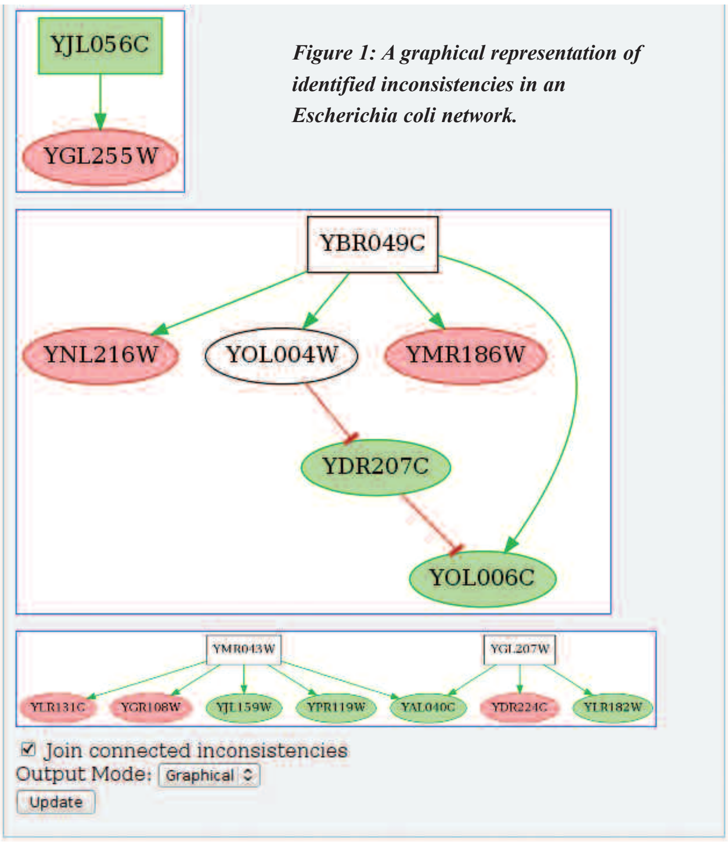
We have proposed an intermediate approach to perform diagnosis on large-scale static datasets. We use a Sign Consistency Model (SCM), imposing a collection of constraints on experimental measurements together with information on cellular regulations between network components.

The main advantage of SCM lies in its global approach to confronting networks and data, since the model allows the propagation of static information along the network and localization of contradictions between distant nodes. In contrast to available probabilistic methods, this model is particularly well-suited for dealing with qualitative knowledge (for instance, reactions lacking kinetic details) as well as incomplete and noisy data. Indeed, SCM is based on influence (or signed interaction) graphs, a common representation for a wide range of dynamical systems, lacking or abstracted from detailed quantitative descriptions.

By combining SCM with efficient Boolean constraints solvers, we address the problem of detecting, explaining, and repairing contradictions (called inconsistencies) in large-scale biological networks and datasets by introducing a declarative and highly efficient approach based on Answer Set Programming [1]. Moreover, our approach enables the prediction of unobserved variations and has shown an accuracy of over 90% on the entire network of E.Coli along with published experimental data. Notably, such genome-wide predictions can be computed in a few seconds.

From the application perspective, the distinguishing novel features of our approach are as follows: (i) it is fully automated, (ii) it is highly efficient, (iii) it deals with large-scale systems in a global way, (iv) it detects existing inconsistencies between networks and datasets, (v) it diagnoses inconsistencies by isolating their source, (vi) it offers a flexible concept of repair to overcome inconsistencies in biological networks, and finally (vii) it enables prediction of unobserved variations (even in the presence of inconsistency).

The efficiency of our approach stems from advanced Boolean Constraint Technology, allowing us to deal with



☑ Join connected inconsistencies
Output Mode: Graphical

Figure 1: A graphical representation of identified inconsistencies in an Escherichia coli network.

problems consisting of millions of variables. Although the basic tools [1] are implemented in C++ we have improved their accessibility by providing a Python library as well as a corresponding Web service [2].

Our project is a joint effort between the Knowledge Representation and Reasoning group [3] at the University of Potsdam and the SYMBIOSE Team [4] at IRISA and INRIA in Rennes. Our techniques have been developed in strong collaboration with the Max-Planck-Institute for Molecular Plant Physiology in Potsdam within the GoFORSYS Project [5] as well as Institut Cochin, Paris [6]. The members of the group include Sylvain Blachon, Martin Gebser, Carito Guziolowski, Jacques Nicolas, Max Ostrowski, Torsten Schaub, Anne Siegel, Sven Thiele, and Philippe Veber.

Links:

[1] http://en.wikipedia.org/wiki/Answer_set_programming
[2] http://potassco.sourceforge.net
[3] http://www.cs.uni-potsdam.de/bioasp/sign_consistency.html
[4] http://www.cs.uni-potsdam.de/wv
[5] http://www.irisa.fr/symbiose
[6] http://www.goforsys.org
[7] http://www.cochin.inserm.fr

Please contact:

Torsten Schaub
Universität Potsdam, Germany
E-mail: torsten@cs.uni-potsdam.de

Anne Siegel
CNRS/IRISA, Rennes, France
E-mail: Anne.Siegel@irisa.fr

MCMC Network: Graphical Interface for Bayesian Analysis of Metabolic Networks

by Eszter Friedman, István Miklós and Jotun Hein

The Data Mining and Web Search Group at the SZTAKI in collaboration with the Genome Analysis and Bioinformatics Group at the Department of Statistics, University of Oxford, developed a Bayesian Markov chain Monte Carlo tool for analysing the evolution of metabolic networks.

“Nothing in biology makes sense except in the light of evolution”. The famous quote by Theodosius Dobzhansky (1900-1975) has been the central thesis of comparative bioinformatics. In this field, the biological function, structure or rules are inferred by comparing entities (DNA sequences, protein sequences, metabolic networks, etc.) from different species. The observed differences between the entities can be used for predicting the underlying function, structure or rule that would be too expensive and laborious to infer directly in lab. These comparative methods have been very successful in silico approaches, for example, in protein structure prediction. The idea can be used for inferring metabolic networks, too.

Metabolic networks are under continuous evolution. Most organisms share a common set of reactions as a part of their metabolic networks that relate to essential processes. A large proportion of reactions present in different organisms, however, are specific to the needs of individual organisms or tissues. The regions of metabolic networks corresponding to these non-essential reactions are under continuous evolution. By comparing metabolic networks from different species, we can find out which parts of the metabolic network are essential (ie those that are common in all networks) and which are non-essential (ie those that are missing in at least one of the networks). Sometimes there is more than one possible metabolic network that can synthesise or degrade a specific chemical. These alternative solutions can be transformed into each other, and the ensemble of all possible reactions form a complicated network (see Figure 1).

The central question is: what are the possible evolutionary pathways through which one metabolic network might evolve into another? This question is especially important to understand in the fight against drug-resistant bacteria. Drugs that are designed to protect us against illness-causing bacteria block an enzyme that catalyzes one of the reactions of the metabolic network of the bacteria. The bacteria, however, can avoid the effects of the drug by developing an alternative metabolic pathway. If we understand how the alternative pathways evolve we may be able to design a combination of drugs from which the bacteria cannot escape through the development of alternative pathways.

Analysis of past events is always coupled with some uncertainty about the nature and order of events that unfolded. It is therefore crucial to infer the evolution of metabolic networks using statistical methods that properly handle the uncertainty that inevitably occurs during analysis. Bayesian methods collect the a priori knowledge into an ensemble of distributions of random variables, set up a random model describing the changes, and calculate the posterior probabilities of what could happen. The relationship between the prior and posterior probabilities is described by the Bayes theorem as shown in Figure 2. Since the integral in the denominator is typically hard to calculate, and the Bayes theorem is often written in the form shown in Figure 3.

The Bayesian theorem in this form can be used in Monte Carlo methods to sample from the posterior distribution. The Markov chain Monte Carlo (MCMC) method sets up a Markov chain that converges to the desired distribution. After convergence, samples from the Markov chain follow the prescribed distribution.

MCMC Network implements the above-described Bayesian MCMC framework for inferring metabolic networks. We model the evolution of networks with a time-continuous Markov

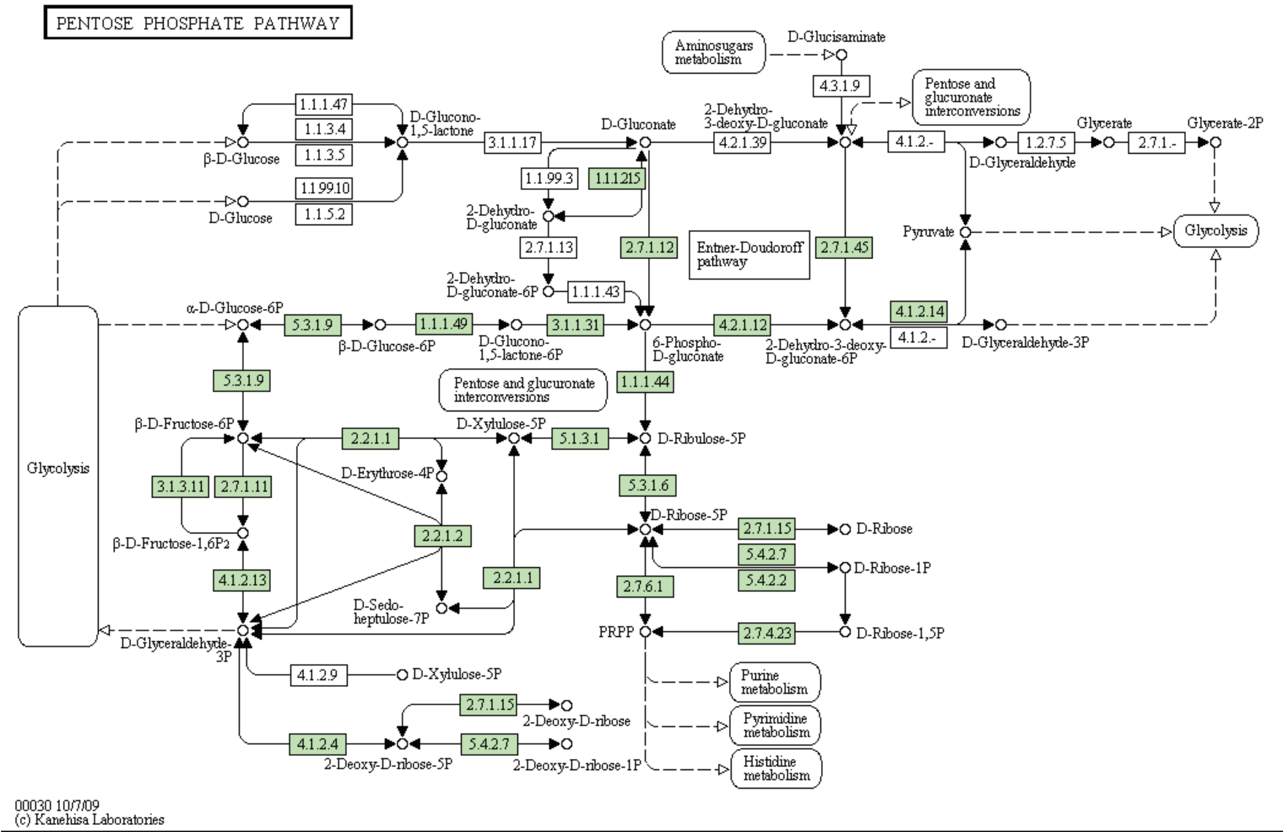


Figure 1: The pentose phosphate pathway of Yersinia pseudotuberculosis strain YPIII. From the KEGG database, http://www.kegg.com

model. In this model, chemical reactions can be deleted from or inserted into the reaction network. Constraints such that the reaction network must be functioning after inserting or deleting a reaction are added to the model. The ??parameter set contains the insertion and deletion rates of the chemical reactions. We know a priori that the important parts of the reaction networks are usually compact parts of the network. The insertion and deletion rates on these important parts are smaller. We can express this knowledge with a Markovian Random Field (MRF) prior distribution. In MRF, the prior probability for a small insertion and deletion rate of a reaction is higher if the neighbour reactions also have small insertion and deletion rates. Our approach is the first implementation of a network evolution in which both the variation of insertion and deletion rates and correlation between the rates of neighbour chemical reactions are considered.

MCMC Network is not only the first implementation of a sophisticated model for chemical network evolution, it also has a user-friendly graphical interface. XML files downloaded from the Kyoto Encyclopedia of Genes and Genomes (http://www.kegg.com/) can be directly used as input files of the program. On the graphical interface, the users can monitor the progress of the Markov chain. The results of the analysis can be visualized on the graphical interface, and can also be saved as a text file for further analysis.

$$P(Y,\Theta\,|\,D) = \frac{P(Y,D\,|\,\Theta)P(\Theta)}{\int_{\Theta'} P(Y,D\,|\,\Theta')P(\Theta')d\Theta'}$$

Figure 2: The relationship between the prior and posterior probabilities. D is the observed data (the metabolic networks), Y is the hidden data (the metabolic networks in the ancestral species that we cannot observe) and Θ is the set of parameters describing the evolution of metabolic networks (the rates with which the reactions of the network are deleted or inserted). P(Y, Θ | D) is the probability that the observed networks evolved from the set of networks Y in a mode described with parameters Θ. P(Y, D | Θ) (called likelihood) is the probability that the random model with parameters Θ generates the observed and unobserved data, and P(Θ) is the prior distribution of the parameters.

$$P(Y,\Theta\,|\,D) \propto P(Y,D\,|\,\Theta)P(\Theta)$$

Figure 3: Bayes theorem in the form as used in Bayesian statistics. Here ∝ stands for 'proportional to'. The theorem is also used in a narrative form posterior ∝ likelihood x prior.

In the last decade, Bayesian methods have been extremely successful for analysing sequence evolution. The evolution of metabolic networks can now also be analysed within the Bayesian framework using our software. There is a continuous debate about the major governing rules in the evolution of networks. For example, it is still unclear which rules result in networks being scale-free. Our software is a useful tool for answering such questions.

Links:

MCMC Network:
http://www.ilab.sztaki.hu/~feszter/mcmcNetwork/

KEGG database:http://www.kegg.com/

Data Mining and Web Search Group:
http://datamining.sztaki.hu

Genome Analysis & Bioinformatics Group: http://www.stats.ox.ac.uk/research/genome

Please contact:

Eszter Friedman
SZTAKI, Hungary
Tel: +36 1 279 6172
E-mail: feszter@info.ilab.sztaki.hu

Drug Dissolution Modelling

by Niall M. McMahon, Martin Crane and Heather J. Ruskin

Recent and continuing work in Dublin City University aims to demonstrate the utility of mathematical and numerical methods for pharmaceutics.

Controlled drug delivery is important for many illnesses. It is often important to ensure that a fixed concentration of drug is available throughout administration of the drug. One type of drug delivery system is the typical tablet that delivers drug as its surface dissolves. Dissolution testing of drug delivery systems is critically important in pharmaceutics. There are three reasons for carrying out dissolution tests: (i) to ensure manufacturing consistency, (ii) to understand factors that affect how the drug enters the blood stream and (iii) to model drug performance in the body. One industry aim is to reduce uncertainty during the development of a new drug. Mathematical and computational simulation will help to achieve this.

Work in Trinity College Dublin, and elsewhere, identified three examples of dissolution physics that are not captured by the well-known Higuchi model. These are: (1) pH changes close to the surface of the dissolving tablet, (2) the effect of particle sizes and (3) the complex hydrodynamics found in typical test devices. The effect of particle size is interesting; large particles of fast-dissolving non-drug additives increase the drug dissolution rate. Understanding how dissolution properties affect drug delivery rates during dissolution seemed a good place for theoreticians to start.

The 1998 Parallel Simulation of Drug Dissolution (PSUDO) project in the Centre for High-Performance Computing at Trinity College Dublin was formed to make improved dissolution models. The team modelled simple cylindrical compacts dissolving in a typical test apparatus. The tablets contained equally-spaced, alternating layers of drug and inert material. This configuration was used because it was a simple starting point and techniques used to model this system might be applied to more complex designs. Finite element code also existed to model dissolution from a layered surface.

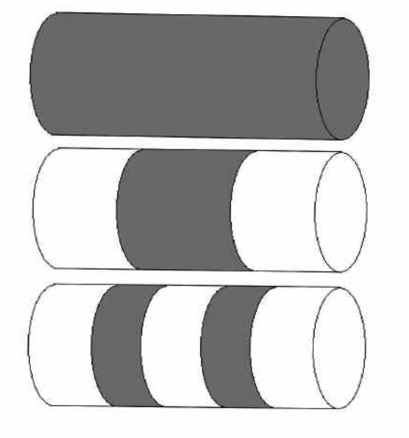


Figure 1: 1-, 3- and 5-layer cylindrical tablets. Dark coloured layers consist of drug. Light coloured layers consist of inert materials.

The useful results from the PSUDO project prompted: (i) continued experimental work, in Trinity College Dublin, (ii) efforts to understand the hydrodynamics of typical test devices, again mostly in Trinity College Dublin, and (iii) improvements to the analytical and numerical models of mass transfer; this work was led by our team in Dublin City University. In addition to the authors of this article, the team included Dr. Ana Barat who investigated the use of probabilistic, Monte Carlo based simulation.

Current work

Recent work by the Centre for Scientific Computing & Complex Systems Modelling (Sci-Sym) at Dublin City University has sought to improve on the PSUDO models of dissolution from multi-layer tablets in pharmaceutical test devices; this work applies to the initial and final phases of a dissolving tablet's life. Results from numerical finite difference models were used to assess how changing the surface boundary conditions, in particular, affects the estimate for drug release in the initial stages of dissolution. Experimental data from our colleagues in the School of Pharmacy and Pharmaceutical Sciences in Trinity College Dublin was used to benchmark our results. How one particular surface boundary condition is implemented can lead to relative errors in the calculated drug dissolution rate of about 10% and consequently, is important. However, our work suggests that much of the physics is captured by the current models.

The second part of our work looked at dissolution from small spherical particles of drug moving about in the test apparatus; this relates to one possible end-state of tablet dissolution, ie fragmentation. This work involved inserting particles into a flow field, again provided by our colleagues in Trinity, and estimating the dissolution rate, due mostly to convection. One finding was that for particles of diameter 100 microns and smaller, radial diffusion dominates as a

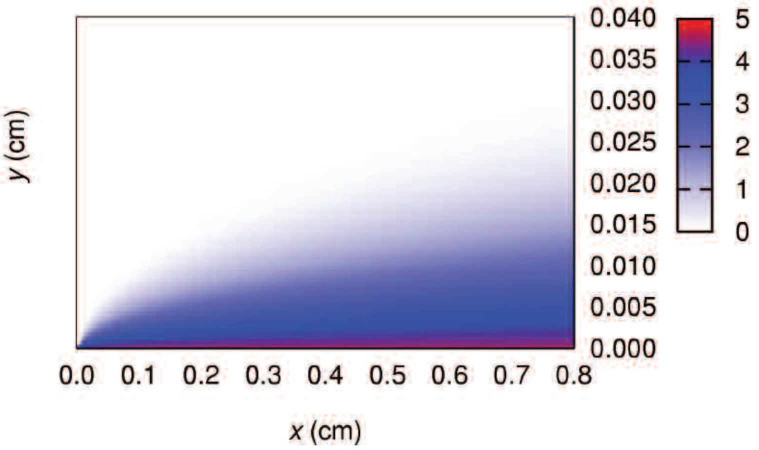


Figure 2: Drug concentration close to the surface of a 1-layer tablet calculated using finite differences. The surface of the tablet is at the bottom of the image and the solvent is flowing from left to right. The concentration scale runs from 0 (no drug present) to 5 (saturated concentration of drug).

mass transfer mechanism. This may help simplify future models.

Most of our work was implemented using Python scripts; we used Crank-Nicolson finite difference schemes, second order accurate in space, first order in the time-like sense. The velocity field within the test apparatus was produced by D'Arcy et al using the commercial CFD code Fluent. LPA and Verlet integration schemes were used to calculate the motion of a particle through the USP velocity field. Mass transfer rates were calculated using empirical correlations.

Future work

In the near future, computing capability, industry needs and a growing acceptance of the utility of cross-disciplinary interaction by professions with strong traditions of experiment, eg, pharmacists, will see the rapid expansion in the use of simulation in pharmaceutical research and development. At present half of the top 40 pharmas use computer simulation. By 2015 it is expected that computer simulation will be a core pharmaceutical research tool.

These two studies, ie (1) drug dissolution from a multi-layered tablet in a dissolution test apparatus and (2) the mass transfer from small particles moving in the test apparatus, each form a part of a future complete framework for simulating drug dissolution in a test apparatus. Work is being continued in DCU by Ms. Marija Bezbradica, a PhD candidate funded by an IRCSET Enterprise Partnership Scholarship.

Acknowledgements

The National Institute for Cellular Biotechnology (NICB) supported much of this research. In addition, the authors acknowledge the additional support of Dublin City University and the Institute for Numerical Computation and Analysis (INCA). Professor Lawrence Crane of INCA was deeply involved in this work as an advisor. Dr. Anne-Marie Healy and Dr. Deirdre D'Arcy et al. of the School of Pharmacy and Pharmaceutical Sciences at Trinity College, Dublin, worked with us and allowed us to access and use their computed flow-fields.

Links:

Centre for Scientific Computing & Complex Systems Modelling, Dublin City University: http://www.sci-sym.dcu.ie/

School of Pharmacy and Pharmaceutical Sciences, Trinity College Dublin:
http://www.pharmacy.tcd.ie/

Institute for Numerical Computation and Analysis (INCA):
http://www.incaireland.org/

Please contact:

Niall McMahon,
Dublin City University and INCA - Institute for Numerical Computation and Analysis, Ireland.
E-mail: nmcmahon@computing.dcu.ie

Martin Crane,
Dublin City University, Ireland.
E-mail: mcrane@computing.dcu.ie

Heather Ruskin,
Dublin City University
E-mail: hruskin@computing.dcu.ie

Computational Modelling and the Pro-Drug Approach to Treating Antibiotic-Resistant Bacteria

by James T. Murphy, Ray Walshe and Marc Devocelle

As antibiotic resistance continues to be a major problem in the health-care sector, research has begun to focus on different methods of treating resistant bacterial infections. One such method is called the* β*-lactamase-dependent pro-drug delivery system. This involves delivering inactive precursor drug molecules that are activated by the same system that normally confers resistance on bacterial cells. In theory this approach seems very promising as it would exploit one of the bacteria's main resistance strategies. However, the complex system dynamics involved are difficult to understand by straightforward experimental observations. Therefore, new computational models and tools are needed to analyse the complex system dynamics involved in this approach to treating antibiotic resistant bacteria such as MRSA. We give an overview here of our work in this area.

