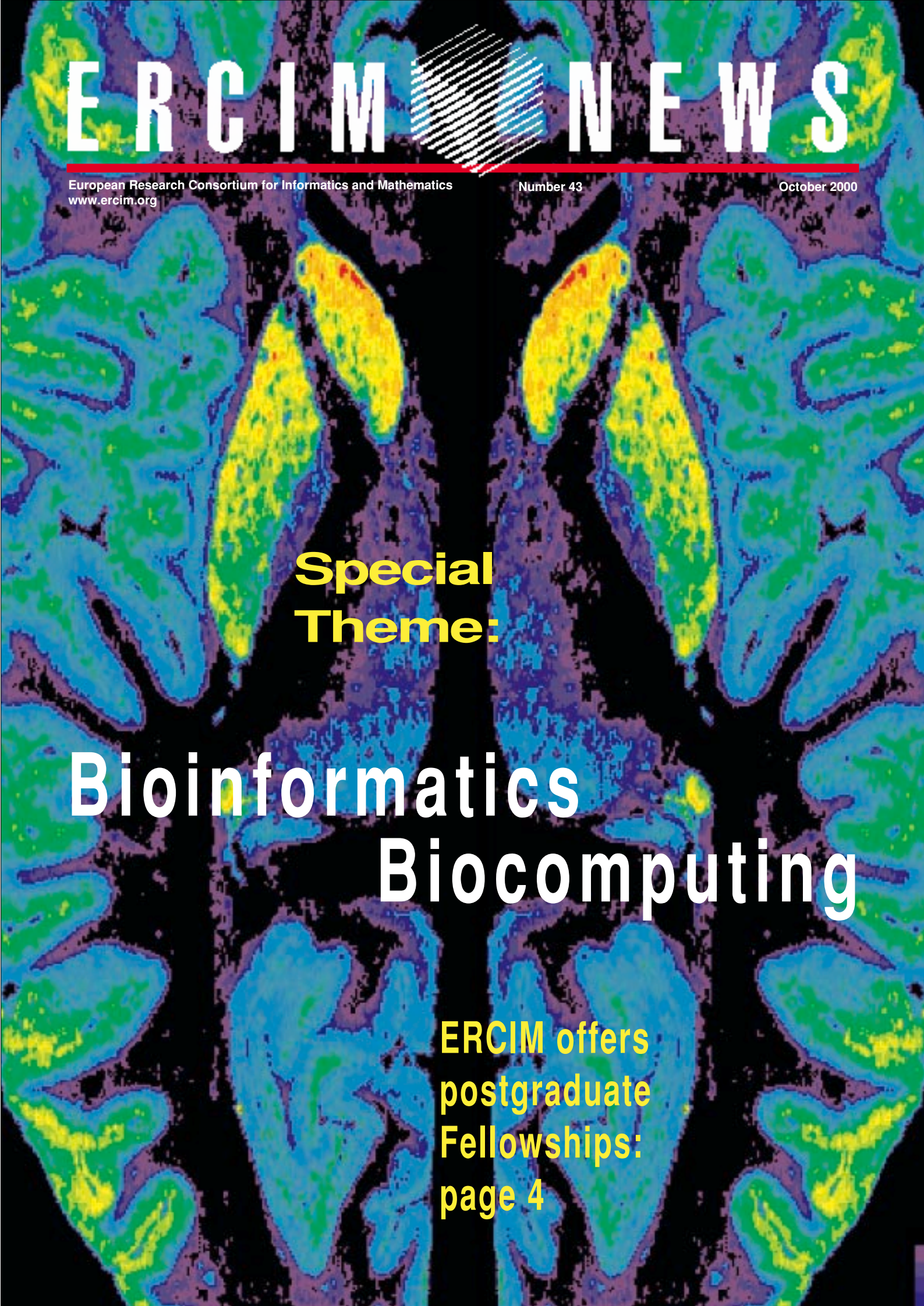
ERCIM NEWS

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

Number 43

October 2000



Special Theme:

Bioinformatics Biocomputing

ERCIM offers postgraduate Fellowships:

page 4

CONTENTS

Cover image: Dopamine-D1 receptors in the human brain, by Håkan Hall, Karolinska Institutet, see article 'Human Brain Informatics – Understanding Causes of Mental Illness' on page 24

Some revolutions in science often come when least you expect them. Others are forced upon us. Bioinformatics is a revolution forced by the extraordinary advances in DNA sequencing technologies, in our understanding of protein structures and by the necessary growth of biological databases. Twenty years ago pioneers such as Doug Brutlag in Stanford and Roger Staden in Cambridge began to use computational methods to analyse the very small DNA sequences then determined. Pioneer efforts were made in 1974 by Bart Barrell and Brian Clarke to catalogue the first few nucleic acid sequences that had been determined. A few years later, in the early 1980's, first the European Molecular Biology Laboratory (EMBL) and then the US National Institutes of Health (NIH) established computerised data libraries for nucleic acid sequences. The first release of the EMBL data library was 585,433-bases; it is 9,678,428,579 on the day I write this, and doubling every 10 months or so.