A computational model has been developed previously, called Micro-Gen, which simulates the growth of bacterial cells in culture and their interactions with anti-microbial drug molecules. The program uses an agent-based modelling approach whereby the individual bacterial cells are represented by unique software agents that are capable of flexible, autonomous action within a simulated environment. The agent-based approach means that the system as a whole can exhibit a complex behaviour that is more than the "sum of its parts". This approach allows the unique dynamics within a bacterial colony to be simulated by taking into account subtle differences within a population or across an environment.

Micro-Gen has been used in previous studies to examine the resistance mechanisms involved in the response of bacterial populations to antibiotic treatment. Studies showed that it could accurately predict the minimum inhibitory concentrations (MIC, a simple laboratory measure of antibiotic efficacy) for various common antibiotics, including penicillin G and cephalothin, against methicillin-resistant Staphylococcus aureus (MRSA) bacteria. However, an important strength of the model is that it can be used to examine new approaches for treating antibiotic-resistant bacteria and give insight into potential novel drug treatment strategies. This can aid in rational drug design by allowing a greater understanding of the underlying

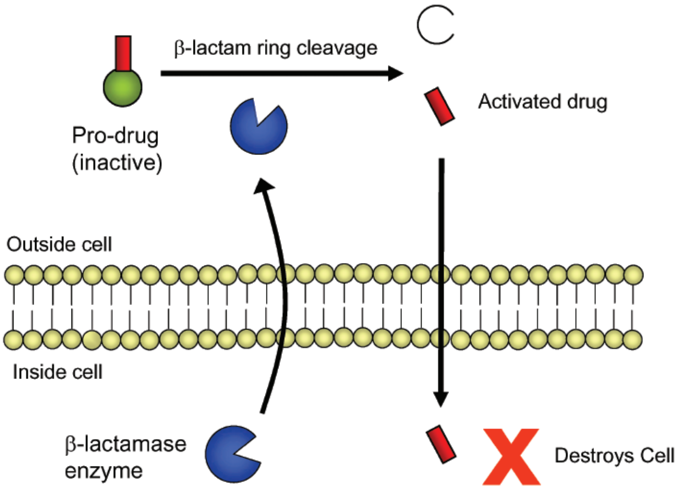


Figure 1: Simple schematic representation of activation of pro-drug molecule as a result of cleavage by a β-lactamase enzyme (blue) released by the bacterial cell. Activation results in the release of the active anti-microbial component (red) which enters the cell and destroys it.

mechanistic principals determining response to treatment.

Recent research has focussed on adapting the existing model to explore a novel approach to treating antibiotic-resistant bacteria called the β-lactamase-dependent pro-drug delivery system. This approach involves administering a substrate-like pro-drug molecule that contains a β-lactam ring structure (a structure contained in many common antibiotics including penicillin and its derivatives). The presence of a β-lactam ring structure means that the pro-drug effectively mimics a normal antibiotic, such as penicillin, and is susceptible to cleavage by β-lactamase enzymes released from antibiotic-resistant bacterial cells. However, this cleavage does not result in the destruction of the drug (as is the case with the antibiotics) but instead triggers the selective release of a molecule with anti-microbial properties that kills or inhibits the growth of the bacterial cells.

This is considered a promising therapeutic approach because many bacterial species have evolved to produce β-lactamase enzymes in response to prolonged clinical exposure to β-lactam antibiotics. The bacterial cells produce β-lactamase enzymes as a defence mechanism because the enzymes cleave the β-lactam ring structure present in the antibiotic molecules, rendering them inactive. In the USA, it is estimated that greater than 95% of all S. aureus bacterial isolates possess resistance to penicillin, due to the expression of β-lactamase. Therefore, by designing pro-drugs that specifically target these bacteria it would introduce an evolutionary selective pressure contrary to that of existing β-lactam antibiotics.

The dynamics between the negative evolutionary selective pressure from pro-drugs and positive selective pressure from β-lactam antibiotics would be an important factor to consider when assessing the possible evolution of drug resistance in bacteria in response to these two different therapeutic strategies. However, the complex interplay of biophysical, pharmacokinetic, pharmacological and epidemiological factors that would contribute to this is difficult to predict. Therefore it is important to develop novel modelling approaches that can be used to analyse the complex system dynamics involved and to develop theories that can then be tested in the lab. That is the purpose of our research: to develop new models that may be suitable to approach this problem.

A detailed description of the Micro-Gen model is beyond the scope of this article, but for further information the reader is directed towards the publication list on our website (see link) documenting the overall program structure along with an analysis of the mechanistic basis for its output. Micro-Gen is an agent-based model, which means that it represents the system from the bottom-up, looking at the individual components of a population and how their interactions contribute to the overall “emergent” dynamics of a population. In this case, the individual bacterial cells are represented by software agents that store physical traits such as their energy state or amount of antibiotic damage. The agents also have behavioural rules associated with them that dictate their actions during the simulation. The environment of the simulations is represented by a discrete, two-dimensional grid containing diffusible elements such as nutrients, β-lactamase enzymes and the anti-microbial (pro-) drug compounds.

The key interaction of the model that determines the response to pro-drug treatment is the reaction between the β-lactamase enzyme and the pro-drug (see Figure 1). This reaction is the activation step that triggers the release of an active drug compound. The success of the pro-drug approach requires the rapid and specific release of the active drug compound in the vicinity of the bacterial cells. The model contains a quantitative representation of this reaction based on Michaelis-Menten kinetic theory.

It is early days in the research of pro-drug compounds and only limited laboratory data is available. However, the power of the computational approach for elucidating the mechanisms of action of novel drug compounds can be an important asset to have. In conjunction with laboratory testing, important insights can be made into the complex interplay of the different components in the pro-drug delivery system using an agent-based modelling approach. Initial testing with our own model has involved carrying out simulations based on data from real-life pro-drug candidates in order to compare the model output to experimental results (in press). This has demonstrated its usefulness in providing an integrated understanding of the unique dynamics of the pro-drug delivery system in order to assess its effectiveness as a viable alternative treatment strategy for microbial infectious diseases.

Links:
http://www.computing.dcu.ie/~jamurphy/
http://sci-sym.computing.dcu.ie/

Please contact:

James T. Murphy
Dublin City University, Ireland.
Tel: +353 1 700 6741
E-mail: jamurphy@computing.dcu.ie

Computational Systems Biology in BIOCHAM

by François Fages, Grégory Batt, Elisabetta De Maria, Dragana Jovanovska, Aurélien Rizk and Sylvain Soliman

The application of programming concepts and tools to the analysis of living processes at the cellular level is a grand challenge, but one that is worth pursuing, as it offers enormous potential to help us understand the complexity of biological systems. To this end, the Biochemical Abstract Machine (Biocham) combines biological principles with formal methods inspired by programming, to act as a modelling platform for Systems Biology. It is being developed by the Contraintes research team at INRIA.

Biologists use diagrams to represent complex systems of interaction between molecular species. These graphical notations encompass two types of information: biochemical reactions (eg, protein complexation, modification, binding to a gene, etc.) and regulations (of a reaction or transcription). Based on these structures, mathematical models can be developed by equipping such molecular interaction networks with kinetic expressions leading to quantitative models. There exist two general categories of quantitative model: ordinary differential equations (ODEs), which offer a continuous interpretation of the kinetics; and continuous-time Markov chains (CTMCs) which provide a stochastic interpretation of the kinetics.

The Systems Biology Markup Language (SBML) uses a syntax of reaction rules with kinetic expressions to define such reaction models in a precise way. Nowadays, an increasing collection of models of various biological processes is available in this format in model repositories, such as www.biomodels.net, and an increasing collection of ODE simulation or analysis software platforms are now compatible with SBML.

Since 2002, we have been investigating the transposition of programming concepts and tools to the analysis of living processes at the cellular level. Our approach relies on a logical paradigm for systems biology which is based on the following definitions:

- biological model = (quantitative) state transition system
- biological properties = temporal logic formulae
- biological validation = model-checking
- model inference = constraint solving.

Our modelling software platform Biocham (http://contraintes.inria.fr/biocham) is based on this paradigm. An SBML model can be interpreted in Biocham at three abstraction levels:

- the continuous semantics (ODE on molecular concentrations)
- the stochastic semantics (CTMC on numbers of molecules)
- the Boolean semantics (asynchronous Boolean state transitions on the presence/absence of molecules).

Of the three levels, the Boolean semantics is the most abstract. It can be used to analyse large interaction networks without known kinetics. These formal semantics have been related within the framework of abstract interpretation, showing for instance, that the Boolean semantics is an abstraction of the stochastic semantics, ie that all possible stochastic behaviours can be checked in the Boolean semantics, and that if a Boolean behaviour is not possible, it cannot be achieved in the quantitative semantics for any kinetics.

The use of model-checking techniques, developed over the last three decades for the analysis of circuits and programs, is the most original feature of Biocham. The temporal logics used to

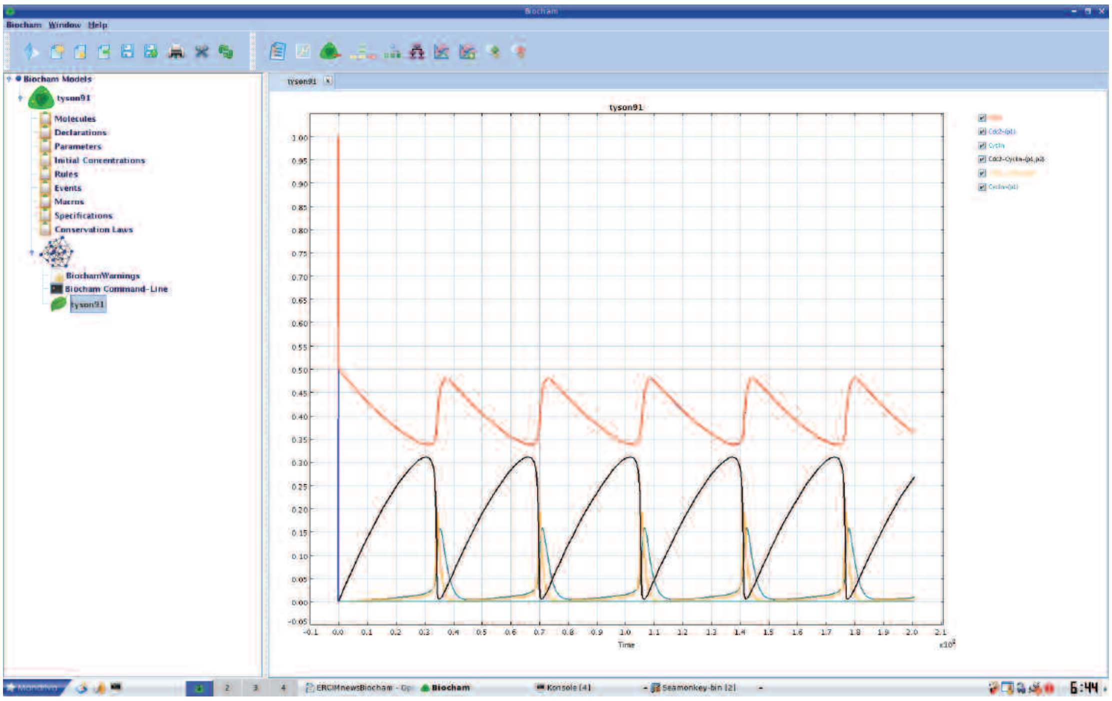


Figure 1: Screenshot of BIOCHAM.

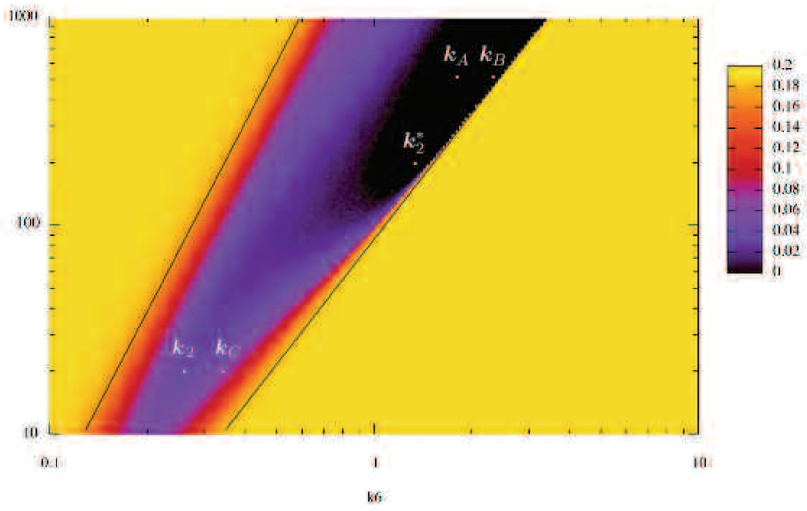


Figure 2: Continuous satisfaction degree of a temporal logic formula for oscillations with a constraint of amplitude in a landscape obtained by varying two parameters.

formalize the properties of the behaviour of the system are respectively the Computation Tree Logic, CTL, for the Boolean semantics, and a quantifier-free Linear Time Logic with linear constraints over the reals, LTL(R), for the quantitative semantics.

Biocham has been used to query large Boolean models of the cell cycle (eg Kohn's map of 500 species and 800 reaction rules) by symbolic model-checking, formalize phenotypes in temporal logic in a much more flexible way than by curve fitting, search parameter values from temporal specifications, measure the robustness of a system with respect to temporal properties, and develop in this way, quantitative models of cell signalling and cell cycle for cancer therapies.

For some time, an important limitation of this approach was due to the logical nature of temporal logic specifications and their Boolean interpretation by true or false. By generalizing model-checking techniques to temporal logic constraint solving, a continuous degree of satisfaction has been defined for temporal logic formulae, opening up the field of model-checking to optimization in high dimension.

Biocham is freely distributed under the GPL license. In our group, it is currently used in collaboration with biologists at INRA for developing models of mammalian cell signalling, at INSERM and in the European consortium EraSysBio C5Sys for developing models of cell cycle for optimizing cancer chronotherapies, and since 2007 has been used in MIT's iGem competition, for designing synthetic biology circuits.

Link:
http://contraintes.inria.fr/Biocham

Please contact:
François Fages
INRIA Paris-Rocquencourt, France
E-mail: Francois.Fages@inria.fr

Prediction of the Evolution of Thyroidal Lung Nodules Using a Mathematical Model

by Thierry Colin, Angelo Iollo, Damiano Lombardi and Olivier Saut

Refractory thyroid carcinomas are a therapeutic challenge owing to some being fast-evolving - and consequently being good candidates for trials with molecular targeted therapies - whilst others evolve slowly. This variation makes it difficult to decide when to treat. In collaboration with Jean Palussière and Françoise Bonichon at the Institut Bergonié (regional centre for the fight against cancer), we have developed a diagnostic tool to help physicians predict the evolution of thyroidal lung nodules.

The evolution of thyroidal lung nodules may be difficult to evaluate. Furthermore, when unwell patients are concerned, physicians try to minimize the use of invasive techniques, restricting treatment (by radiofrequency ablation) to nodules that may become malignant. Thus, an accurate prognosis of each nodule is critical. We propose a numerical method of predicting the actual tumour growth for a specific patient.

Classically, accurate mathematical models describing tumoral growth involve a large number of parameters that cannot always be recovered from experimental data. The model proposed here is tuned for each patient thanks to two medical images following the evolution of a nodule. From this analysis, it is possible to obtain an estimate of the evolution of a targeted nodule using only non-invasive techniques.

Our model describes, not only the volume of the tumour, but also its localization and shape. It takes into account nutrient concentration, cell-cycle regulation and evolution of populations of cells, as well as mechanical effects. Our prediction relies on parameters estimation using temporal series of MRI or scans. The approach uses optimization techniques and Proper Orthogonal Decomposition (POD) to estimate the parameters of the chosen mathematical model (adapted to the type of cancer studied) that best fit with the real evolution of the tumour shown on the images.

Our model is a simplified Darcy-type model describing the evolution of various cellular densities (proliferative and quiescent cancer cells, healthy tissue…) as well as nutrients distribution or mechanical variables using Partial Differential Equations (PDE). This

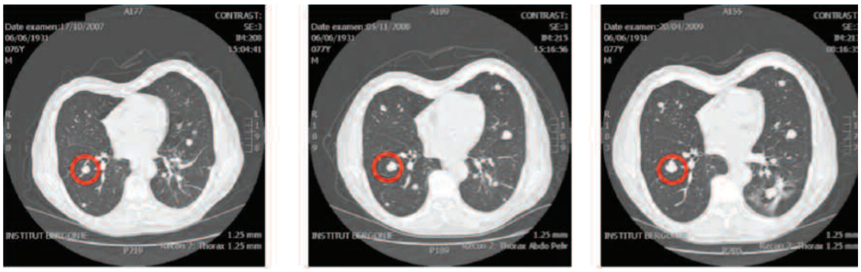


Figure 1: Evolution of the an untreated thyroidal nodule in the lung.

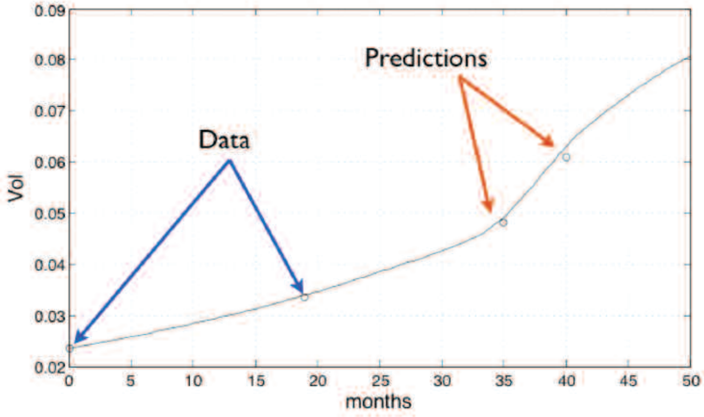


Figure 2: results of a prediction of the evolution of the nudule based on the mathematical model.

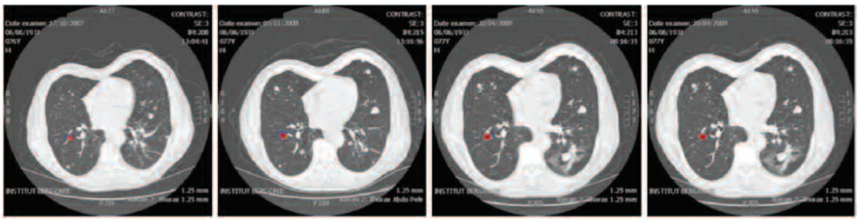


Figure 3: In addition to computing the volume of the tumour, the model can be used to predict the localization of the tumour (plotted in red).

parametric model is sufficiently accurate to take into account the main physical features of tumor growth but simple enough to have its parameters recovered.

Essentially, the technique may be summarized as follows. Initially, the nodule under investigation is marked by the physicians on successive CT scans. From these images we recover the geometry of the lung and the shape of the nodule at different times. From the initial shape of the nodule, we run numerous numerical simulations of our mathematical model using a large set of parameter values. A basis using a POD approach is extracted from this large collection of numerical results. This time-consuming process can be efficiently performed on a High Performance Parallel Architecture since the direct simulations can be run concurrently and the POD extraction uses a parallel algorithm to be as fast as possible. The last step of the procedure consists in solving an inverse problem based on a Newton method to recover the parameters that best fit the available medical images of the nodule. Once these parameters are determined, a prediction is simply obtained by running the numerical code.

We tested our technique on several test cases, one of which is presented in Figure 1. For a patient with an untreated thyroidal nodule in the lung, four computerised tomography (CT) scans were available (we show three of them illustrating the evolution of the nodule below).

We used the first two scans to perform the data assimilation and recover the parameters of the mathematical model adapted to the patient and to this nodule. Once these parameters are determined, the model is used in order to obtain predictions on the evolution of this nodule. The results of this prediction are plotted in Figure 2. The measured volume of the nodule from the CT scan is plotted using circles; the continuous line represents the volume computed using our tuned mathematical model.

Only the first two scans were used for data assimilation; the others are shown for the purpose of comparison. In addition to computing the volume of the tumour,our model enables usto predict its localization, as shown in the Figure 3 where the computed tumour is plotted in red.

For oncologists the development of such tools is of interest in therapy planning (and in the evaluation of an antitumoral treatment). For example a slowly evolving tumour prediction could reinforce the decision to wait without specific treatment. In the opposite case the simulation can support the decision to start a radiofrequency thermal ablation (for example) or a molecular targeted therapy.

Links:
INRIA Resarch Team MC2
http://www.math.u-bordeaux1.fr/MAB/mc2/
Institut Bergonié
http://www.bergonie.org/.

Please contact:

Olivier Saut,
INRIA, France
Tel +33 5 40002115
E-mail: Olivier.Saut@inria.fr

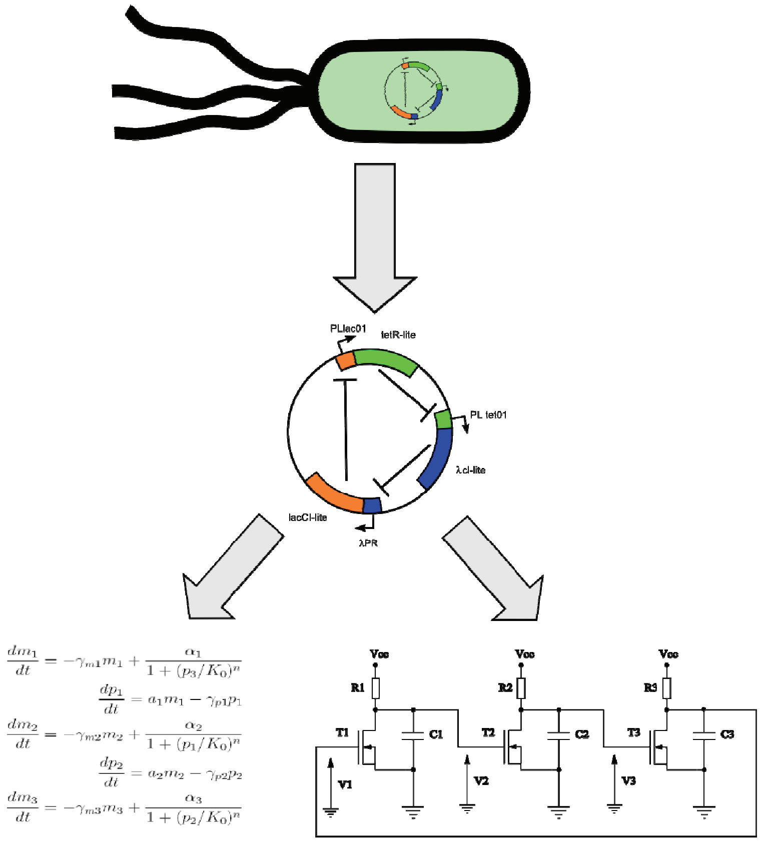
Dynamical and Electronic Simulation of Genetic Networks: Modelling and Synchronization

by Alexandre Wagemakers and Miguel A. F. Sanjuán

Cells can be considered to be dynamical systems that possess a high level of complexity due to the quantity and range of interactions that occur between their components, particularly between proteins and genes. It is important to acquire an understanding of these interactions, since they are responsible for regulating fundamental cellular processes.

The climax of the genome project, the most notable success of which, has been the complete sequencing of the human genome, was the identification of all the genes that comprise the genetic material of an organism. This achievement led to a new phase of the project: the postgenomic era. Research is now focusing on understanding the organization of, and interactions between proteins, the product of gene expression. Each protein is in charge of a function which can induce changes in other molecules in the cell, such as enzymes or even hormones. These molecules can be viewed as the nodes of a network where the interactions are the links. Thus we can view the system as a complex network of regulation interaction which is responsible for the functioning of the cell.

Recently, the design and the construction of artificial networks has been proposed as a means of studying biological processes, such as oscillations of the metabolism. These networks, simpler than the natural ones, can contribute to the understanding of the molecular basis of a specific function. Simple mathematical models can be constructed in order to perform qualitative and numerical analyses, and synthetic genetic networks can even be synthesized in a laboratory. These works, among others, gave birth to the so-called synthetic biology which integrates several scientific fields such as nonlinear dynamics, physics of complex systems and molecular bioengineering. This is a newly emerging field with a strong interdisciplinary component in which the future advances seem very promising.



$$\frac{dm_1}{dt} = -\gamma_{m1}m_1 + \frac{\alpha_1}{1+(p_3/K_0)^n}$$
$$\frac{dp_1}{dt} = a_1m_1 - \gamma_{p1}p_1$$
$$\frac{dm_2}{dt} = -\gamma_{m2}m_2 + \frac{\alpha_2}{1+(p_1/K_0)^n}$$
$$\frac{dp_2}{dt} = a_2m_2 - \gamma_{p2}p_2$$
$$\frac{dm_3}{dt} = -\gamma_{m3}m_3 + \frac{\alpha_3}{1+(p_2/K_0)^n}$$

Figure 1: Genetic networks in living cells can firstbe identified with molecular genetics techniques. Once the network is identified a mathematical model is developed and analysed using nonlinear dynamics methods and electronic modelling.

The paradigmatic examples of synthetic genetic networks are the genetic toggle switch and the represilator. The genetic toggle switch is the combination of two mutually repressing genes forming a bistable system whose state can be changed with an external signal. One can say that this genetic switch has memory, since it remains in its current state until an external inducer acts again. The second paradigmatic system is the repressilator, which is in fact, a genetic oscillator. In this system three repressor genes are placed in a ring, with each repressor inhibiting the production of the following protein with a given delay. This configuration leads to oscillations in the expression of the three proteins.

Our work in this field, in the Nonlinear Dynamics, Chaos and Complex Systems Group at the Universidad Rey Juan Carlos (URJC), consists mainly of the application of nonlinear dynamics techniques to the modelling and simulation of genetic networks. The evolution of the protein concentration of a particular gene can be represented mathematically with a set of ordinary differential equations. Once these equations are defined, we can apply methods from nonlinear dynamics such as phase space analysis, bifurcation diagrams and stability analysis to understand and predict the behavior of any synthetic genetic network. Furthermore these tools are useful for the design and study of laboratory experiments.

We also have proposed an alternative way to design and analyse the genetic networks with analog electronic circuits. The dynamics of a regulatory genetic network can be simulated with very simple nonlinear circuits based on MOSFET transistors. Our circuits allow a one-to-one correspondence between the structure of the genetic and electronic networks, and their analog character extends this correspondence to the full dynamical behavior. An obvious benefit of this approach is that the electronic circuits are easier to implement experimentally than genetic circuits. We have applied this technique successfully in studies of the dynamics and synchronisation of populations of genetic networks, such as the repressilator and the toggle-switch. Synchronisation of a population of such units has been thoroughly studied, with the aim of comparing the role of global coupling with that of global forcing on the population. We have also analysed a method for the prediction of the synchronisation of a network of electronic repressilators, based on the Kuramoto model. Our research indicates that nonlinear circuits of this type can be helpful in the design and understanding of synthetic genetic networks.

Link:
http://www.fisica.escet.es

Please contact:
Alexandre Wagemakers
Universidad Rey Juan Carlos, Madrid, Spain
Tel: +34 91 4888242
E-mail:
alexandre.wagemakers@urjc.es

Formal Synthetic Immunology

by Marco Aldinucci, Andrea Bracciali and Pietro Lio'

The human immune system fights pathogens using an articulated set of strategies whose function is to maintain in health the organism. A large effort to formally model such a complex system using a computational approach is currently underway, with the goal of developing a discipline for engineering "synthetic" immune responses. This requires the integration of a range of analysis techniques developed for formally reasoning about the behaviour of complex dynamical systems. Furthermore, a novel class of software tools has to be developed, capable of efficiently analysing these systems on widely accessible computing platforms, such as commodity multi-core architectures.

Computational approaches to immunology represent an important area of systems biology where both multi-scale and time and spatial dynamics play important roles. In order to explore how different modelling and computational techniques can be better integrated to support in silico experiments, a collaboration amongst researchers of Italian and British institutions has been established. In the long term we anticipate that this will lead to the engineering of synthetic immune responses.

In silico experiments involve the reproduction of the dynamics occurring between the immune system and pathogens. These models can be used to simulate the mechanisms and the emerging behaviour of the system, to test new hypotheses and to predict their effects, hence, they will need to be able to measure, analyse and formally reason about the system behaviour. By means of such a virtual lab one can try to design novel "synthetic" responses and drug treatments that may then be implemented in the organic world.

We are exploring the feasibility of combining two research trends, which will certainly be further developed in the near future. The first trend draws from formal methods, ie a set of description techniques that allows qualitative and quantitative aspects of system behaviour to be described and formally analysed. Generally, models consist of populations of agents/entities, or aggregate abstractions of them, eg viruses or lymphocytes. These play a part in the economy of the whole system by means of the possible behaviour they exhibit. The overall behaviour of the system emerges from the interaction between agents and can be observed by means of simulations that account for either probabilistic or averaged evolutions from given initial conditions. Also, the system can be observed in its transient or equilibrium dynamics, and its properties, when precisely expressed in a formal language, can be verified (model checked) against the system behaviour. In particular the formal analysis can be tightly coupled with stochastic simulations in order to improve the information obtained from the simulation results. These approaches, particularly the stochastic ones under certain hypotheses, are extremely computationally expensive.

The second trend aims to enhance the precision and effectiveness of these virtual labs. A good opportunity to achieve this is provided by the recent design shift towards multi-core architectures that make high computational power diffusely available. To exploit this power, however, software tools need to be redesigned to match the new architectures, which may suffer from inefficiencies of the shared memory subsystem due to poor memory access hot-spots and caching behaviour. Within the FastFlow programming framework [1], which provides a cache-aware, lock-free approach to multi-threading, we have developed StochKit-FF (see application page at [1]), an efficient parallel version of StochKit, a reference stochastic simulation tool-kit.

In our reference scenario we use aspects of the HIV dynamics within a patient, as a model infectious disease. Many different viral strains contribute to infection progression; they use different cell receptors (CCR5 and CXCR4) and consequently several anti-HIV therapies target these receptors. Progression of the disease (AIDS) is dependent on interactions between viruses and cells, the high mutation rate of viruses, the immune response of individuals and the interaction between drugs and infection

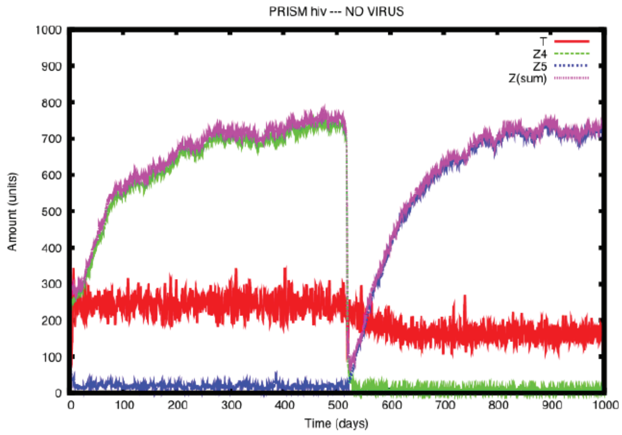


Figure 1: A single stochastic simulation over the [1-1000] day interval. Levels of cells from the immune system during HIV infection are shown on the y-axis. Importantly, the decay of T below about 200 represents the passage to AIDS. This emerges at about day 500.

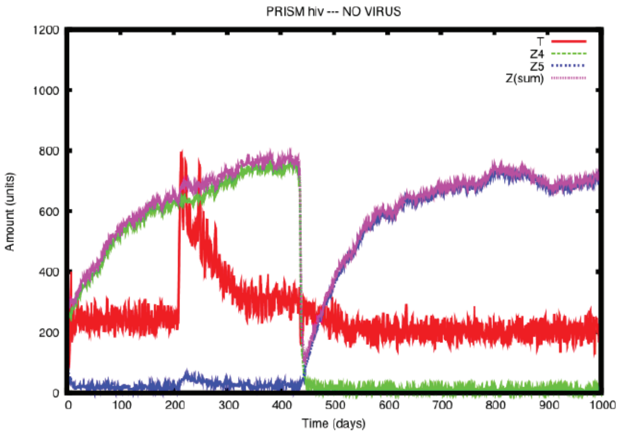


Figure 2: The same scenario as that represented in Figure 1 with the addition of a treatment period starting at about day 200. The effect of the treatment can be "visually" appreciated in terms of an higher amount of T (and Z5) during the treatment period, with respect to the same time interval in Figure 1.

dynamics. These phenomena are inherently stochastic, subject both to intrinsic and extrinsic noises. Some simulation results are shown in Figure 1 and Figure 2.

StockKit-FF allows us to run multiple concurrent stochastic simulations efficiently. Results based on a large number of simulations are much more informative than a few probabilistic outcomes. Since the cost of saving comprehensive information about each simulation can become prohibitive, StochKit-FF allows the on-line parallel reduction of solution trajectories from different simulation runs by way of user-defined statistical estimators, such as average and variance. The relative speed of StockKit-FF with respect original StockKit for the HIV model is reported in Figure 3.

Furthermore, we are quantitatively assessing different scenarios regarding the immune system response. Probabilistic model checking, for instance, can provide a greater understanding of infection models, by shifting the attention from an informative, but empirical, analysis of the graphs produced by simulations towards more precise quantitative interpretations. For instance, this technique can allow us to determine the probability of immune system failure with or without a support therapy over a given time interval. Beyond the visual evidence, it is possible to determine that such a probability for the cases in Figure 1 and Figure 2 is 0.377 vs. 0.855 (results obtained by the PRISM model checker [2], see [3] for details).

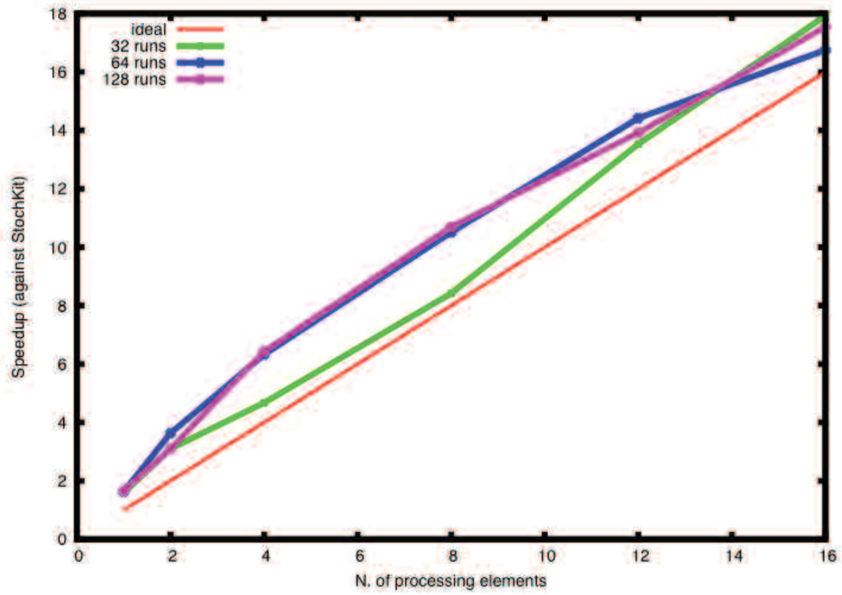


Figure 3: Speedup of StockKit-FF against original StochKit for 32, 64, and 128 runs of the HIV stochastic simulation. StockKit-FF runs are concurrently executed; their outputs are reduced by way of the average and variance functions. StockKit-FF exhibits a super-linear speedup, ie it is always more than n-fold faster than StockKit when running on a n-core platform.

Currently, we are investigating more structured ways of collecting information from multiple runs and the possibility of exploiting such multiple runs as approximated model knowledge for model checking purposes. Efficient simulations will make possible an effective investigation of relevant characteristics of the virus attack and the immune defence, while the formal analysis will provide the key tool to identify best anti-viral drug therapy.

This project is a British-Italian collaboration (Universities of Cambridge, Torino and Stirling, ISTI-CNR), partially supported by HPC-Europe, EMBO and the CNR project RSTL-XXL.

Links:
[1] http://mc-fastflow.sourceforge.net
[2] http://www.prismmodelchecker.org/
[3]
http://www.biomedcentral.com/1471-2105/11/S1/S67

Please contact:

Andrea Bracciali
ISTI-CNR, Italy and
Computing Science and Mathematics,
University of Stirling, UK
E-mail: braccia@isti.cnr.it,
braccia@cs.stir.ac.uk

Pietro Lio'
Computer Laboratory
University of Cambridge
E-mail: pl219@cam.ac.uk

An Ontological Quantum Mechanics Model of Influenza

by Wah-Sui Almberg and Magnus Boman

Seasonal flu has prevailed in the temperate zones for 400 years without adequate scientific explanation. We suggest a novel approach to modelling influenza building on the ideas and theories of David Bohm.

Our work originates from a project that has been underway since 2002. The project takes a multi-disciplinary approach to modelling the spread of infectious disease, employing research from medicine, computer science, statistics, mathematics, and sociology. This approach led not only to a large programming effort, hosted by the Swedish Institute for Infectious Disease Control, but also to the formation of the Stockholm Group for Epidemic Modeling (S-GEM). The micro-model built by the project group holds nine million individuals, corresponding to the entire population of Sweden. Variations of this model have been used for smallpox and influenza, and results have been published within academia. The model will go open source in 2010.

Most mathematical work in the project, as in epidemiology in general, focuses on traditional macro-models, in the Susceptible Infectious Recovered (SIR) family. Additionally, the project considers, micro- and meso- effects such as individual movement (between work and home, for example) and family structures (for instance, where a person lives, and with whom). Whilst being tentative and somewhat speculative, the work described in this short overview is potentially extremely fruitful. It targets the enigma of seasonal influenza, and how it relates to the prospect of future lethal pandemics. Resting on the foundation of our experience of micro-meso-macro modelling in the burgeoning field of computational epidemiology, we propose a novel way of modelling influenza. This work is carried out in cooperation between Stockholm University, the Royal Institute of Technology and the Swedish Institute of Computer Science (SICS). While our international network is extensive, we would welcome any input from the ERCIM partners on our work, to be presented as a PhD thesis by Wah-Sui Almberg in 2010.

Figure 1: Depiction of the Influenza virus protein (neuraminidase) structure, using the NASA Ribbons program. Image courtesy of NASA.

Infectious disease is the cause of more human deaths than any other individual factor, and influenza kills a larger number of people than any other infectious disease. Even in less severe years, more than a million deaths worldwide can be attributed to influenza. Adding to the burden of the direct suffering caused by the contraction of the disease, influenza may have negative social consequences, including a drastic fall in productivity and possible societal conflicts; effects that may cause further spread of the disease and make it more difficult to care for the sick.

Currently, little is known about the transmission process of influenza. In the case of the thoroughly investigated H5N1 virus, for instance, the World Health Organisation acknowledged that no scientific explanation for its unusual pattern of spread could be identified. Furthermore we do not know which subtype—there are 16 hemagglutinin (H) and 9 neuraminidase (N) subtypes—the next pandemic will have. Hence, it is very hard to develop a vaccine and determine its efficacy beforehand.

In the temperate zones, the epidemic spread of influenza is strongly governed by forces that coincide with season, with epidemic activity usually peaking in winter. In these regions, summer outbreaks are uncommon and do not develop into epidemics. There are various explanations for this phenomenon which consider the role of climate, school semesters, and complex network transmission effects. We hypothesise that a unification of the micro-, meso-, and macro- modelling within a single model would give new insights into the social dynamics involved in the spread of influenza. Since the micro and macro-modellers traditionally represent opposite and competing sides, however, and since their synthesis is a long-unresolved problem, we opt for a radically different kind of unification. David Bohm, encouraged by Einstein, outlined an approach that could resolve the micro-macro dichotomy. Originally, his aim was to bring quantum mechanics (a micro perspective) and the theory of relativity (a macro perspective) under one theoretical umbrella, but the philosophy and the mathematics that emerged suggest a more general applicability.

The ontological theory developed by Bohm is a kind of algebra that was originally designed to describe thought and consciousness, but it has also proven valuable for understanding the social dynamics of financial systems. Essentially, Bohm's scheme is one of participation rather than of interaction. In a set of individuals constituting a population, every individual participates in every other individual's existence and actions; that is, the totality is present implicitly in every individual, and every individual makes a mark on the totality. Bohm refers to this holistic concept as the implicate order. The concept of wholeness makes it impossible for primary laws to be summarized in a simple set of statements, since every aspect of reality enfolds all other aspects of it in the implicate order. In contrast, the explicate order, which current laws of all natural sciences are based upon, refers to the apparent reality of things. Starting with the explicate order, Bohm extracted the operations displacement (D) and rotation (R) in Euclidean space. Founding a description of reality on D and R, instead of on static elements such as straight lines, circles, etc, Bohm created a system in which process and movement are primary, and spatial coordinates of a secondary nature. As these operators are multiplied and added, orders and measures can be defined. Bohm then added transformations to these basic operations, as is customary, but then he introduced an operator called metamorphosis (M). Using M, sets of transformations (E) can be translated to other Es. This is mathematically stated as $E' = MEM^{-1}$. Adding a unit and zero operation, Bohm had an algebra for implicate order. The implicate order is, to date, an incomplete theory, but important steps in that direction were taken by Bohm, together with Hiley, when they developed their complete theory of ontological quantum mechanics. This is the theory that we are now attempting to use for modelling influenza.

Links:
S-GEM: http://s-gem.se/
Synthetic Populations:
http://dsv.su.se/syntpop

Please contact:
Magnus Boman
SICS and KTH/ICT/SCS
E-mail: mab@sics.se, mab@kth.se

Wah-Sui Almberg
Stockholm University
E-mail: wahsui@dsv.su.se

Epidemic Marketplace: An e-Science Platform for Epidemic Modelling and Analysis

by Fabrício A. B. da Silva, Mário J. Silva and Francisco M. Couto

The Epidemic Marketplace is a distributed data management platform where epidemiological data can be stored, managed and made available to the scientific community. The Epidemic Marketplace is part of a computational framework for organising and distributing data for epidemic modelling and forecasting, dubbed Epiwork. The platform will assist epidemiologists and public health scientists in sharing and exchanging data.

The Epidemic Marketplace is an e-Science platform for collecting, storing, managing and providing epidemic semantically annotated data collections. In recent years, the availability of a huge volume of quantitative social, demographic and behavioural data has spurred an interest in the potential of innovative technologies to improve disease surveillance systems, by providing faster and better geo-referenced outbreak detection capabilities. These capabilities depend on the availability of finely-tuned models, which require accurate and comprehensive data. However, the increasing amount of data introduces the problem of data integration and management. New solutions are needed to ensure that data are correctly stored, managed and made available to the scientific community.

The Epidemic Marketplace is part of a European research effort, the Epiwork project, a four-year project started in 2009. Epiwork supports multidisciplinary research aimed at developing the appropriate framework of tools and knowledge needed for the design of epidemic forecast infrastructures, to be used by epidemiologists and public health scientists. The project is a truly interdisciplinary effort, anchored to the research questions and needs of epidemiology research by the participation of epidemiologists, public health specialists, mathematical biologists and computer scientists. The Epidemic Marketplace is the Epiwork data integration platform, where epidemiological data can be stored, managed and made available to investigators, thus fostering collaboration. The objectives of Epiwork in which the Epidemic Marketplace will play a direct role: (1) the development of large scale, data driven computational models endowed with a high level of realism and aimed at epidemic scenario forecast; (2) the design and implementation of original data-collection schemes motivated by identified modelling needs, such as the collection of real-time disease incidence. This is achieved with the use of innovative Internet and ICT applications; (3) the set up of a computational platform for epidemic research and data sharing that will generate important synergies between research communities and states.

The architectural requirements of the Epidemic Marketplace are directly related to the objectives of the Epiwork project and have been defined according to the feedback from its partners. The main functional requirements of the Epidemic Marketplace are:

- Support the sharing and management of epidemiological data sets. Registered users should be able to upload annotated data sets, and a data set quality assessment mechanism should be available.
- Support the seamless integration of multiple heterogeneous data sources. Users should be able to have a unified view of related data sources. Data should be available from streaming, static and dynamic sources.
- Support the creation of a virtual community for epidemic research. The platform will serve as a forum for discussion that will facilitate the sharing of data between providers and modellers.
- Distributed Architecture. The Epidemic Marketplace should implement a geographically distributed architecture deployed in several sites for improved data access performance, availability and fault-tolerance.
- Support secure access to data. Access to data should be controlled. The marketplace should provide single sign on, distributed federated authorization and multiple access policies, customizable by users.
- Support data analysis and simulation in grid environments. The Epidemic Marketplace will provide data analysis and simulation services in a grid environment.

- Workflow. The platform should provide workflow support for data processing and external service interaction.