Michael Ashburner, Joint-Head of the European Bioinformatics Insitute:

"What we so desperately need, if we are going to have any chance of competing with our American cousins over the long term in bioinformatics, genomics and science in general, is a European Science Council with a consistent and science led policy with freedom from political and nationalistic interference."

Bioinformatics is a peculiar trade since, until very recently, most in the field were trained in other fields – computer science, physics, linguistics, genetics, etc. The term will include database curators and algorithmists, software engineers and molecular evolutionists, graph theorists and geneticists. By and large their common characteristic is a desire to understand biology through the organisation and analysis of molecular data, especially those concerned with macromolecular sequence and structure. They rely absolutely on a common infrastructure of public databases and shared software. It has proven, in the USA, Japan and Europe, to be most effective to provide this infrastructure by a mix of major public domain institutions, academic centres of excellence and industrial research. Indeed, such are the economies of scale for both data providers and data users that it has proved to be effective to collect the major data classes, nucleic acid and protein sequence, protein structure co-ordinates, by truly global collaborative efforts.

In Europe the major public domain institute devoted to bioinformatics is the European Bioinformatics Institute, an Outstation of the EMBL. Located adjacent to the Sanger Centre just outside Cambridge, this is the European home of the major international nucleic acid sequence and protein structure databases, as well as the world's premier protein sequence database. Despite the welcome growth of national centres of excellence in bioinformatics in Europe these major infrastructural projects must be supported centrally. The EBI is a major database innovator, eg its proposed ArrayExpress database for microarray data, and software innovator, eg its SRS system. Jointly with the Sanger Centre the EBI produces the highest quality automatic annotation of the emerging human genome sequence (Ensembl).

To the surprise of the EBI, and many others, attempts to fund these activities at any serious level through the programmes of the European Commission were rebuffed in 1999. Under Framework IV the European Commission had funded databases at the EBI; despite an increased funding to the area of 'infrastructure' generally the EBI was judged ineligible for funding under Framework Programme V. Projects internationally regarded as excellent, such as the ArrayExpress database, simply lack funding. The failure of the EC to fund the EBI in 1999 led to a major funding crisis which remains to be resolved for the long term, although the Member States of EMBL have stepped in with emergency funds, and are considering a substantial increase in funding for the longer term.

It is no coincidence that the number of 'start-up' companies in the fields of bioinformatics and genomics in the USA is many times that in Europe. There there is a commitment to funding both national institutions (the budget of the US National Center for Biotechnology Information is three-times that of the EBI) and academic groups. What we so desperately need, if we are going to have any chance of competing with our American cousins over the long term in bioinformatics, genomics and science in general, is a European Science Council with a consistent and science led policy with freedom from political and nationalistic interference. The funding of science through the present mechanisms in place in Brussels is failing both the community and the Community.

Michael Ashburner

ERCIM has launched the 2001/2002 Fellowship Programme

ERCIM offers postdoctoral fellowships in leading European information technology research centres. The Fellowships are of 18 months duration, to be spent in two research centres. Next deadline for applications: 31 October 2000.

The ERCIM Fellowship Programme was established in 1990 to enable young scientists from around the world to perform research at ERCIM institutes. For the 2001/2002 Programme, applications are solicited twice with a deadline of 31 October 2000 and 30 April 2001.

Topics

This year, the ERCIM Fellowship programme focuses on the following topics:

- Multimedia Systems
- Database Research
- Programming Language Technologies
- Constraints Technology and Application
- Control and Systems Theory
- Formal Methods
- Electronic Commerce
- User Interfaces for All
- Environmental Modelling
- Health and Information Technology
- Networking Technologies
- E-Learning
- Web Technology, Research and Application
- Software Systems Validation
- Computer Graphics
- Mathematics in Computer Science
- Robotics
- others.

Objectives

The objective of the Programme is to enable bright young scientists to work collectively on a challenging problem as fellows of an ERCIM insitute. In addition, an ERCIM fellowship helps widen and intensify the network of personal relations and understanding among scientists. The Programme offers the opportunity:

- to improve the knowledge about European research structures and networks
- to become familiar with working conditions in leading European research centres
- to promote co-operation between research groups working in similar areas in different laboratories, through the fellowships.

Selection Procedure

Each application is reviewed by one or more senior scientists in each ERCIM institute. ERCIM representatives will select the candidates taking into account the quality of the applicant, the overlap of interest between applicant and the hosting institution and the available funding.

Conditions

Candidates must:

- have a PhD degree (or equivalent), or be in the last year of the thesis work with an outstanding academic record
- be fluent in English
- be discharged or get deferment from military service
- start the grant before October 2001.

Fellowships are of 18 months duration, spent in two of the ERCIM institutes. ERCIM offers a competitive salary which may vary depending on the country. Costs for travelling to and from the institutes will be paid. In order to encourage the mobility, a member institution will not be eligible to host a candidate of the same nationality.