The main non-functional requirements that have been identified for the Epidemic Marketplace are:

- Interoperability: The Epidemic Marketplace must interoperate with other software. Its design must take into account the future possibility that systems developed by other researchers worldwide may need to query the Epidemic Marketplace catalogue for access to its datasets.
- Open-source: All software packages, and new modules required for the implementation and deployment of the Epidemic Marketplace should be open source.
- Standards-based: To guarantee software interoperability and the seamless integration of all geographically dispersed sites of the Epidemic Marketplace, the system will be built according to standards defining web services, authentication and metadata.

The Epidemic Marketplace can be defined as a distributed virtual repository, a platform supporting transparent, seamless access to distributed, heterogeneous and redundant resources. It is a virtual repository because data can be stored in systems that are external to the Epidemic Marketplace, and it provides transparent access because several heterogeneities are hidden from its users. The Epidemic Marketplace is composed of a set of interconnected data management nodes geographically distributed, sharing common canonical data models, authorization infrastructure and access interfaces.

As shown in Figure 1, each Epidemic Marketplace node has the following modules:

- Repository: stores epidemic data sets and an epidemic ontology to characterize the semantic information of the data sets.
- Mediator: a collection of web services that will provide access to internal data and external sources, based on a catalogue describing existing epidemic databases through their metadata, using state-of-the-art semantic-web/grid technologies.
- MEDcollector: retrieves information relating to real-time disease incidences from publicly available data sources, such as social networks. After retrieval, the collector groups the incidences by subject and creates data sets to store in the repository.
- Forum: allows users to post comments on integrated data from other modules, fostering collaboration among modellers.

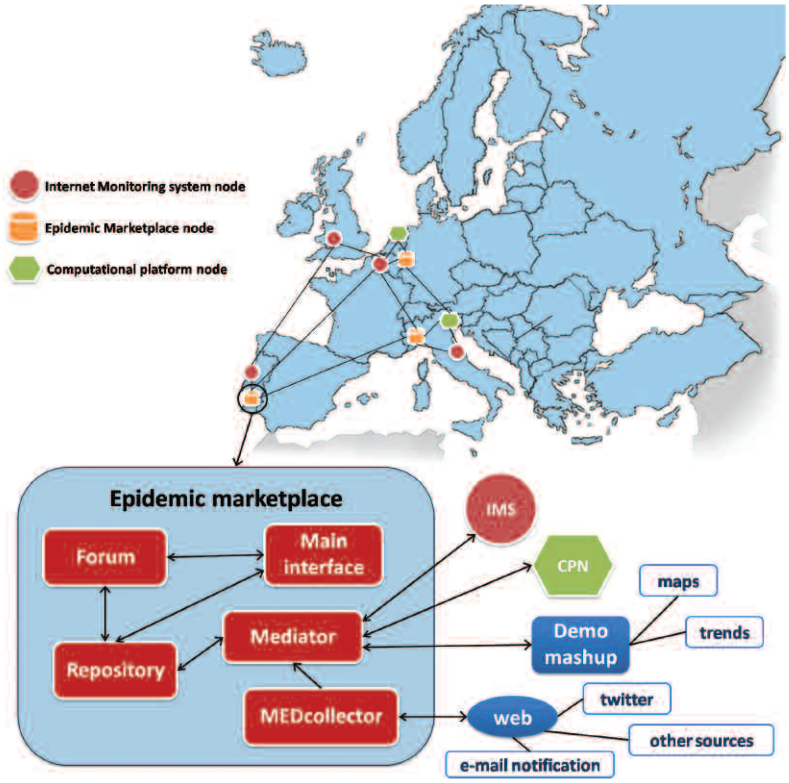


Figure 1: An envisioned deployment of the Epidemic Marketplace distributed among several locations. Currently, only the Lisbon node has been deployed. Each Epidemic Marketplace node is composed of four modules: repository, MEDcollector, forum and mediator. The mediator will be the contact point for other applications, such as Internet Monitoring System nodes (eg Gripenet), and for clients that show data in a graphical and interactive way using geographical maps and trend graphs.

A first prototype version of the Epidemic Marketplace is already in use internally. This prototype implements several of the main features of the outlined architecture such as data management and sharing support and secure access to data, and is currently being populated with epidemic data collections. Several open-source tools and open standards are being used in the Epidemic Marketplace implementation and deployment process, such as the Fedora Commons for the implementation of the main features of the repository. Access control in the platform uses the XACML, LDAP and Shibolleth standards. The front-end of this first prototype is based in Muradora, but the next version will also include a front-end based in the Drupal content management system.

The Epiwork project is funded by the European Commission under the Seventh Framework Programme. The authors want to thank the Epiwork project partners, the students that have been working in the initial implementation of the Epidemic Marketplace (Luis Filipe Lopes, Patricia Sousa, Hugo Ferreira and João Zamite), the CMU-Portugal partnership and FCT (Portuguese research funding agency) for its support.

Links:
Epiwork Project: http://www.epiwork.eu
Epidemic Marketplace:
http://epiwork.di.fc.ul.pt/
Fedora Commons:
http://www.fedora-commons.org/
Shibolleth: http://shibboleth.internet2.edu/
Muradora:
http://www.fedora-commons.org/confluence/display/MURADORA/Muradora
Drupal: http://drupal.org
Gripenet: http://www.gripenet.pt

Please Contact
Mario J. Silva
Universidade de Lisboa, Portugal
E-mail: epiwork@lasige.di.fc.ul.pt

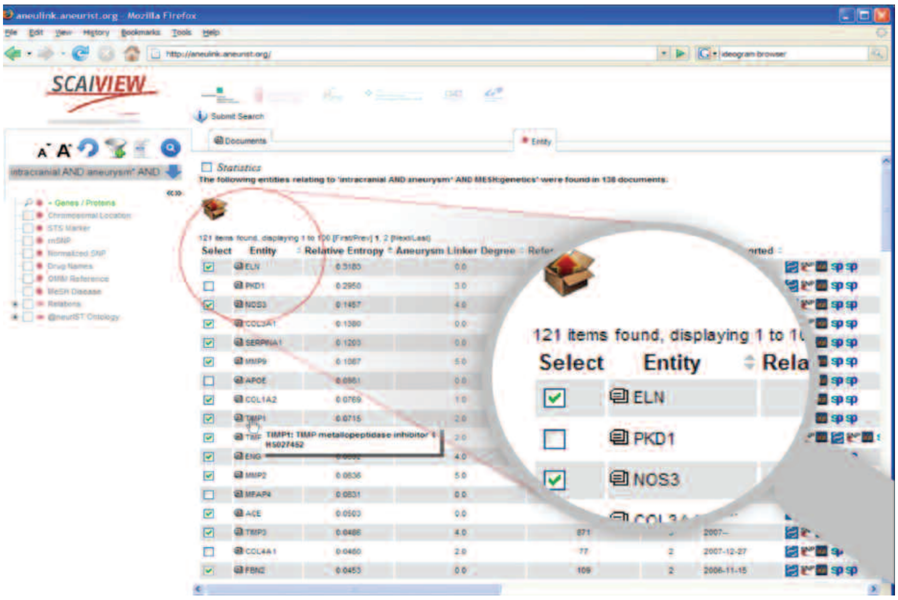
Cross-Project Uptake of Biomedical Text Mining Results for Candidate Gene Searches

by Christoph M. Friedrich, Christian Ebeling and David Manset

From intracranial aneurysms to paediatric diseases - A biomedical text mining service developed in the European IP-project @neurIST has been integrated into a 3D Knowledge Browser developed in the European IP-project Health-e-Child and can be used for candidate gene searches in different diseases.

Most biomedical knowledge can be found in unstructured form in publications. Every day approximately 2000 new citations are added to the PubMed database, a repository of more than 20 million biomedical citations. Even within specific disease areas it is impossible for scientists to stay up to date. Text mining is seen as a solution to this problem. In the European Integrated-project @neurIST (FP6, 1/2006-4/2010), which was concerned with integrated biomedical informatics for the management of intracranial aneurysms, among other data mining solutions, a biomedical text mining system called SCAIView has been developed. This Knowledge Discovery System, depicted in Figure 1, can answer questions like: "Which genes or gene variations, are concerned with intracranial aneurysms?" Or "Which co-morbidities are mentioned together with Alzheimers disease?" or "Which pathways are involved in diabetes?".

The key technologies that enable SCAIView are called semantic search and ontology search. The core of the service is a precompiled index that holds a copy of the PubMed database for fulltext searches and additional information on named entities, which



☐ *Statistics*

The following entities relating to 'intracranial AND aneurysm* AND MESH:genetics' were found in 138 documents.

121 items found, displaying 1 to 100 [First/Prev] 1, 2 [Next/Last]

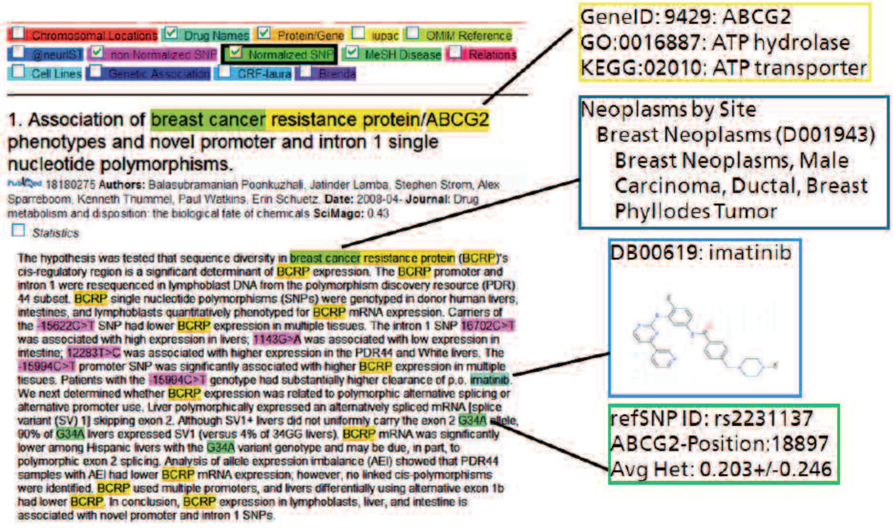
Select	Entity	Relative Entropy	Aneurysm Linker Degree
☑	ELN	0.3180	0.0
☐	PKD1	0.2950	3.0
☑	NOS3	0.1457	4.0
☑	COL3A1	0.1380	0.0
☑	SERPINA1	0.1203	0.0
☑	MMP9	0.1087	5.0
☐	APOE	0.0861	0.0
☑	COL1A2	0.0769	1.0
☑	TIMP1	0.0715	2.0
☑	TIMP		2.0
☑	ENG	[illegible]	4.0
☑	MMP2	0.0636	5.0
☐	MFAP4	0.0631	0.0
☑	ACE	0.0503	0.0
☑	TIMP3	0.0486	4.0
☐	COL4A1	0.0460	2.0
☑	FBN2	0.0453	0.0

TIMP1: TIMP metallopeptidase inhibitor 1 HS027452

121 items found, displaying 1 to 1(

Select	Entity	Rela
☑	ELN	
☐	PKD1	
☑	NOS3	

Figure 1: The search interface of SCAIView.


☐ Chromosomal Locations ☑ Drug Names ☑ Protein/Gene ☐ iupac ☐ OMIM Reference
☐ @neurIST ☑ non Normalized SNP ☑ Normalized SNP ☑ MeSH Disease ☐ Relations
☐ Cell Lines ☐ Genetic Association ☐ CRF-laura ☐ Brenda

1. Association of breast cancer resistance protein/ABCG2 phenotypes and novel promoter and intron 1 single nucleotide polymorphisms.

PubMed 18180275 **Authors:** Balasubramanian Poonkuzhali, Jatinder Lamba, Stephen Strom, Alex Sparreboom, Kenneth Thummel, Paul Watkins, Erin Schuetz. **Date:** 2008-04- **Journal:** Drug metabolism and disposition: the biological fate of chemicals **SciMago:** 0.43

☐ *Statistics*

The hypothesis was tested that sequence diversity in breast cancer resistance protein (BCRP)'s cis-regulatory region is a significant determinant of BCRP expression. The BCRP promoter and intron 1 were resequenced in lymphoblast DNA from the polymorphism discovery resource (PDR) 44 subset. BCRP single nucleotide polymorphisms (SNPs) were genotyped in donor human livers, intestines, and lymphoblasts quantitatively phenotyped for BCRP mRNA expression. Carriers of the -15622C>T SNP had lower BCRP expression in multiple tissues. The intron 1 SNP 16702C>T was associated with high expression in livers; 1143G>A was associated with low expression in intestine; 12283T>C was associated with higher expression in the PDR44 and White livers. The -15994C>T promoter SNP was significantly associated with higher BCRP expression in multiple tissues. Patients with the -15994C>T genotype had substantially higher clearance of p.o. imatinib. We next determined whether BCRP expression was related to polymorphic alternative splicing or alternative promoter use. Liver polymorphically expressed an alternatively spliced mRNA [splice variant (SV) 1] skipping exon 2. Although SV1+ livers did not uniformly carry the exon 2 G34A allele, 90% of G34A livers expressed SV1 (versus 4% of 34GG livers). BCRP mRNA was significantly lower among Hispanic livers with the G34A variant genotype and may be due, in part, to polymorphic exon 2 splicing. Analysis of allele expression imbalance (AEI) showed that PDR44 samples with AEI had lower BCRP mRNA expression; however, no linked cis-polymorphisms were identified. BCRP used multiple promoters, and livers differentially using alternative exon 1b had lower BCRP. In conclusion, BCRP expression in lymphoblasts, liver, and intestine is associated with novel promoter and intron 1 SNPs.

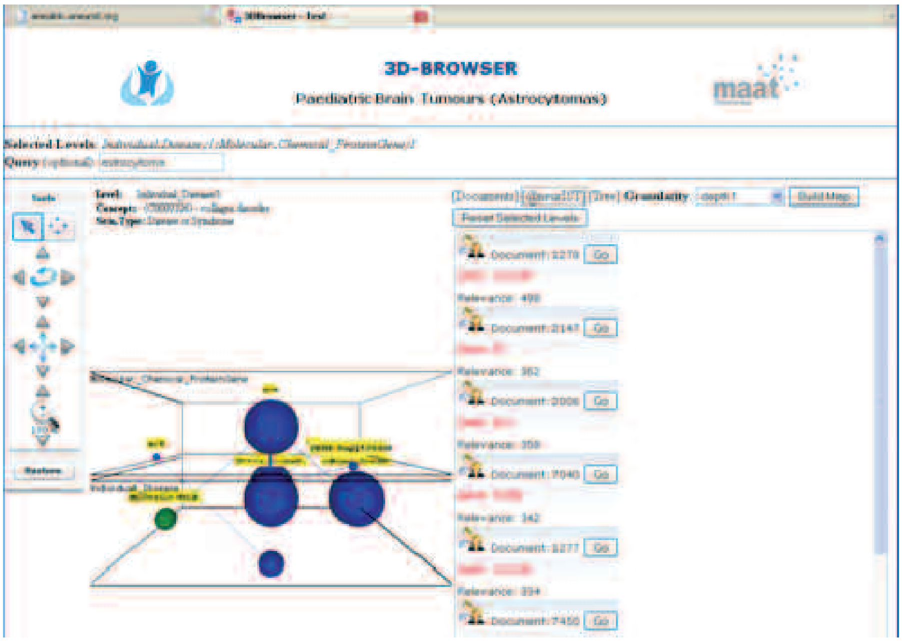
GeneID: 9429: ABCG2
GO:0016887: ATP hydrolase
KEGG:02010: ATP transporter

Neoplasms by Site
Breast Neoplasms (D001943)
Breast Neoplasms, Male
Carcinoma, Ductal, Breast
Phyllodes Tumor

DB00619: imatinib

refSNP ID: rs2231137
ABCG2-Position:18897
Avg Het: 0.203+/-0.246

Figure 2: Highlighting and Enrichment through Named Entity Recognition.



3D-BROWSER

Paediatric Brain Tumours (Astrocytomas)

Selected Levels: Individual_Disease / Molecular_Chemical_ProteinGene/

Query (optional): astrocytoma

[Documents] [Tree] Granularity: depth1 [Build Map]

[Reset Selected Levels]

Document: 2279 [Go]

Relevance: 499

Document: 2147 [Go]

Relevance: 362

Document: 2006 [Go]

Relevance: 358

Document: 7040 [Go]

Relevance: 342

Document: 2277 [Go]

Relevance: 334

Document: 7450 [Go]

Figure 3: SCAIView searches integrated into the 3D Knowledge Browser.

occur within the text. A named entity is a semantic entity such as a drug name or gene name which is found by Named Entity Recognition (NER). In the NER process all synonyms of a semantic entity are used for an approximative search and, where possible, the found entity is mapped to a unique database identifier. This process is necessary as some genes or diseases have up to 250 different name variants that occur in text. The information is enriched with data from biomedical databases and ontologies (see Figure 2), so that finally a query for all genes which are on a certain pathway and are co-mentioned with a disease can be submitted.

Most of the named entities in SCAIView are searched with dictionaries but some entities cannot be enumerated beforehand. For these cases rule- and machine-learning based entity recognisers have been implemented. The recognition process involves the machine learning from examples or rules provided by humans. One example for this might be: gene variations like: "... polymorphisms in TIMP-3 (249T>C, 261C>T)". In this example the result of the search is an identifier to a polymorphism in a database or the identifier of a probe on a Microarray for Genome wide association studies.

Evaluation of Retrieval Performance

Every Search engine is only as good as the results are valid, novel, relevant and useful for the user. The evaluation of SCAIView was based on the question for candidate genes of a disease. The gold standard for "Intracranial Aneurysms" was an expert review on the topic and a Cochrane Report that has been produced in the course of the @neurIST project. We found all candidate genes with our search and they have been distributed among the top ranking hits. Additionally we found other novel candidate genes, which have not been mentioned in the reviews. Other successful evaluations have been conducted for Alzheimer's disease, schizophrenia and Parkinson's disease.

Integration into the 3D Knowledge Browser

The integrated European project Health-e-Child (HeC) (FP6, 2/2006-4/2010) has developed a 3D Knowledge Browser, which is specially suited for multi-scale and multi-level searches in biomedical Knowledge Sources. It lacked a search engine for discovering links between HeC data and gene-based information published in external sources. Therefore we conducted a joint research project and integrated the SCAIView results via Webservices into the 3D Knowledge Browser. Now entity searches, for instance for candidate genes, can be displayed next to database searches like a retrieval from SwissProt. In Figure 3 a screenshot of the resulting interface is given. In the Health-e-Child project, this has been used to search for candidate genes in paediatric heart diseases, inflammatory diseases, and brain tumours.

Links:
http://www.aneurist.org
http://www.health-e-child.org

Please contact:

Christoph M. Friedrich
Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Germany
Tel: +49 2241 142502
E-mail: friedrich@scai.fraunhofer.de

Ontologies and Vector-Borne Diseases: New Tools for Old Illnesses

by Pantelis Topalis, Emmanuel Dialynas and Christos Louis

Vector-borne diseases are illnesses that are characterized by the transmission of infectious agents between humans via the bites of arthropod vectors, most prominently mosquitoes. It is hoped that recent scientific advances, especially in the areas of high throughput biological research and bioinformatics will assist in alleviating the global burden caused by these diseases.

Vector-borne diseases have been central in shaping human history by playing a crucial role in events such as the siege of Syracuse, and the death of prominent leaders, such as Alexander the Great. More than one million deaths annually, resulting from hundreds of millions of cases, can be attributed to malaria alone, which, in spite of considerable efforts, still represents one of the menaces faced the poorest populations in the tropics. The fight against this disease is hampered by increasing resistance exhibited by both Plasmodium, the malaria parasite, against drugs, and by vector anopheline mosquitoes against insecticides used for public health purposes. Moreover, the lack of vaccines and the less-than-perfect infrastructure in the countries most affected have not helped improve the situation. Nevertheless, the advent and expansion of molecular biology and genomics over the last decade has provided new knowledge on all three "partners" of the vector-borne diseases (human, pathogen and vector). Combined with the recent research boom in bioinformatics, this knowledge could potentially be put to use for the development of specific tools that might help control the diseases. The potential eradication of malaria, a notion that was abandoned 40 years ago, is now being discussed again as a possibility.

In the frame of this enhanced effort to fight vector-borne diseases our group has joined several international networks that share the same goals, most notably the BioMalPar and its successor EVIMalaR, both Networks of Excellence in the frame of the FP6 and FP7 programmes of the European Union, and the VectorBase (VB; PI: Frank Collins, University of Notre Dame) project funded by the National Institute of Allergy and Infectious Diseases of the USA. While the European networks concentrate on "bench" biological research, the latter is a dedicated bioinformatics project that is responsible for the development of a comprehensive database of genetic, genomic and other bio-data that are pertinent to disease vectors. The key contents of VB include the genomic sequence information on several vectors, first of which is Anopheles gambiae, the most important malaria vector in Africa. Additional information hosted by VB includes, among others, data on gene expression, specific bioinformatics tools and a section on insecticide resistance, IRBase (see below), which was constructed by the Cretan team. In addition to these, VB is also the home of

Mosquitoes from an Anopheles gambiae laboratory colony, "resting" in the authors' laboratory.

a set of ontologies, all in OBO format, that describe different aspects of vector-borne diseases, which are the main focus of interest of the IMBB group in the frame of VB. These can be used to promote interoperability between databases and facilitate complex queries in different but related datasets.

The first two ontologies to be built were those relating to the anatomy of mosquitoes (TGMA) and of ticks (TADS), both species groups that include several major vectors. These ontologies were primarily constructed in order to help the annotation of biological experiments dealing with these species. The next ontology developed, MIRO, covers the domain of insecticide resistance, a feature that is of extreme importance for the control of both agricultural pests and disease vectors. In the frame of VB, MIRO drives IRBase, the corresponding database, which collects data supplied by both individuals and organizations such as the World Health Organization, with the aim of making them freely available to the worldwide community of field entomologists and public health workers. Finally, a large malaria ontology (IDOMAL) is also part of the VB effort. This, developed in the frame of the international consortium IDO that works on the construction of a top-level infectious disease ontology, is the first step in an ambitious plan to construct ontologies for all major vector-borne diseases such as African and American trypanosomiasis, lymphatic filariasis, yellow fever and more.

The common thread between all ontologies that are being constructed de novo is that all follow the rules set by the OBO Foundry, a loose collaboration between teams developing open biological ontologies. Moreover, we build our ontologies on the format laid down by BFO, the basic formal ontology. Both of these conditions guarantee the maximum of interoperability and orthogonality, thus enhancing their usability.

In addition to the "classical" potential usage of our existing and new/forthcoming ontologies in driving databases, their availability under the criteria briefly described make it possible to use them in more practical ways, for example, in driving epidemiological decision support systems. Given the state of public health in most third world countries, we believe that a functional improvement based on bioinformatics tools such as the specialized ontologies on which we work is a practical way to help the poverty-stricken populations of the tropics.

Links:
http://www.vectorbase.org
http://anobase.vectorbase.org/ontologies

Please contact:
Christos Louis
IMBB-FORTH, Greece
Tel: +30 281 039 1119?
E-mail: louis@imbb.forth.gr

Unraveling Hypertrophic Cardiomyopathy Variability

by Catia M. Machado, Francisco Couto, Alexandra R. Fernandes, Susana Santos, Nuno Cardim and Ana T. Freitas

Hypertrophic cardiomyopathy is a disease characterized by a high genetic heterogeneity with variable clinical presentation, thus rendering the possibility of personalized treatments highly desirable. This can be achieved through the integration of genomic and clinical data with Semantic Web technologies, combined with the identification of correlations between the data elements using data mining techniques.

Hypertrophic cardiomyopathy (HCM) is an autosomal dominant genetic disease that may afflict as many as one in 500 individuals, being the most frequent cause of sudden death among apparently healthy young people and athletes. It is an important risk factor for heart failure disability at any age and is characterized by a hypertrophied, non-dilated left ventricle, myocyte disarray and interstitial fibrosis.

Since the disease is characterized by a variable clinical presentation and onset, its clinical diagnosis is difficult prior to the development of severe or even fatal symptoms. Therefore, its early diagnosis is extremely important.

In terms of genetic manifestation, more than 640 mutations in more than 20 genes have been associated with HCM phenotype. The detection of these mutations in the genome of the patient can greatly improve the disease diagnosis. However, genetic diagnosis using dideoxi-sequencing techniques is hampered by these high numbers of genes/mutations and by the fact that some patients present the clinical manifestations of the disease but none of the mutations is found under the current genetic testing (which indicates that even more mutations and/or genes might be involved). Moreover, HCM's severity may not be the same for two individuals, even if direct relatives, since the presence of a given mutation can have a benign pattern in one individual and result in sudden cardiac death in another.

These complicating factors indicate that a joint strategy among clinicians, biologists and bioinformaticians may be advantageous. Biologists need to identify and characterize new mutations with high throughput techniques, thus enabling clinicians to count on this information, when articulated with the clinical findings, to provide the best possible treatment for each individual patient. Bioinformaticians have the tools needed to expeditiously integrate and analyse the data, thus providing the necessary articulation between the different types of data and the identification of patterns that might shed light into the disease variability.

In more concrete terms, the data elements that need to be integrated corre-

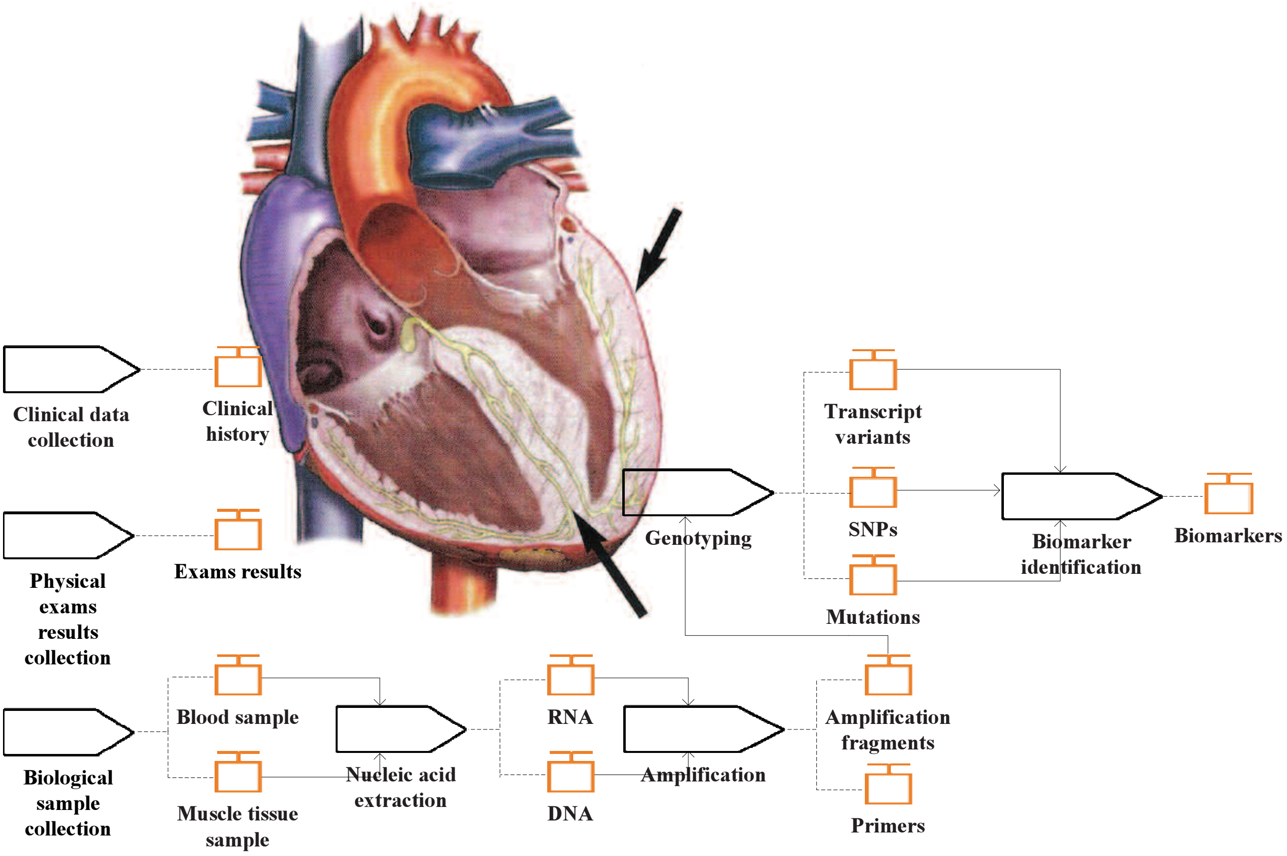


Figure 1: Model of a human heart showing the thickening of the left ventricular wall, and the HCM characterization workflow, comprising all modelling activities.

spond to the presence/absence of each mutation in the genome of the patients (genotype data) and to the clinical elements upon which the clinicians rely to provide a diagnose (phenotype data). The latter normally include the results from physical examinations (eg electrocardiogram, echocardiogram), as well as the clinical history of the individual (eg age at diagnosis, sudden deaths in the family).

The data integration procedure in this context is a very demanding task considering that genotype and phenotype data correspond to transversal domains, normally stored under heterogeneous formats and on different physical locations. The approach proposed here is based on Semantic Web technologies, previously identified as suitable for this type of task since they make it possible to integrate, share and reuse data in an application- and domain-independent manner.

Upon completion of the data integration phase, the data will be analysed with data mining techniques in order to infer genotype-phenotype correlations, or, more specifically, to develop models for the association between the presence of certain mutations and the resulting physical traits. Although a large number of studies have been conducted on the linkage between specific mutations and the risk of specific illnesses (eg cystic fibrosis), models for the more general case of genotype to phenotype association in the presence of high disease complexity, both genetic and clinical, remains largely unexplored. Supervised machine learning techniques, such as decision trees and support vector machines, offer the potential to identify more complex relationships than those identified using simple correlation analysis, the standard practice in genotype-phenotype association models. Standard statistical analysis may identify correlations between one or a small set of specific mutations, but in more complex cases these correlations will not be significant enough to lead to concrete diagnosis methods. The models obtained using data mining techniques are expected to be of great interest both in terms of their predictive ability and their practical usability for doctors.

The ultimate goal of this work is to develop a clinical characterization system that, upon introduction of a new patient's data, will provide an indication of whether the patient is suffering from HCM based on the existing model.

Currently in the integration stage, this work is being conducted in the LASIGE group at the Departamento de Informática of the Universidade de Lisboa and in the KDBIO group at the Instituto de Engenharia de Sistemas e Computadores of the Instituto Superior Técnico (both in Lisbon, Portugal). The data used are provided by the six other Portuguese institutions listed below:

- Phenotype data – Hospitais da Universidade de Coimbra (Coimbra), Centro de Cardiologia da Universidade de Lisboa, Hospital da Luz and Hospital de Sta. Cruz (all in Lisbon)
- Genotype data – Centro de Química Estrutural of the Instituto Superior Técnico of the Universidade Técnica de Lisboa and the Faculdade de Farmácia of the Universidade de Lisboa (both in Lisbon).

Link:
http://kdbio.inesc-id.pt

Please contact:
Ana Teresa Freitas
KDBIO research group,
INESC-ID/IST Lisbon, Portugal
Tel: +351 213100394
E-mail: atf@inesc-id.pt

Continuous Evolutionary Automated Testing for the Future Internet

by Tanja E. J. Vos

The Future Internet will be a complex interconnection of services, applications, content and media, on which our society will become increasingly dependent for critical activities such as social services, learning, finance, business, as well as entertainment. This presents challenging problems during testing; challenges that simply cannot be avoided since testing is a vital part of the quality assurance process. The European funded project FITTEST (Future Internet Testing) aims to attack the problems of testing the Future Internet with Evolutionary Search Based Testing.

The Future Internet (FI) will be a complex interconnection of services, applications, contents and media, possibly augmented with semantic information, and based on technologies that offer a rich user experience, extending and improving current hyperlink-based navigation. Our society will be increasingly dependent on services built on top of the FI, for critical activities such as public utilities, social services, government, learning, finance, business, as well as entertainment. As a consequence, the applications running on top of the FI will have to satisfy strict and demanding quality and dependability standards.

FI applications will be characterized by an extreme level of dynamism. Most decisions made at design or deploy time are deferred to the execution time, when the application takes advantage of monitoring (self-observation, as well as data collection from the environment and logging of the interactions) to adapt itself to a changed usage context. The realization of this vision involves a number of technologies, including:

- Observational reflection and monitoring, to gather data necessary for run-time decisions.
- Dynamic discovery/composition of services, hot component loading and update.
- Structural reflection, for self-adaptation.
- High configurability and context awareness.
- Composability into large systems of systems.

While offering a major improvement over the currently available Web experience, complexity of the technologies involved in FI applications makes testing them extremely challenging. Some of the challenges are described in Table 1. The European project FITTEST (ICT-257574, September 2010-2013) aims at addressing the testing challenges in Table 1 by developing and evaluating an integrated environment for continuous evolutionary automated testing, which can monitor the FI application under test and adapt to the dynamic changes observed. Today's traditional testing usually ends with the release of a planned product version with testing being discontinued after application delivery. All testware (test cases as well as the underlying behavioural model) are constructed and executed before delivering the software. These testwares are fixed, because the System Under Test (SUT)has a fixed set of features and functionalities, since its behaviour is determined before its release. If the released product needs to be updated because of changed user requirements or severe bugs that need to be repaired, a new development cycle will start and regression testing needs to be done in order to ensure that the previous functionalities still work with the new changes. The fixed testwares, designed during post-release testing, need to be adapted manually in order to cope with the changed requirements, functionalities, user interfaces and/or work-flows of the system (See Figure 1 for a graphical representation of traditional testing).

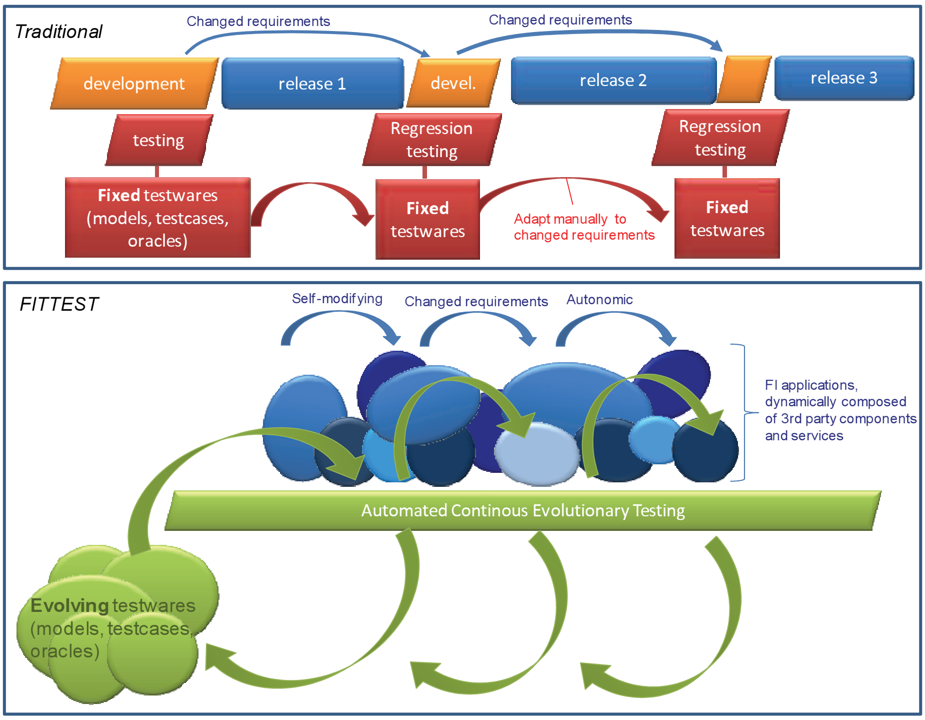

Traditional
Changed requirements
Changed requirements
development
release 1
devel.
release 2
release 3
testing
Regression testing
Regression testing
Fixed testwares (models, testcases, oracles)
Fixed testwares
Adapt manually to changed requirements
Fixed testwares
FITTEST
Self-modifying
Changed requirements
Autonomic
FI applications, dynamically composed of 3rd party components and services
Automated Continous Evolutionary Testing
Evolving testwares (models, testcases, oracles)

Figure 1: Traditional testing versus FITTEST testing.

1	Dynamism, self-modification and autonomic behaviour	FI applications are highly autonomous; their correct behaviour for testing cannot be specified and modelled precisely at design-time.
2	Low observability	FI applications are composed of an increasing number of third-party components and services, accessed as a black box, which are hard to test.
3	Asynchronous interactions	FI applications are highly asynchronous and hence harder to test. Each client submits multiple requests asynchronously; multiple clients run in parallel; server-side computations are distributed over the network and concurrent.
4	Time and load dependent behaviour	For FI applications, timing and load conditions make it hard to reproduce errors during debugging.
5	Huge feature configuration space	FI applications are highly customizable and self-configuring, having a huge number of configurable features, such as context, user, environment-dependent configurable parameters that need to be tested.
6	Ultra-large scale	FI applications are often systems of systems; traditional testing adequacy criteria, like coverage, cannot be applied.

Table 1: Future Internet Challenges.

FI testing will require continuous post-release testing since the application under test does not remain fixed after its release. Services and components could be dynamically added by customers and the intended use could change significantly. Therefore, testing has to be performed continuously after deployment to the customer.

Within FI application, services and components could be dynamically added by customers, and the intended use could change significantly. Since FI applications do not remain fixed after its release, they will require ongoing, continuous testing even after deployment to the customer.

The FITTEST testing environment will integrate, adapt and automate various techniques for continuous FI testing (eg dynamic model inference, model-based test case derivation, log-based diagnosis, oracle learning, classification trees and combinatorial testing, concurrent testing, regression testing, etc.). To make it possible for the above mentioned techniques to deal with the huge search space associated with FI testing, evolutionary search based testing will be used. Search Based Software Testing (SBST) is the FITTEST basis for test case generation, since it can be adopted even in the presence of dynamism and partial observability that characterize FI applications. The key ingredients required by search algorithms are:

- an objective function, that measures the degree to which the current solution (eg test case.) achieves the testing goal (eg, coverage, concurrency problems, etc.)
- operators that modify the current solution, producing new candidate solutions to be evaluated in next iterations of the algorithm.

Such ingredients can usually be obtained even in the presence of high dynamism and low observability. Hence, it is expected that most techniques that will be developed to address the project's objectives will take advantage of a search based approach. For example, model-inference for model-based testing requires execution scenarios to limit under-approximation. We can generate them through SBST, with the objective of maximizing the portion of equivalent states explored. Oracle learning needs training data that are used to infer candidate specifications for the system under test. We can select the appropriate executions through SBST, using a different objective function (eg, maximizing the level of confidence and support for each inferred property). We will generate classification-trees from models and use SBST to generate test cases for classification-trees. Input data generation and feasibility check are clearly good candidates to resort to search based algorithms, with an objective function that quantifies the level to which a test case gets close to satisfying a path of interest. In concurrency testing, test cases that produce critical execution conditions are given a higher objective value defined for this purpose. Coverage targets can be used to define the objective function to be used in coverage and regression testing of ultra-large scale systems. SBST represents the unifying conceptual framework for the various testing techniques that will be adapted for FI testing in the project. Implementations of such techniques will be integrated into a unified environment for continuous, automated testing of FI applications. Quantification of the actually achieved project objectives will be obtained by executing a number of case studies, designed in accordance with the best practices of empirical software engineering.

The FITTEST project is composed of partners from Spain (Universidad Politecnica de Valencia), UK (University College London), Germany (Berner & Mattner), Israel (IBM), Italy (Fondazione Bruno Kessler), The Netherlands (Utrecht University), France (Softeam) and Finland (Sulake).

Link:
http://www.pros.upv.es/fittest/

Please contact:

Tanja E.J. Vos,
Universidad Politécnica de Valencia / SpaRCIM
Tel: +34 690 917 971
E-mail: tvos@pros.upv.es

Robust Image Analysis in the Evaluation of Gene Expression Studies

by Jan Kalina

Gene expression data are typically analysed by standard automated procedures that tend to be vulnerable to outlying values. In a project carried out by the Centre of Biomedical Informatics of the Ministry of Education, Youth and Sports of the Czech Republic, we use alternative approaches, based on robust statistical methods, to measure differential gene expression in cardiovascular patients. Our results are applicable to personalized medicine.

Recent research in the area of molecular bioinformatics and genetics, conducted by the Centre of Biomedical Informatics aims to find the optimal set of genes for diagnostics and prognosis of cardiovascular diseases. Since 2006 we have been using whole-genome beadchip microarray technology to measure gene expression. The sample of peripheral blood is taken from each patient, the ribonucleic acid (RNA) is isolated and applied on the microarray. This technology allows to measure the gene expression as the gene activity leading to synthesis of proteins and consequent biological processes. The Municipal Hospital in Èáslav takes blood samples from two groups of patients: one group with acute myocardial infarct (AMI) or cerebrovascular accident (CVA) (as examples of ischemic diseases); and a control group (patients hospitalized with a different cause without a manifested ischemic disease). This whole-genome analysis examines the entire set of human genes with different microbeads corresponding to different genes randomly distributed on the surface of the microarray, which contains 12 separate physical strips (Figure 1) for samples from different patients.

The standard approach in the preprocessing of the data tends to be vulnerable to outlying observations. The raw data are scanned images with a high fluorescence intensity corresponding to highly expressed genes. To compute the bead-level data for particular microbeads a cascade of transformations is computed, including the local estimation of background, image sharpening and smoothing by averaging, estimating foreground, background correction, data normalization and outlier deletion. The initial steps are strongly influenced by local noise in the neighbourhood of particular microbeads and the resulting biased values are passed on to the next steps of the analysis.The outlier deletion is computed only at the end of the procedure. Therefore the differential expression analysis is sensitive to random or systematic errors in the original data.

As an alternative to the processing of the scanned images with gene expression measurements we propose a more robust approach which involves searching for systematic artifacts in the data. The methods are based on robust statistics applied to image analysis which enables outliers to be deleted at each step of the procedure. The method is also robust to specific properties of the neighbourhood of particular pixels. A careful normalization of bead-level data is computed only after deleting the outlying values. At the same time these methods allow fast computing and are computationally feasible. We propose that standard software for analysing gene expression data be modified to incorporate this new approach.

Figure 1: Beadchip for genome-wide expression analysis containing twelve strips.

The outcome of this unique project is the ability to demonstrate which genes are more strongly expressed in patients with acute myocardial infarct or cerebrovascular accident compared to controls. The significance of differential expression of particular genes is acquired by means of statistical hypothesis testing. Clinical and biochemical data recorded for each patient contribute to our understanding of the genetic predisposition to cardiovascular disease. The study of gene expression profiling has allowed the Centre of Biomedical Informatics to patent an oligonucleotide microarray as the main result of the whole study, directly applicable to disease diagnosis, prognosis, prediction and treatment. This technology containing an optimal set of genes provides an invaluable contribution towards the development of a personalized and predictive medical care, in keeping with the new paradigm of data-driven evidence-based medicine.

We believe that the development and use of robust analysis methods is becoming increasingly important in the area of bioinformatics. Next generation (Next-Gen) sequencing, the new low noise approach to genetic analysis, is currently undergoing rapid development, producing huge data sets. Robust statistical methods are therefore becoming ever more crucial for fast and reliable data analysis. It is vital, when designing real-time image analysis systems for Next-Gen technologies, that such methods are adaptive and tailor-made for the particular task, allowing the user to tune parameters. The future of molecular bioinformatics will therefore require more precise and well-considered robust image analysis.

Links:
http://www.euromise.org/cbi/cbi.html
http://www.euromise.org/homepage/department.html

Please contact:
Jan Kalina, Centre of Biomedical Informatics, Institute of Computer Science, Academy of Sciences of the Czech Republic / CRCIM
Tel.: +420 266053099
E-mail: kalina@euromise.cz

Autonomous Production of Sport Video Summaries

by Christophe De Vleeschouwer

Video production cost reduction is an ongoing challenge. The FP7 'APIDIS' project (Autonomous Production of Images based on Distributed and Intelligent Sensing) uses computer vision and artificial intelligence to propose a solution to this problem. Distributed analysis and interpretation of a sporting event are used to determine what to show or not to show from a multi-camera stream. This process involves automatic scene analysis, camera viewpoint selection, and generation of summaries through automatic organization of stories. APIDIS provides practical solutions to a wide range of applications, such as personalized access to local sport events on the web or automatic log in of annotations.

Today, individuals and organizations want to access dedicated contents through a personalized service that is able to provide what they are interested in, at a time convenient to them, and through the distribution channel of their choice. To address such demands, cost-effective and autonomous generation of sports team video contents from multi-sensored data becomes essential, ie to generate on-demand football or basket ball match summaries.

APIDIS is a research consortium developing the automatic extraction of intelligent contents from a network of cameras and microphones distributed around a sports ground. Here, intelligence refers to the identification of salient segments within the audiovisual content, using distributed scene analysis algorithms. In a second step, that knowledge is exploited to automate the production and personalize the summary of video contents.

Specifically, salient segments in the raw video content are identified based on player movement analysis and scoreboard monitoring. Player detection and tracking methods rely on the fusion of the foreground likelihood information computed in each camera view. This overcomes the traditional hurdles associated with single view analysis, such as occlusions, shadows and changing illumination. Scoreboard monitoring provides valuable additional input which assists in identifying the main focal points of the game.

In order to produce semantically meaningful and perceptually comfortable video summaries, based on the extraction of sub-images from the raw content, the APIDIS framework introduces three fundamental concepts: "completeness", "smoothness" and "fineness". These concepts are defined below..Scene analysis algorithms then select temporal segments and corresponding viewpoints in the edited summary as two independent optimization problems according to individual user preferences (eg in terms of preferred player or video access resolution). To illustrate these techniques, we consider a basket-ball game case study, which incorporates some of the latest research outputs of the FP7 APIDIS research project.

Multi-view player detection, recognition, and tracking

The problem of tracking multiple people in cluttered scenes has been extensively studied, mainly because it is common to numerous applications, ranging from sport event reporting to surveillance in public spaces. A typical problem is that all players in a sports team have a very similar appearance. For this reason, we integrate the information provided by multiple views, and focus on a particular subset of methods that do not use color models or shape cues of individual people, but instead rely on the distinction of foreground from background in each individual camera view to infer the ground plane locations occupied by people.

Figure 1 summarizes our proposed method. Once players and referee have been localized, the system has to decide who's who. To achieve this, histogram analysis is performed on the expected body area of each detected person. Histogram peak extraction enables assignment of a team label to each detected player (see bounding boxes around the red and blue teams). Further segmentation and analysis of the

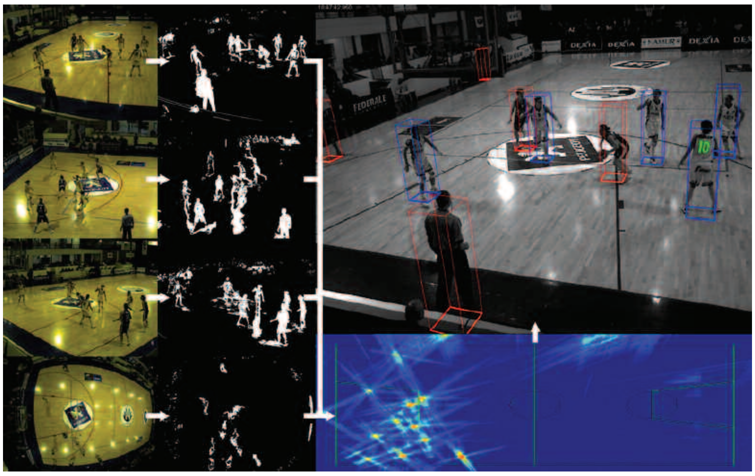


Figure 1: On the left, the foreground likelihoods are extracted from each camera. They are projected to define a ground occupancy map (bottom right in blue) used for player detection and tracking, which in turns supports camera selection.

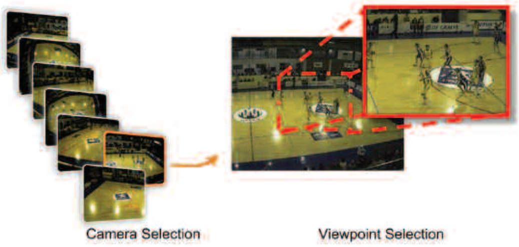


Figure 2: Camera selection and field of view selection.

regions composing the expected body area permits detection and recognition of the digits printed on the players' shirts when they face the camera.

Event recognition

The main events occurring during a basketball game include field goals, violations, fouls, balls out-of-bounds and free-throws. All these events correspond to 'clock-events', ie they cause a stop, start or re-initialization of the 24" clock, and they can occur in periods during which the clock is stopped.

An event tree is built on the basis of the clock and scoreboard information. When needed, this information is completed by visual hints, typically provided as outcomes of the player (and ball) tracking algorithms. For instance, an analysis of the trajectories of the players can assist in decision-making after a start of the 24'' clock following a 'rebound after free-throw' or a 'throw-in' event.

Autonomous production of personalized video summaries
To produce condensed video reports of a sporting event, the system selects the temporal segments corresponding to actions that are worth being included in the summary based on three factors:

- Completeness stands for both the integrity of view rendering in camera/viewpoint selection, and that of story-telling in summary.
- Smoothness refers to the graceful displacement of the virtual camera viewpoint, and to the continuous story-telling resulting from the selection of contiguous temporal segments. Preserving smoothness is important to avoid distracting the viewer from the story with abrupt changes of viewpoint.
- Fineness refers to the amount of detail provided about the rendered action. Spatially, it favours close views. Temporally, it implies redundant story-telling, including replays. Increasing the fineness of a video does not only improve the viewing experience, but is also essential in guiding the emotional involvement of viewers through the use of close-up shots.

The ability to personalize the viewing experience through the application of different parameters for each end user was appreciated during the first subjective tests. The tests also revealed that viewers generally prefer the viewpoints selected by the automatic system than those selected by a human producer. This is, no doubt, partly explained by the severe load imposed on the human operator with an increasing number of cameras..

The author thanks the European Commission and the Walloon Region for funding part of this work through the FP7 APIDIS and WIST2 WALCOMO projects, respectively.

Link:
http://apidis.org/

Please contact:

Christophe De Vleeschouwer
Université Catholique de Louvain (UCL), Belgium
Tel: +32 1047 2543
E-mail: christophe.devleeschouwer@uclouvain.be

IOLanes: Advancing the Scalability and Performance of I/O Subsystems in Multicore Platforms

by Angelos Bilas

Data storage technology today faces many challenges, including performance inefficiencies, inadequate dependability and integrity guarantees, limited scalability, loss of confidentiality, poor resource sharing, and increased ownership and management costs. Given the importance of both direct-attached and networked storage systems for modern applications, it becomes imperative to address these issues. Multicore CPUs offer the promise of dealing with many of the underlying limitations of today's I/O architectures. However, this requires careful consideration of architectural and systems issues and complex interactions in the I/O stack, all the way from the application to the disk. "IOLanes" (Advancing the Scalability and Performance of I/O Subsystems in Multicore Platforms) is new a EC-funded project led by FORTH-ICS that addresses these issues.

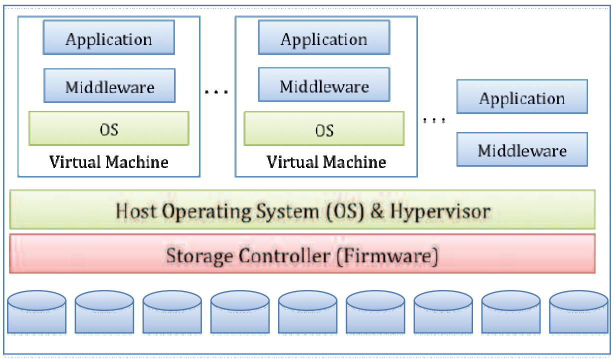


Figure 1: Layers in the I/O path of existing systems.

IOLanes targets three major challenges: (i) dealing with performance and scalability issues of the I/O stack on multicore architectures, (ii) addressing I/O performance and dynamic resource management issues in virtualised, single-host environments, and (iii) examining on-loading and off-loading tradeoffs for advanced functions that are becoming essential in modern storage systems, eg compression, protection, encryption, error correction.

Concept

Our lives are becoming increasingly dependent on electronic information that is processed and stored in computer systems. Individuals, businesses, and organizations cannot survive in today's competitive economy without the use of digital information stored in electronic data storage systems. Data storage is perhaps the most critical and valuable component of today's computing infrastructures. However, critical and valuable as they may be, existing data storage systems today fall short of applications and users' needs in four main respects:

1. *Performance:* The storage system continues to be the performance bottleneck in most computer systems, due to the processor-disk performance gap. The recent advent of multicore processors and the steadily increasing number of cores per chip has resulted in a decrease in the effective storage bandwidth available per core.
2. *Dependability and Data Integrity:* For the last three decades there has been a dramatic increase of storage density in magnetic and optical media, which has led to significantly lower cost per capacity unit (€/GB). Furthermore, organizations and individuals have taken advantage of this trend by creating and storing ever-increasing amounts of digital data. Storing and handling these unprecedented amounts of data has led to increased failures; and failures in the storage subsystem can be unnerving.
3. *Data Confidentiality:* The sensitivity of digital data for individuals and organisations requires careful protection of all digital assets. Data should always be stored in eg an encrypted form to prevent leakage to anyone who can obtain access to the operating systems or physical access to the storage device(s). However, such solutions remain, to date, exotic, are not used in most systems, and even when they can be used, they are extremely costly in terms of processing cycles and power.
4. *Resource and content consolidation:* In order to minimise both cost and complexity in storage systems there is an increasing need to consolidate both (a) storage resources and (b) data and content. The first, results in better use of available resources, mainly via sharing among multiple applications and amortizing peak needs. The second simplifies as much as possible, management procedures and mechanisms that incur important costs to modern storage infrastructures.

A basic enabler for building the future storage systems is to take advantage in the I/O stack of the performance potential of multicore CPUs and at the same time deal with their shortcomings.

Challenges

Data storage processing is typically structured as an "I/O path" that takes data from the application to the storage device itself when we need to store it or retrieve it. The I/O path in existing virtualized systems traverses through several layers of the system: the application and middleware layer, the virtual machine (VM) layer, the host operating system (OS) layer and the embedded storage controller firmware layer. In non-virtualized systems, there are no virtual machines and applications are executed on the host OS. This generic layered structure is shown in Figure 1.

IOLanes aims to analyse and address challenges throughout the I/O path. It will analyse and address inefficiencies associated with these layers on multicore CPUs, by designing an I/O stack that minimizes unnecessary overheads and scales with the number of cores. Since storage systems are perhaps the most critical component of modern computing infrastructures, the proposed work will benefit many I/O-intensive applications that support activities of businesses, organizations, and individuals alike. Our work will result in systems that are able to perform multi-GBytes/s I/O in virtualized environments, supporting advanced functionality, and allowing scalability with the number of cores, as multicore architectures evolve over time.

The project is carried out by ICS-FORTH; Univ. Politécnica Madrid, Barcelona Supercomputing Center. Spain; IBM Israel - Science and Technology LTD, Israel; Intel Performance Learning Solutions Ltd., Ireland, and Neurocom S.A. , Greece. The IOLanes project is part of the portfolio of the Embedded Systems Unit – G3, Directorate General Information Society (http://cordis.europa.eu/ist/ embedded).

Link: http://www.iolanes.eu

Please contact:
Angelos Bilas
FORTH-ICS, Greece
E-mail: bilas@ics.forth.gr

InfoGuard: A Process-Centric Rule-Based Approach for Managing Information Quality

by Thanh Thoa Pham Thi, Juan Yao, Fakir Hossain and Markus Helfert

High quality data helps the data owner to save costs, to make better decisions and to improve customer service and thus, information quality impacts directly on businesses. Many tools for data parsing and standardization, data cleansing, or data auditing have been developed and commercialized. However identifying and eliminating root causes of poor IQ along the information life cycle and within information systems is still challenging. Addressing this problem we developed an innovative approach and tool which helps identifying root causes of poor information quality.

In a recent survey by Gartner information quality (IQ) related costs are estimated as much as millions of dollar (SearchDataManagement.com, "Poor data quality costing companies millions of dollars annually", 2009.). Meanwhile Thomson Reuters and Lepus survey in 2010 (http://thomson-reuters.com/content/press_room/tf/tf_gen_business/2010_03_02_lepus_survey) revealed "77% of participants intend to increase spending on projects that address data quality and consistency issues" which are "key to risk management and transparency in the financial crisis and the market rebuild".

In order to improve and maintain IQ, effective IQ management is necessary, which ensures that the raw material (or data) an organisation creates and collects is as accurate, complete and consistent as possible. Furthermore the information product which is manufactured from raw data by transforming, assembling processes must also be accurate, com-

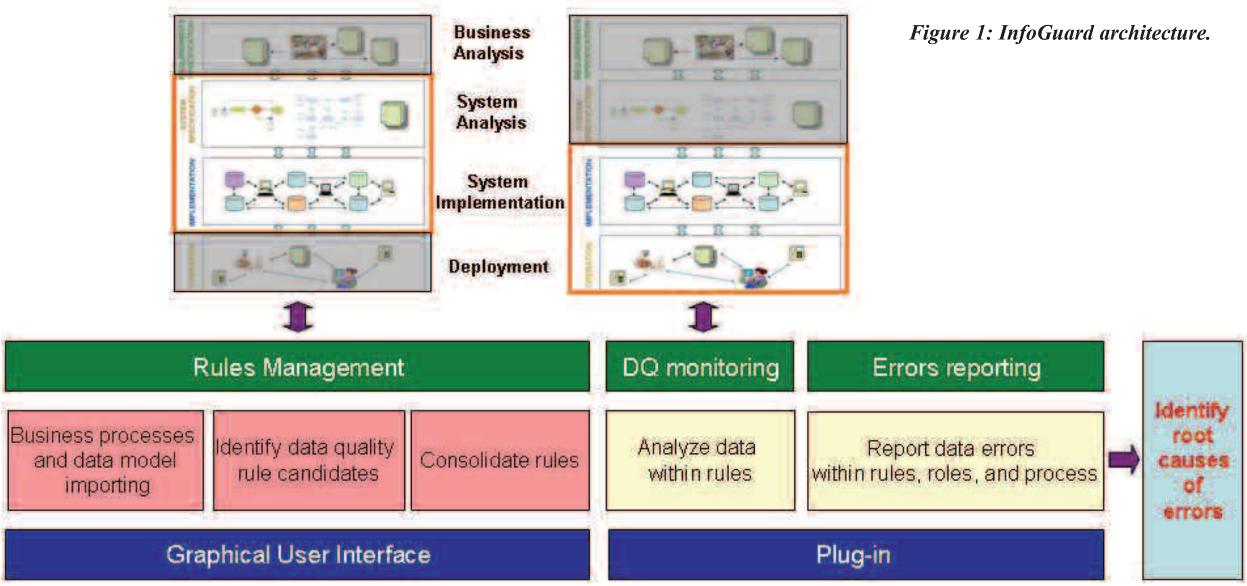

Business Analysis
System Analysis
System Implementation
Deployment
Rules Management
DQ monitoring
Errors reporting
Business processes and data model importing
Identify data quality rule candidates
Consolidate rules
Analyze data within rules
Report data errors within rules, roles, and process
Identify root causes of errors
Graphical User Interface
Plug-in

Figure 1: InfoGuard architecture.

plete and consistent. Thereby IQ management usually involves processes, tools and procedures with defined roles and responsibilities.

Total IQ Management is a cycle within 4 phases. The defining phase aims to define characteristics of information products and involved raw data, define IQ requirements for that information. The measuring phase focuses on IQ assessments and measurement against IQ requirements based on metrics such as data accuracy, completeness, consistency and timeliness. The result of this phase is analyzed in the analyzing phase which provides organizations with causes of errors and plans or suitable actions for improving IQ. The improving processes are taken place in the improving phase.

Many tools and techniques have been developed in order to measure the consistency of data, mostly involving some forms of business rules. Current approaches and tools for IQ auditing and monitoring often validate data against a set of business rules. These rules constrain the data entry or are used to report violations of data entries. However, the major challenge of these approaches is the identification of business rules, which can be complex and time consuming. Although simple rules can be relatively easy identifies, dynamic rules along business processes are difficult to discover.

Funded by Enterprise Ireland under the Commercialization programme, our research at the Business Informatics Group at Dublin City University focuses on addressing those challenges. A recent project -InfoGuard- aims to indentify and design business rules by combining business processes and data models. This project complements our already established rule-based data analyzing module for discovering root causes of data errors by indicating involved processes and involved organizational roles.

Our approach is based on business process documentation and data models extracted from legacy systems. The principles of our approach are to analyze rules based on the consistency between business process, data and responsibilities of organizational roles. Therefore, apart from business rules set out by the organizations, we can help users to identify IQ rules which come from correlations between business processes and data, particularly dynamic rules, and authority rules on data and business process performing. In other words, this approach makes clear the imprecision of the data model regarding to the business process model and vice-versa which are root causes of errors at the system specification level.

One of the most powerful features of our approach/tool is to support the rule identification by relating graphically business processes and data models. Business process elements and data elements concerning a rule are highlighted as long as designing/displaying the rule. By this way, the context and the rational of a rule is presented, which facilitates the rule management.
Once IQ rules are obtained and stored in a rule repository, the analyzing service can be called to detect incorrect data. The tool then produces error reports after analyzing process. The report includes incorrect data, the involved rules, the involved process and the involved organizational roles. Figure 1 illustrates the architecture of our approach.

In the future we will apply our approach to a broader scale of applications such as rules across applications, across databases. We will need data profiling across databases for total data quality analyzing.

Link:
http://www.computing.dcu.ie/big

Please contact:

Markus Helfert
School of Computing
Dublin City University
E-mail: markus.helfert@computing.dcu.ie

Integration of an Electrical Vehicle Fleet into the Power Grid

by Carl Binding, Olle Sundström,Dieter Gantenbein and Bernhard Jansen

The sudden peak of oil prices in 2008 together with the depletion of oil reserves, and the ecological impact of today's combustion engines have sparked renewed interest in fully electrical vehicles (EV) and plug-in hybrid EVs (PHEV). Based on typical driving patterns, it is believed that EVs could substantially alter the energy mix used for individual transportation, current limitations of battery technology not withstanding. In addition, in electrical power grids with a high percentage of fluctuating, renewable, energy sources (wind, solar) a larger fleet of EVs can absorb peaks of power production, and attenuate power troughs by acting as a distributed energy resource (DER), feeding power from its accumulator into the grid. In the context of the Danish EDISON project, we are investigating the potential impact of an EV fleet on electrical distribution grids and how to optimally integrate these resources into the power grid.

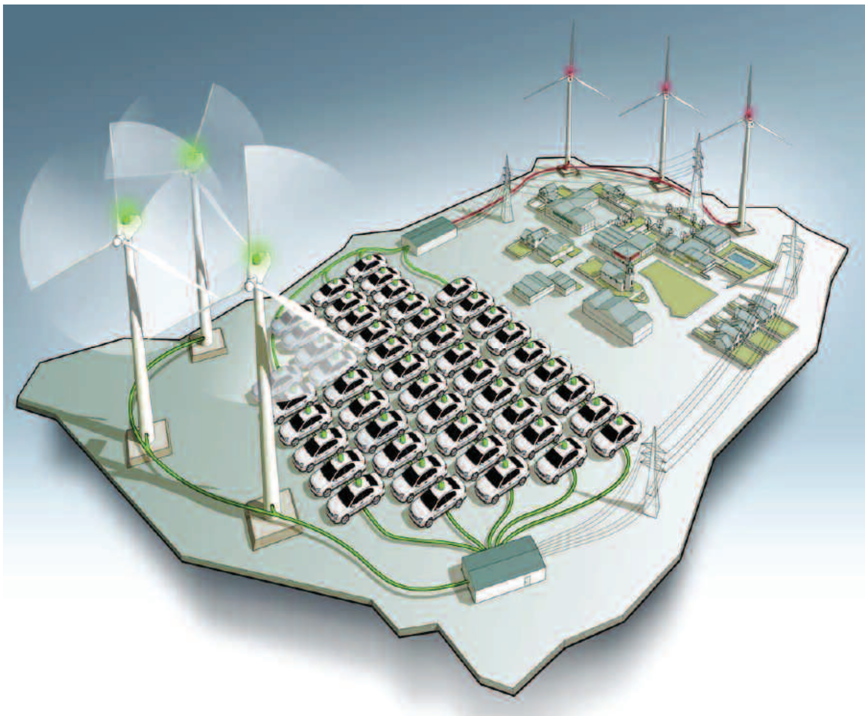


Figure 1: The EDISON Concept (Picture by courtesy of muff-illustration.ch).

In the recent past, several motor vehicle vendors have announced PHEVs or EVs for the US, the Japanese, and the European markets. The expected electrical range of these vehicles varies between 50 to 100 km, which is considered enough to cover a large percentage of daily driving patterns, in particular commuting to and from work.

The basic concept is to integrate an electrical accumulator into the vehicle which is recharged during stop-overs, preferably during periods in which electrical energy from environmentally friendly sources is abundant and exceeds the non-EV load in a power grid. Conversely, one can also imagine EVs feeding back excessive energy, when not needed by the car's operator, into the grid during times of peak demand when insufficient power is supplied by wind or solar. Hence, the vehicle's accumulator acts as a buffer for electrical energy. Since electrical power grids require continuous balancing of generated and consumed power, this energy buffering is of technical and economical interest as a supply of balancing power to the power grid. In particular, in grids with intermittent power supplies and limited balancing power supplies (hydro, gas turbines) additional balancing power capabilities are appealing.

The goals of the Danish EDISON project are to build a pilot system to aggregate a fleet of EVs and have this aggregator, or virtual power plant (EV-VPP), act as an entity capable of delivering and absorbing balancing power [2]. Clearly, given the power volumes of typical EVs charging (which varies from 3 kW upwards) versus the power delivered by a typical wind-turbine (1.5-3 MW), such aggregation is needed in order to achieve some balancing power impact.

Figure 1 illustrates the overall concept. CO_2 neutral wind-energy is fed into the EV fleet and can be consumed by the traditional domestic and industrial loads when power supply from intermittent, renewable, sources is reduced.

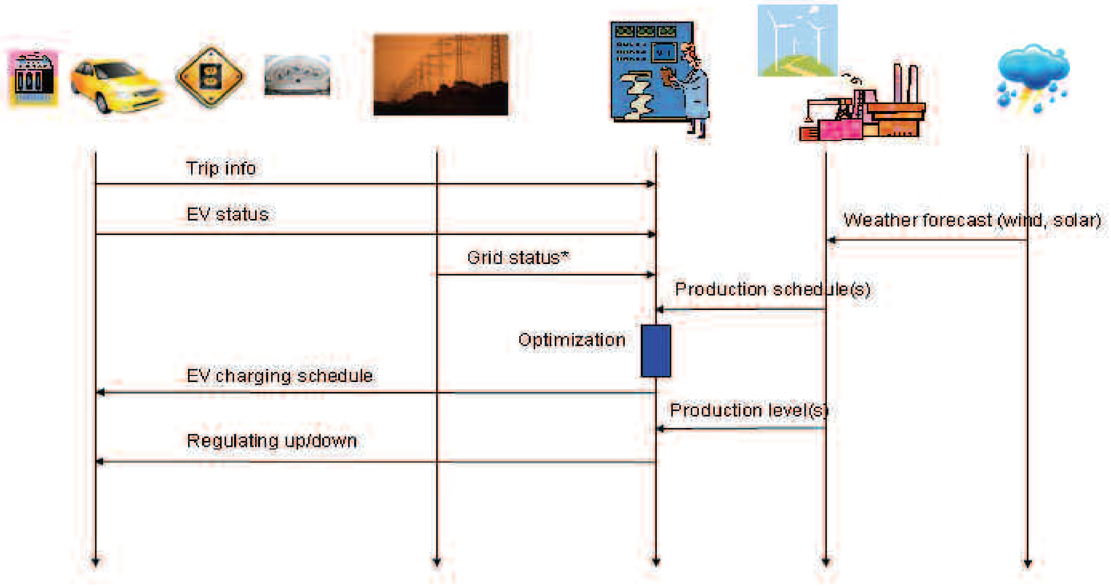
There are several challenges which need to be addressed by such an undertaking:

- *Resource management:* The EVs – as managed resources - need to maintain appropriate levels of state-of-energy in their accumulators to satisfy their user's requirements. Sufficient electrical power must be available when the vehicles are parked in order to recharge the batteries. When feeding power from the EV into the grid, enough energy for the next trip must remain stored in the accumulator. Ideally these requirements should be economically beneficial for the end user, ie he or she will want to charge the batteries in times of cheap power and possibly re-sell such power at a higher price during power troughs within the grid. A grid utility might have different objectives regarding pricing – a global cost optimization for provision of energy to the EV fleet is its likely objective.

- *Grid integration:* Smart charging or discharging must be aligned with the electrical grid's power generation ability and non-EV, conventional, loads. Based on forecasts of available renewable power, EVs can optimize their charging behavior under the assumption that energy costs are low in times of abundant renewable power and vice versa. The energy requirements for each EV can either be estimated or obtained from the EV users in order to perform planning for individual EVs in relation to overall power supply and non-EV load in the grid.

- *Communications:* Reliable, secure, timely and responsive communications are needed to exchange information between distributed energy producers, mobile energy con-

Figure 2: EV-VPP Architecture - Message Flow.



sumers, and the EV-VPP. The grid's status, as well as its static parameters, is required for a comprehensive optimization and this information needs to be relayed to the VPP.

- *Billing and customer relationship management:* Last but certainly not least, end-users will have to be billed for power consumed, and power fed from the EV's battery into the grid must be accounted for. This is based on metering information which can be transmitted from the EV's charging spot into the grid and utilities IT infrastructure. End-users will need to be provided with the ability to register, to deregister, and to manage their profiles with the EV-VPP. To enable larger reach of EVs, roaming between VPPs also needs to be provided; similar to cellular telephone systems this requires appropriate and interconnected back-end infrastructure.

Our approach to solving the above integration problem is centered on two key aspects. First we postulate the architecture of an EV-VPP. As shown in Figure 2, the data flows into the VPP, originating from various entities in the power grid: the loads (ie, the EVs), the transport and distribution grids, and the power generation. The VPP's core task is to derive the optimal charging plans for all vehicles under its control and disseminate these to the EVs.

The second key postulate is that we perform a centralized, global, optimization of the charging behavior. Using information gathered from energy consumers and generators as well as grid operators, we set up a linear programming optimization which computes a charging plan for each EV which minimizes the global energy costs for total EV recharging. We also address potential distribution grid constraints by including power flow bottlenecks in the planning of the charging. The computed charging plan is eventually pushed back to the EVs and its execution is then controlled during the daily grid operations; possible overriding control messages are generated in case of non-anticipated deviations from the plans.

We are currently simulating the above optimizations using a synthesized grid patterned after the Bornholm (DK) power grid [1]. We include approximately 11,000 low-voltage access points and simulate the motion of EVs over that grid. For each EV, we maintain historical trip data which we use in order to estimate future energy needs. The generation estimates are based on wind-data and conventional generator resources. We have been able to achieve reasonable performance using a fleet of over 1,000 vehicles simulated on conventional lap-top computers.

Our current work focuses on integrating more available real-world trip and grid data into our simulation. We are also investigating the convergence and complexity issues associated with the impact of the grid constraints into the planning for EV charging. Based on our work so far, we believe that our approach delivers optimal grid and resource utilization. Our investigations continue.

Links:

[1] Wind power system of the Danish island of Bornholm: Model set-up and determination of operation regimes http://www.elektro.dtu.dk/English/about_us/staff/staff_lists/staff.aspx?lg=showcommon&id=240821

[2] The EDISON Project http://www.edison-net.dk/

[3] IBM ILOG CPLEX V12.1: User's Manual for CPLEX, ftp://public.dhe.ibm.com/software/websphere/ilog/docs/optimization/cplex/ps_usrmancplex.pdf

Please contact:

Carl Binding
IBM Zurich Research Laboratory, Switzerland
E-mail: cbd@zurich.ibm.com

Graph Transformation for Consolidation of Creativity Sessions Results

by Peter Dolog

Graph transformation approach for consolidation of creativity sessions results is part of the FP7 EU/IST project idSpace: Tooling of and training for collaborative, distributed product innovation. The goal of graph transformation approach is to provide a tool for merging results of various sessions (such as brainstorming sessions), which are represented as graphs, when the session participants- are physically distributed.

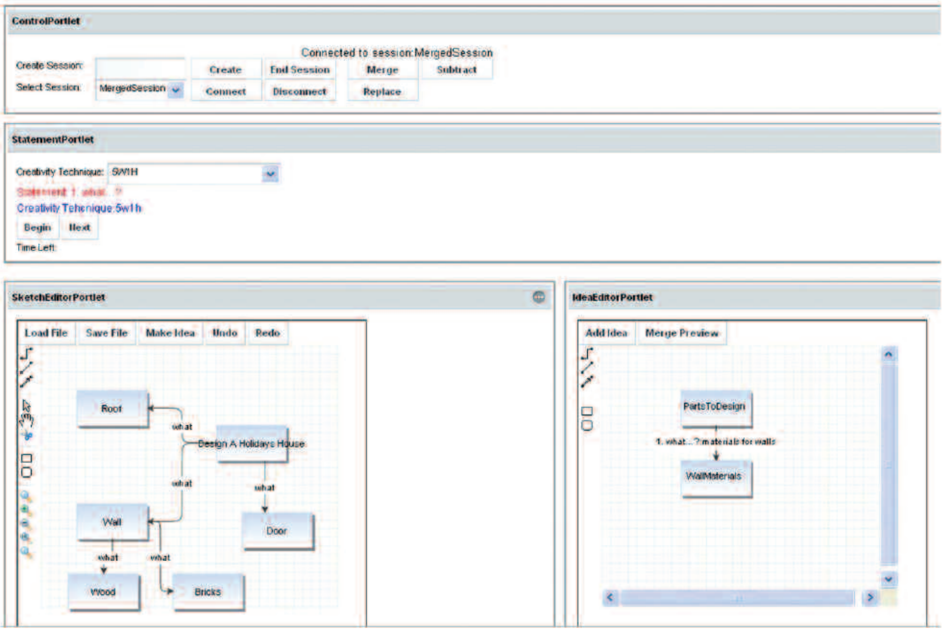
Creativity is a mental and social process involving the discovery of new ideas or concepts, or new associations of the creative mind between existing ideas or concepts. It has been studied from various perspectives including behavioural psychology, social psychology, psychometrics, cognitive science, artificial intelligence and philosophy. Elicitation of such ideas is usually supported by one or more so called "creativity techniques" which are usually performed in a group.

We are dealing with creativity in the context of discovery of ideas (and associations between them) applied to new product development. From this perspective, each meeting or session which focuses on new product development results in some kind of model of ideas with associations between them. One example of such a model is a mind map (a graph where nodes represent ideas and edges between them define relationships between them).

New product development usually does not end with one session. It typically consists of several sessions performed with several participating teams in different geographical locations. This is especially relevant in situations in which a new product needs to be assembled from components of physically distributed suppliers within a supply chain. As there are usually multiple models resulting from several sessions, there is a natural requirement to consolidate the results with electronic support.

The idSpace project started in April 2008 and finished in March 2010. The idSpace project aims to provide a web-based platform which supports the distributed creative sessions and ideation. The results of ideation sessions are represented as graphs. The follow up processing requires consolidation of the results into a unified model. As the graph structure is adopted, the consolidation can be done with the help of series of graph transformations. Graph transformations are used to represent merge, subtract and replace operations.

The web based tools (see Figure 1) are based on lightweight portlet technology, which is based on Liferay Portal Technology. This method integrates sketch editor for free drawing in ideation sessions, idea editor for preserving ideation results based on MxGraph software, statement portlet for association label suggestions according to chosen creativity technique, portlets for graph transformations (merge, subtract, and replace buttons in Figure 1) and other portlets realized by different members of the idSpace consortium. The idSpace consortium consists of Open University of the Netherlands, Aalborg University, University of Cyprus, Landesinitiative Neue Kommunikationswege Mecklenburg-Vorpommern e.V, University of Pireaus Research Center, University of Hildesheim, Morpheys Software, Space Applications, and Extreme Media Solutions.


Connected to session:MergedSession

Create Session:

Select Session: MergedSession

Creativity Technique: 5W1H

Statement 1: what...?

Creativity Tehenique:5w1h

Time Left:

Figure 1: Portlets for graph transformation.

Let us exemplify the idea graph using the 5W1H creativity technique . The 5W1H is one of the creativity techniques explored in idSpace project, it stands for the six question words (what, why, where, when, who and how). A typical scenario in which the "5WH1" creativity technique may be applied is shown in Figure 1, using "holiday house simple ideation session" results. The aim of this session was to find new ideas on better improved materials for such a house. One way in which this technique may be applied is to use the answers arising from the 5W1H questions as input for later questions.

The idea behind creativity techniques is to encourage lateral thinking, because it is a common human behavior to stick to one line of thought that may be preferred at first glance. This line of thought is steared by the statements, guidelines, and questions from the creativity techniques. Each generated idea (as for example in Figure 1) can be connected to another by an association with a label taken from the set of 5WH1 questions (labels on edges in Figure 1 for example). This can be modeled as graph. Further, each idea represented as a node can be supported by a sketch drawn in a sketch editor. Once a group is finished with the idea graph, a consolidation session might be called upon where the graph transformation of existing graphs, including the one generated in the latest session, are employed to achieve a final agreed upon suggestion for a solution.

The benefit of such an approach is that teams get user-friendly support for consolidation and preservation of the creativity session results. The results are far more than just snapshots of white boards; they can be used for later processing, such as automatic discovery of hidden implicit relations between ideas, recommendations of related existing ideas, guiding through an information space of ideas and so on. As a consequence more people are able to contribute ideas, to manage creativity and its results more transparently and efficiently, and to benefit from previous knowledge constructed in other creative sessions.

Link:
http://www.idspace-project.org

Please contact:
Peter Dolog
Aalborg University, Denmark
E-mail: dolog@cs.aau.dk

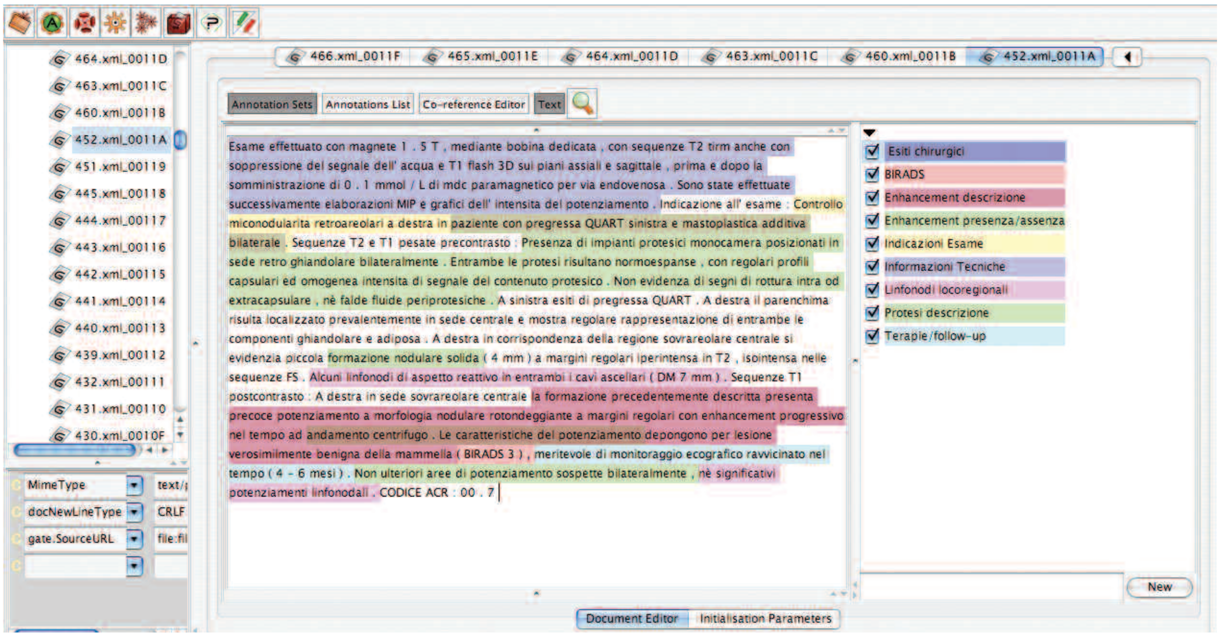
Extracting Information from Free-text Mammography Reports

by Andrea Esuli, Diego Marcheggiani and Fabrizio Sebastiani

Researchers from ISTI-CNR, Pisa, aim to effectively and efficiently extract information from free-text mammography reports, as a step towards the automatic transformation of unstructured medical documentation into structured data.

Information Extraction is the discipline concerned with the extraction of natural language expressions from free text, where these expressions instantiate concepts of interest in a given domain. For instance, given a corpus of job announcements, one may want to extract from each announcement the natural language expressions that describe the nature of the job, annual salary, job location, etc. Put another way, information extraction may be seen as the activity of populating a structured information repository (such as a relational database, where "job", "annual salary" and "job location" count as attributes) from an unstructured information source such as a corpus of free text. Another example of information extraction is searching free text for named entities, ie, names of persons, locations, geopolitical organizations, and the like.

An application of great interest is extracting information from free-text medical reports, such as radiology reports. These reports are unstructured in nature, since they are written in free text by medical personnel. However, applying



Esame effettuato con magnete 1 . 5 T , mediante bobina dedicata , con sequenze T2 tirm anche con soppressione del segnale dell' acqua e T1 flash 3D sui piani assiali e sagittale , prima e dopo la somministrazione di 0 . 1 mmol / L di mdc paramagnetico per via endovenosa . Sono state effettuate successivamente elaborazioni MIP e grafici dell' intensita del potenziamento . Indicazione all' esame : Controllo miconodularita retroareolari a destra in paziente con pregressa QUART sinistra e mastoplastica additiva bilaterale . Sequenze T2 e T1 pesate precontrasto : Presenza di impianti protesici monocamera posizionati in sede retro ghiandolare bilateralmente . Entrambe le protesi risultano normoespanse , con regolari profili capsulari ed omogenea intensita di segnale del contenuto protesico . Non evidenza di segni di rottura intra od extracapsulare , nè falde fluide periprotesiche . A sinistra esiti di pregressa QUART . A destra il parenchima risulta localizzato prevalentemente in sede centrale e mostra regolare rappresentazione di entrambe le componenti ghiandolare e adiposa . A destra in corrispondenza della regione sovrareolare centrale si evidenzia piccola formazione nodulare solida (4 mm) a margini regolari iperintensa in T2 , isointensa nelle sequenze FS . Alcuni linfonodi di aspetto reattivo in entrambi i cavi ascellari (DM 7 mm) . Sequenze T1 postcontrasto : A destra in sede sovrareolare centrale la formazione precedentemente descritta presenta precoce potenziamento a morfologia nodulare rotondeggiante a margini regolari con enhancement progressivo nel tempo ad andamento centrifugo . Le caratteristiche del potenziamento depongono per lesione verosimilmente benigna della mammella (BIRADS 3) , meritevole di monitoraggio ecografico ravvicinato nel tempo (4 - 6 mesi) . Non ulteriori aree di potenziamento sospette bilateralmente , nè significativi potenziamenti linfonodali . CODICE ACR : 00 . 7

☑ Esiti chirurgici
☑ BIRADS
☑ Enhancement descrizione
☑ Enhancement presenza/assenza
☑ Indicazioni Esame
☑ Informazioni Tecniche
☑ Linfonodi locoregionali
☑ Protesi descrizione
☑ Terapie/follow-up

Figure 1: A mammographic report automatically annotated according to the nine concepts of interest.

information extraction would be beneficial, since extracting key data and converting them into a structured (eg, tabular) format would greatly contribute towards endowing patients with electronic medical records that, aside from improving interoperability among different medical information systems, could be used in a variety of applications, from a patient's care to epidemiology and clinical studies.

In recent months we have worked on a system for automatically extracting information from mammography reports written in Italian. There are two main approaches to designing an information extraction system. One is the rule-based approach, which consists of writing a set of rules which relate natural language patterns with the concepts to be extracted from the text. This approach, while potentially effective, is too costly, since it requires a lot of human power to write the rules, which are jointly written by a domain expert - say, an expert radiologist - and a natural language engineer. We have followed an alternative approach, which is based on machine learning. According to this approach, a general-purpose learning software learns to relate natural language patterns with the concepts to be instantiated from a set of manually annotated free texts, ie texts in which the instances of the concepts of interest have been marked by a domain expert. The advantage of this approach is that the human power required to annotate the texts needed to train the system is much smaller than that needed to manually write the extraction rules.

The system we have built uses "conditional random fields" (CRFs) as a learning technique. CRFs were explicitly devised for managing data of a sequential nature, such as text, and have given good results on text-related tasks such as named entity extraction and part-of-speech tagging.

We have tested our system on a set of 405 mammographical reports written (in Italian) by medical personnel of the Radiology Institute of Policlinico Umberto I of Rome, and manually annotated by two expert radiologists of the same institution according to nine concepts of interest (eg, "Followup Therapies"). 336 reports were annotated by one radiologist only, while the other 59 were independently annotated by both radiologists. The presence of reports annotated by both radiologists has allowed us to directly compare the accuracy of our system with human accuracy, by comparing the agreement between the system's and the radiologist's annotations with the agreement between the annotations of the two radiologists. Our experiments, run by 10-fold cross validation, have shown that our system obtains near-human performance: the agreement between system and human, measured by the standard "macroaveraged F1" measure, turned out to be 0.776, while the agreement between the two experts was 0.794 (higher values are better, since 0 and 1 indicate perfect disagreement and perfect agreement, respectively). These results are especially encouraging because no specialized lexical resource was used in the experiments, since no such resource exists for the radiological / mammographic sector for Italian.

Please contact:
Fabrizio Sebastiani
ISTI-CNR, Italy
E-mail: fabrizio.sebastiani@isti.cnr.it

Parental Control for Mobile Devices

by Gabriele Costa, Francesco la Torre, Fabio Martinelli and Paolo Mori

As a result of the rapid increase in mobile device capabilities, we believe that many users will soon migrate most of their tasks from the computer to the smart phone. We thus feel that new, more effective security mechanisms are needed to protect users, especially minors. For this reason, we are currently engineering a software suite designed to monitor all security-relevant activities and, where necessary, block unsuitable material.

In western countries mobile phones now outnumber people. The majority of these phones have embedded cameras and good connectivity support, connecting to the Internet via their mobile operator network or in wireless mode, and communicating directly with other devices through a Bluetooth interface. This means that the gap in functionality between personal computers and smart phones is being continuously reduced. Mobile devices such as smart phones or Personal Digital Assistants (PDAs) are now powerful enough to run applications that are comparable to those running on normal PCs.

The mobile phone is thus now typically employed for many tasks, such as browsing the Internet, reading and writing e-mails, exchanging data (photos, contacts, files, etc) with other devices, and so on. Clearly, these operations bring with them new vulnerabilities and raise new threats to the user's privacy. A particular source of concern is the widespread use of mobile devices by young people and minors who can easily access content that is inappropriate for their age. Filters provided by the mobile telco operators are not sufficient to block such material because they only inspect data sent via the telco network and have no control on direct connections (eg Bluetooth). As a consequence, in recent years, much research has focused on the provision of customised, security support for mobile devices. However, what is still missing is a general security framework that controls applications as well as phone calls and message content.

In order to meet this need, by calling upon the experience gained in the EU project S3MS (Security of Software and Services for Mobile Systems), we have developed a new modular security monitor support for mobile devices which uses a plugin-based, extensible architecture to meet the demands of security requirements which change from device to device and from user to user.

Our system aims to monitor all security-relevant activities involving a mobile device. The basic idea is to develop a minimal security core that evaluates the platform security state, accepts customised security policies, and deals with heterogeneous system events.

The system can be easily extended with new modules. When, following authentication, a Module M is plugged into the system. It injects the list of its security-relevant events and

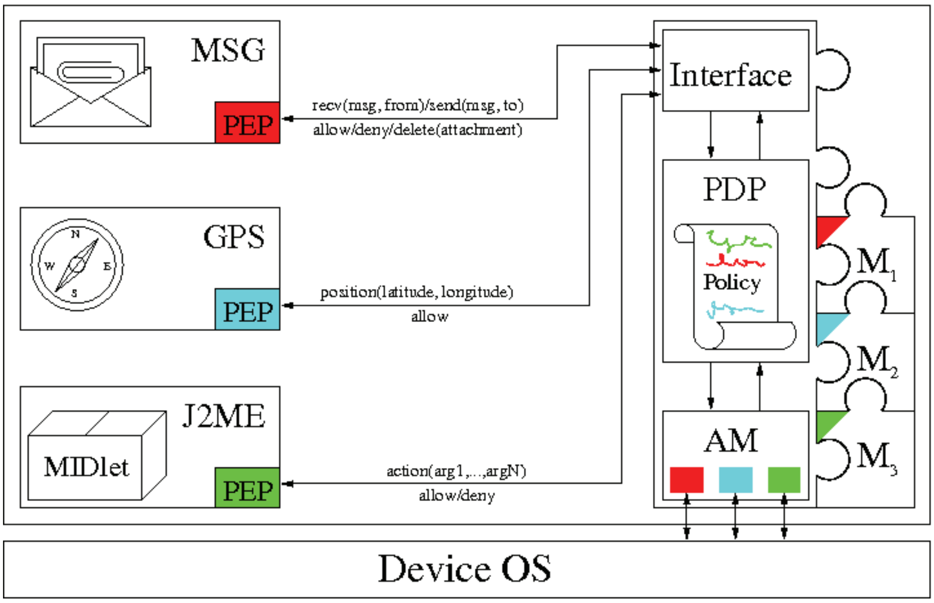


Figure 1: The system core consists of three components: an events interface (Interface), a Policy Decision Point (PDP) and an Attribute Manager (AM). The Interface receives the security events, authenticating and dispatching them to the PDP. The PDP holds the policy defining the currently enforced rules and the security state. Whenever it receives an event from the Interface, it updates the internal state and computes the reaction to be returned. The PDP is able to access some system information through the AM.

the new attributes it needs to handle and then disseminates its own Policy Enforcement Points (PEPs) within the device software.

Each PEP checks the behaviour of a critical system component and enforces certain operations (depending on what the PEP is set to monitor). When a PEP captures a sensitive action, it generates a corresponding event and sends it to the system interface. It will then receive the action to be enforced on its target.

Security policies are defined using an editor only accessible by the device owner. The policy editor provides the user with a human-readable graphic interface so that system policy can be defined or modified without the need for technical skills. Security polcies generated by the editor are passed to the PDP to be loaded and enforced.

A very general class of security rules that can involve multiple modules is thus obtained. The example shown in Figure 1 uses three Modules: M_1 (red) for messages control, M_2 (blue) for position control and M_3 (green) for Java applications control. Each module places its own PEP, declares the necessary attributes and specifies their security actions and reactions (labels above and below the connections, respectively). For instance, the PEP watching the transmission of messages sends an action signal whenever a message is sent or received. The action contains information about the message content (msg) and the sender/recipient (from/to). The possible reactions to this event are: allow (if the policy permits the operation), deny (if the policy prohibits it) or delete an attachment (if the message contains dangerous material). Assessment of the danger of an attachment (eg a picture or a video-clip) is done by automatically classifying its content, either remotely on a server or locally on the device itself. The system maps one of its attributes to the classifier invocation and uses it to determine whether a picture is offensive (eg pornographic) or not.

Another PEP, attached to the system GPS, gives the geographical position of the device whenever it is available. This information can be exploited to define location-based security policies, eg alerting a parentif the device moves outside certain area.

The last PEP monitors security-relevant operations (eg open/close connections or read/write files) performed by Java MIDlets. This PEP provides valuable feedback on which Java program is executed (method MIDlet.start()).

The large variety of information available to our PDPs makes it possible to apply very expressive constraints. In this way parents can tailor a specific security for their children's devices. For instance, with a configuration similar to that shown in Figure 1, rules like: "Do not execute Java games at school" or "Alert me with a SMS if the device receives some pornographic content (and block this content)". Such rules can be composed into highly expressive and effective security policies.

Please contact:
Fabio Martinelli
IIT-CNR, Italy
E-mail: fabio.martinelli@iit.cnr.it

Preschool Information When and Where you Need it

by Stina Nylander

Having children in preschool can be a logistic challenge. Parents need to keep track of dates for events, what to bring, changes in the opening hours of the preschool, and many other things. PreschoolOnline is an attempt to support the information exchange between parents and preschool teachers and provide an opportunity to study how ICT can be used in preschool.

As part of a joint project, carried out by SICS, Squace AB and Nockeby Preschools we have developed PreschoolOnline, a web service also available on mobile phones. The service is organized in communities corresponding to the groups of children in preschool, and allows parents to access information about their child's group and the preschool as well as contact information for other parents in that group. Teachers can add and update information about the preschool, parents can add and update information about themselves and their children, and everyone can post on the notice board.

During the project, PreschoolOnline was deployed in six preschools and used by more than 500 parents in a five month field trial. This gave the researchers an excellent opportunity to study real life use through surveys, interviews, and logs. In general, both parents and teachers held positive attitudes toward the service. PreschoolOnline offered access to information when and where it was needed. Moreover, by providing a means of communication of routine and logistical information the service enabled teachers and parents to spend more time talking about more individual aspects of the children's schooling and development when they met at preschool.

PreschoolOnline showed that even a simple service can improve information exchange between parents and teachers. We are now adding functionality to the service to create not only a tool for disseminating information but also to gather information, such as vacation dates and whether or not parents plan to attend a meeting.

The six original preschools have chosen to continue to use PreschoolOnline after the project has ended, and more than 40 other preschools have tried out the service on their own initiative. Furthermore, the City of Stockholm is interested in using PreschoolOnline for all preschools in the city.

Link:
http://www.forskolanimobilen.se

Please contact:
Stina Nylander
SICS, Sweden
Tel: +46 8 633 15 69
E-mail: stny@sics.se

Call for Papers

HCI International 2011 - 14th International Conference on Human-Computer Interaction

Hilton Orlando, Bonnet Creek,
Orlando, Florida, USA, 9-14 July 2011

jointly with:

- Symposium on Human Interface (Japan) 2011
- 9th International Conference on Engineering Psychology and Cognitive Ergonomics
- 6th International Conference on Universal Access in Human-Computer Interaction
- 4th International Conference on Virtual and Mixed Reality
- 4th International Conference on Internationalization, Design and Global Development
- 4th International Conference on Online Communities and Social Computing
- 6th International Conference on Augmented Cognition
- 3rd International Conference on Digital Human Modeling
- 2nd International Conference on Human Centered Design
- 1st International Conference on Design, User Experience, and Usability

Overview and Areas of Interest

HCI International 2011 jointly with the affiliated conferences, which are held under one management and one registration, invite you to Orlando, Florida, USA to participate and contribute to the international forum for the dissemination and exchange of up-to-date scientific information on theoretical, generic and applied areas of HCI through the following modes of communication: Plenary / Keynote Presentation(s), Parallel Sessions, Poster Sessions, Tutorials and Exhibition.

The Conference focuses on the following major thematic areas:

- Ergonomics and Health Aspects of Work with Computers
- Human Interface and the Management of Information
- Human-Computer Interaction
- Engineering Psychology and Cognitive Ergonomics
- Universal Access in Human-Computer Interaction
- Virtual and Mixed Reality
- Internationalization, Design and Global Development
- Online Communities and Social Computing
- Augmented Cognition
- Digital Human Modeling
- Human Centered Design
- Design, User Experience, and Usability

Deadlines

- *Papers*
  Abstract: 15 October 2010
  Notification: 3 December 2010
  Camera-ready: 4 February 2011

- *Posters*
  Abstract: 11 February 2010
  Notification: 4 March 2011
  Camera-ready: 1 April 2011

- *Tutorials*
  Abstract: 15 October 2010
  Notification: 3 December 2010
  Camera-ready: 6 May 2011

Individuals can appear as co– authors in several papers in HCI International 2011 and the affiliated conferences, but can present only one paper.

Awards

A total of thirteen awards will be conferred during HCI International 2011. A plaque and a certificate will be given to the best paper of each of the eleven Affiliated Conferences / Thematic Areas, amongst which, one will be selected to receive the golden award as the Best HCI International 2011 Conference paper. Moreover, the Best Poster extended abstract will also receive a plaque and a certificate.

Keynote address

Professor Ben Shneiderman is the keynote speaker for HCI International 2011. Ben Shneiderman is a Professor in the Department of Computer Science, founding director of the Human Computer Interaction Lab and Member of the Institute for Advanced Computer Studies at the University of Maryland. The pioneering work of Ben Shneiderman has played a key role in the development of HCI as a new academic discipline, by bringing scientific methods to the study of human use of computers. For more than 30 years, Ben Shneiderman has promoted HCI through his prolific writing and lecturing. He is one of the most frequently referenced researchers in computer science. The title of his keynote address is "Technology-Mediated Social Participation: The Next 25 Years of HCI Challenges".

Proceedings

The HCI International 2011 Conference Proceedings will be published by Springer in a multivolume set in the Lecture Notes in Computer Science (LNCS) and Lecture Notes in Artificial Intelligence (LNAI) series.

Exhibition

The HCI International Conference is an ideal opportunity to exhibit your products and services to an international audience of about 2,000 researchers, academics, professionals and users in the field of HCI. Attendees will be able to examine state-of-the-art HCI technology and interact with manufacturing representatives, vendors, publishers, and potential employers.

Student Volunteers

The HCI International 2011 Program for Student Volunteers gives university students from around the world the opportunity to attend and contribute to one of the most prestigious conferences in the field of Computing and Human-Computer Interaction. Being a Student Volunteer is a great opportunity to interact closely with researchers, academics and practitioners from various disciplines, meet other students from around the world, and promote personal and profession growth. The HCI International Conference's past, present and continued future successes are due in a large part to the skills, talents and dedication of its Student Volunteers.

More information:
http://www.hcii2011.org

Call for Participation

CEDI 2010 - Third Spanish Conference on Informatics

Valencia, Spain, 7-10 September 2010

Several national symposia and workshops have traditionally been organized on specific topics related to computer science and engineering. CEDI joins all of them into a single conference, and this is the third time that CEDI is organized. The main goal is to gather the Spanish research community on computer science and engineering in order to proudly show our scientific advances, discuss specific problems and gain visibility in Spanish society, emphasizing Spain's role in the new age of the Information Society.

The conference is organized in a 'federated conference' format, by joining several more specific symposia (22 and growing), including Artificial Intelligence, Software Engineering and Databases, Programming Languages, Parallelism and Computer Architecture, Computer Graphics, Concurrency, Soft-computing, Pattern Recognition, Natural Language Processing, Ubiquitous Computing and Human-Computer Interaction. In addition to the scientific activities of each symposium, the CEDI conference will feature two invited keynotes, two round tables and social events and a gala dinner at which several awards will be delivered. Through these awards, the research community on informatics will recognize the efforts of some colleagues in promoting informatics research in Spain.

The expected attendance is around 1,500 scientists. The event is funded by the Spanish Ministry of Education, and sponsored by SpaRCIM. The conference will be organized by the Universidad Politécnica de Valencia. The conference chair will be Prof. Isidro Ramos, the Scientific Program Chair will be Prof. José Duato, and the Local Organization Chair will be Prof. Juan Miguel Martinez, all of whom are from the Universidad Politécnica de Valencia. This university is one of the leading technical universities in Spain, and has a strong focus on innovation and research.

More information:
http://www.congresocedi.es/2010/

Call for Participation

W3C Web on TV Workshop

Tokyo, Japan 2-3 September 2010

The demand for access to applications, video, and other network services continues to grow. The Web platform itself continues its expansion to support mobile devices, television, home appliances, in-car systems, and more consumer electronics. To meet the growing demand, the Web platform of the future will require smarter integration of non-PC devices with Web technology so that both hardware and software vendors can provide richer Web applications on various devices at lower costs. This is the first of a planned series on "Web on TV" workshops to bring various communities together to discuss this integration. This first Workshop will be hosted W3C/Keio with the support of the Japanese Ministry of Internal Affairs and Communications. The Workshop will be conducted in Japanese, Korean, Chinese, and English via simultaneous translation. A meeting summary will be available in English. A workshop of this series is planned in Europe early 2011.

More information:
W3C workshop calendar:
http://www.w3.org/2003/08/Workshops/
W3C Web on TV Workshop:
http://www.w3.org/2010/09/web-on-tv/

Call For Participation

ICT 2010

Brussels, 27-29 September 2010

ICT 2010, Europe's biggest event for research in information and communication technologies, will be organised by the European Commission DG Information Society and Media. ERCIM and ERCIM-managed projects will most certainly have representation at this key event through booths and speakers, as in Lyon, France, in 2008.

More information:
http://ec.europa.eu/information_society/events/ict/2010/

Call for Submissions

Special issue of the International Journal of RF Technologies: Research & Applications on

RFID and the Internet of Things in Europe

the ERCIM-managed RACE network RFID (http://www.race-networkrfid.eu/), is designed to become a federating platform to the benefit of all European stakeholders in the development, adoption and usage of RFID. The network solicits submissions to the Special Issue of the Journal for RF Technologies: Research and Applications on research in the fields of RFID and the Internet of Things in Europe.

The mission of the Special Issue is to provide an overview on the ongoing research activities concerning RFID and the Internet of Things. The overall objectives are:

- to provide a European perspective on research and development
- to discuss European stakeholder involvement, ethical issues and governance
- to evaluate the impact of RFID and the Internet of Things on people;
- to present new business developments and;
- to build a bridge between research and practice in Europe.

Target audience

The Special Issue is intended to support a professional audience of researchers, top managers and governmental institutions. So even though we will require a professional scientific contribution the style of writing should address a wider audience. Submissions need to contribute new and original research and review articles. Contributions are specifically welcome from ongoing European research projects.

Deadline for submission: 15 September 2010. All submitted chapters will be reviewed (double-blind review). The Special Issue is scheduled to be published in the second quarter of 2011.

More information:
http://www.iospress.nl/loadtop/load.php?isbn=17545730

http://www.race-networkrfid.eu/

Master of Science in Computational Biology and Biomedicine at Nice

While biological data exhibits a formidable diversity, the past two decades have seen the advent of massive data produced either by high throughput experiments or by measurement devices of increasing accuracy at very different scales ranging from nano to macro. Handling these massive and complex data within a virtuous cycle linking modeling and measurements is one of the major challenges in Computational Biology and Biomedicine.

Starting your career

The aim of this program is to provide excellent academic or industrial career opportunities by offering high level coverage of modeling and computing principles that will enable the challenges to be met and make tomorrow's technological choices in biological, medical computing domains.

Scientific Goal

The goal of this program is to focus on the human being from different perspectives (understanding and modeling functional aspects or interpreting biomedical signals from various devices) and at different scales (from molecules to organs and the whole organism). Courses are organized into 3 categories: 1) bioinformatics, 2) biomedical signal and image analysis and 3) Modeling in neuroscience. All classes will be given in English by several outstanding professors and researchers from research institutes in the campus: Nice Sophia Antipolis University/CNRS (I3S, IPMC, LJAD, INLN) and INRIA.

Scholarships

The scholarship program offers outstanding students the chance to receive a grant covering the living expenses during the first period of the master. Early application increases the chances of receiving financial aid. Therefore, it is in the applicants best interest to submit his/her application at the first round of applications.

Admission Criteria:

Due to the multi-disciplinary nature of the program, the program is designed for those having completed the first-year MSc program at home institution in either computer science, electrical engineering, applied mathematics, mathematical biology, bioinformatics or biophysics.

Please contact:
Frederic Cazals and Pierre Kornprobst
Programme coordinators
E-mail: msc-compbio.coordinator@lists-sop.inria.fr

More information:
http://www.computationalbiology.eu

Spanish Pilot Project for Training and Development of Technology Applied to Systems Biology

The 'KHAOS' research group at the University of Málaga has developed several applications in a pilot project for training and development of technology applied to Systems Biology.

With the appearance of analysis tools in systems biology, sources of information have been growing exponentially, creating a necessity to approach the problem of the integration of existing information. The applications developed in the pilot project solve specific problems facing the researchers working in this area. These applications focus on data integration and Web Services by means of semantics (and the use of data available as Liked Data):

- *BioBroker* - This tool integrates different biological databases using XML as a model of data interchange (Navas-Delgado et al. SPE 2006). This was the starting point for the activities of Khaos in Computational Biology.
- *SB-KOM* - A mediator based on using KOMF as the framework and using optimized scheduling algorithms for the biological data bases (Navas-Delgado et al. Bioinformatics 2008). This mediator is the natural evolution of BioBroker towards the use of semantics when integrating biological data.
- *ASP 3D Model Finder (AMMO Prot)* - An application that enables the user to view the three-dimensional protein structure in the metabolism of Amines, using computer generated methods. This structure is generated using existing information and it is the first tool that took advantage of SB-KOM.
- *SBMM-Assistant* - This assistant allows the user to locate information about metabolic routes (including details about their components) (http://sbmm.uma.es). In addition, it allows the edition of these routes by the users. This tool allows users studying metabolic pathways to retrieve information from many databases and curate them manually.
- Social Pathway Annotation (http://sbmm.uma.es/SPA). This extension of SBMM introduces the use of social networking tools to enable the collaborative curation of pathway models.
- BioSStore. In this work we address the drawbacks presented by bioinformatics services and try to improve the current semantic model introducing the use of the W3C standard Semantic Annotations for WSDL and XML Schema (SAWSDL) and related proposals (WSMO Lite).
- Scientific Mashups. The application of mashups can be useful within biology . Automatic processes are needed to develop biological components which will be applied by end-users to develop specific biological mashups using a mashup platform, ie. EzWeb Platform.

All research conducted by KHAOS is funded by several research grants.

More information:
http://khaos.uma.es

Van Dantzigprijs for Peter Grünwald and Harry van Zanten

Peter Grünwald (left) and Harry van Zanten.

In April, Peter Grünwald (CWI and Leiden University) and Harry van Zanten (Eindhoven University of Technology and EURANDOM) received the Van Dantzig Prize, the highest Dutch award in statistics and operational research. The prize is awarded every five years by the Netherlands Society for Statistics and Operations Research (VVS-OR) to a young researcher who has contributed significantly to developments in these research areas.

More information: http://www.vvs-or.nl

VISITO Tuscany

VISITO Tuscany was presented by ISTI-CNR at a press conference in the historical centre of Pisa. VISITO provides an interactive guide for tourists visiting cities of art via a smart phone application. To receive detailed information on a monument, the user simply takes a photo of it. The system will support tourists in all phases of their trip, from the initial planning to post-visit archiving and sharing over social networks. Leveraging on techniques of image analysis, content recognition and 3D scanning and browsing, on returning home the tourist can relive the highlights of the trip via virtual visits to the monuments photographed. The project, coordinated by Giuseppe Amato, ISTI-CNR, is supported by the Tuscan Region and by the European Regional Development Fund. Project members include IIT-CNR and three private companies: Alinari24Ore, Hyperborea, and 3Logic MK.

More information:
http://www.visitotuscany.it/index.php/en

MoTeCo - A New European Partnership for Development in Mobile Technology Competence

MoTeCo is a new project supported by the European Leonardo da Vinci Programme that will offer a training curriculum in mobile technology, in particular for the Symbian operating system (http://www.symbian.org/). At the moment, such trainings are accessible only in a few European countries located in the north-west of the continent.

The project will assure better access to training in this field by transferring the training methodology from already existing training centres in Europe to new training centres in South and Central Europe. This will be achieved by:

- the establishment of five new training centres for mobile software development
- five training curricula with additional training material (textbooks, workbooks, presentations, e-learning platforms) adapted to specific national and institutional needs
- professional preparation of a group of 21 trainers ready to prepare other trainers and programmers.

More generally, MoTeCo will thus transfer knowledge to the regions of the partners, offering a unique experience for institutions involved in the project. As a main contribution of the project, a new method for transferring the training curriculum developed will be proposed, as well as good practices to be used in similar dissemination processes.

Partners involved in the project are institutions with experience in vocational education. They comprise the company COMARCH S.A., the University of Málaga, AGH University of Science and Technology, Universidad Politécnica de Cartagena and the Grenoble Institute of Technology. This project will help the partners to establish a stable platform to realise the project and a good basis for future educational projects. Co-operation of private training companies and universities offers the chance to develop the unique potential of IT and educational standards to realise common projects. This assures the creation of new attractive vocational training adapted to the commercial use of the technology.

More information:
http://moteco.kt.agh.edu.pl/index.php

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
http://www.aarit.at/

Consiglio Nazionale delle Ricerche, ISTI-CNR
Area della Ricerca CNR di Pisa,
Via G. Moruzzi 1, 56124 Pisa, Italy
http://www.isti.cnr.it/

Czech Research Consortium
for Informatics and Mathematics
FI MU, Botanicka 68a, CZ-602 00 Brno, Czech Republic
http://www.utia.cas.cz/CRCIM/home.html

Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
http://www.cwi.nl/

Danish Research Association for Informatics and Mathematics
c/o Aalborg University,
Selma Lagerlöfs Vej 300, 9220 Aalborg East, Denmark
http://www.danaim.dk/

Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
http://www.fnr.lu/

FWO
Egmontstraat 5
B-1000 Brussels, Belgium
http://www.fwo.be/

FNRS
rue d'Egmont 5
B-1000 Brussels, Belgium
http://www.fnrs.be/

Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
http://www.ics.forth.gr/

Fraunhofer ICT Group
Friedrichstr. 60
10117 Berlin, Germany
http://www.iuk.fraunhofer.de/

Institut National de Recherche en Informatique
et en Automatique
B.P. 105, F-78153 Le Chesnay, France
http://www.inria.fr/

Irish Universities Association
c/o School of Computing, Dublin City University
Glasnevin, Dublin 9, Ireland
http://ercim.computing.dcu.ie/

Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and
Electrical Engineering, N 7491 Trondheim, Norway
http://www.ntnu.no/

Portuguese ERCIM Grouping
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, nº 378,
4200-465 Porto, Portugal

Polish Research Consortium for Informatics and Mathematics
Wydzia³ Matematyki, Informatyki i Mechaniki,
Uniwersytetu Warszawskiego, ul. Banacha 2, 02-097 Warszawa, Poland
http://www.plercim.pl/

Science and Technology Facilities Council,
Rutherford Appleton Laboratory
Harwell Science and Innovation Campus
Chilton, Didcot, Oxfordshire OX11 0QX, United Kingdom
http://www.scitech.ac.uk/

Spanish Research Consortium for Informatics and Mathematics,
D3301, Facultad de Informática, Universidad Politécnica de Madrid,
Campus de Montegancedo s/n,
28660 Boadilla del Monte, Madrid, Spain,
http://www.sparcim.es/

Swedish Institute of Computer Science
Box 1263,
SE-164 29 Kista, Sweden
http://www.sics.se/

Swiss Association for Research in Information Technology
c/o Professor Daniel Thalmann, EPFL-VRlab,
CH-1015 Lausanne, Switzerland
http://www.sarit.ch/

Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
http://www.sztaki.hu/

Technical Research Centre of Finland
PO Box 1000
FIN-02044 VTT, Finland
http://www.vtt.fi/

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.eu

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 edition*** ☐ ***online edition (email required)***

Name: ....

Organisation/Company: ....

Address: ....

....

Postal Code: ....

City: ....

Country ....

E-mail: ....

You can also subscribe to ERCIM News and order back copies by filling out the form at the ERCIM Web site at http://ercim-news.ercim.eu/