Links:
Detailed description and online application form: http://www.ercim.org/activity/fellows/

Please contact
Aurélie Richard – ERCIM Office
Tel: +33 4 92 38 50 10
E-mail: aurelie.richard@ercim.org

Poster for the 2001/2002 ERCIM Fellowship Programme.

Irish Government invests over €635 Million in Basic Research

Science Foundation Ireland, the National Foundation for Excellence in Scientific Research, was launched by the Irish Government to establish Ireland as a centre of research excellence in strategic areas relevant to economic development, particularly Biotechnology and Information and Communications Technologies (ICT). The foundation has over €635 million at its disposal.

The Technology Foresight Reports published in 1999 had recommended that the Government establish a major fund to develop Ireland as a centre for world class research excellence in strategic niches of Biotechnology and ICT. As part of its response, the Government approved a Technology Foresight Fund of over €635 million for investment in research in the years 2000-2006. Of this fund, €63 million has been allocated to set up a new research and development institute located in Dublin in partnership with Massachusetts Institute of Technology. The new institute will be known as Media Lab Europe and will specialise in multimedia, digital content and internet technologies.

This Fund is part of a €2.5 billion initiative on R&D that the Irish Government has earmarked for Research, Technology and Innovation (RTI) activities in the National Development Plan 2000-2006. Science Foundation Ireland is responsible for the management, allocation, disbursement and evaluation of expenditure of the Technology Foresight Fund. The Foundation will be set up initially as a sub-Board of Forfás, the National Policy and Advisory Board for Enterprise, Trade, Science, Technology and Innovation.

Speaking at the launch of the first call for Proposals to the Foundation on the 27th of July, 2000, Ireland's Deputy Prime Minister, Mary Harney, said that the Irish Government was keen to establish Ireland as a centre of research excellence in ICT and Biotechnology. "We wish to attract the best scientific brains available in the international research community, particularly in the areas of Biotechnology and ICT, to develop their research in Ireland. The large amount of funding being made available demonstrates the Irish government's commitment to this vitally important project."

Advisory Panels with international experts in Biotechnology and Information and Communications Technologies (ICT) have been set up to advise on the overall strategy of Science Foundation Ireland, including Dr. Gerard van Oortmerssen, chairman of ERCIM, who comments: "The decision by the Irish Government to make a major strategic investment in fundamental research in ICT shows vision and courage and is an example for other European countries. It is fortunate that Ireland recently joined ERCIM. We are looking forward to co-operating with a strong research community in Ireland."

The aim of the first call of the proposals is to identify and fund, at a level of up to €1.3 million per year, a small number of outstanding researchers and their teams who will carry out their work in public research organisations in Ireland. Applications are invited not only from Irish scientists at home and abroad but also from the global research community. Selection will be by an international peer review system. The funding awards will cover the cost of research teams, possibly up to 12 people, over a three to five year period. The SFI Principal Investigator and his/her team will function within a research body in Ireland; either in an Irish University, Institute of Technology or public research organisation. International co-operation will be encouraged.

This initiative will be observed with interest by other European countries, and investment decisions made now will have far-reaching effects for future research in Ireland.

At the launch of the First Call for Proposals on 27th July were: L to R: Mr. Paul Haran, Secretary General, Dept. of Enterprise, Trade and Employment; Ms. Mary Harney, T.D., Deputy Prime Minister and Minister for Enterprise, Trade and Employment; Mr. Noel Treacy, TD., Minister for Science, Technology and Commerce and Mr. John Travers, Chief Executive Officer, Forfás.

Links:
Science Foundation Ireland: http://www.sfi.ie
Media Lab Europe: http://www.mle.ie

Please contact:
Josephine Lynch – Science Foundation Ireland
Tel: +353 1 6073 200
E-mail: info@sfi.ie, Josephine.Lynch@forfas.ie

Bioinformatics: From the Pre-genomic to the Post-genomic Era

by Thomas Lengauer

Computational Biology and Bioinformatics are terms for an interdisciplinary field joining information technology and biology that has skyrocketed in recent years. The field is located at the interface between the two scientific and technological disciplines that can be argued to drive a significant if not the dominating part of contemporary innovation. In the English language, Computational Biology refers mostly to the scientific part of the field, whereas Bioinformatics addresses more the infrastructure part. In other languages (eg German) Bioinformatics covers both aspects of the field.

The goal of this field is to provide computer-based methods for coping with and interpreting the genomic data that are being uncovered in large volumes within the diverse genome sequencing projects and other new experimental technology in molecular biology. The field presents one of the grand challenges of our times. It has a large basic research aspect, since we cannot claim to be close to understanding biological systems on an organism or even cellular level. At the same time, the field is faced with a strong demand for immediate solutions, because the genomic data that are being uncovered encode many biological insights whose deciphering can be the basis for dramatic scientific and economical success. With the pre-genomic era that was characterized by the effort to sequence the human genome just being completed, we are entering the post-genomic era that concentrates on harvesting the fruits hidden in the genomic text. In contrast to the pre-genomic era which, from the announcement of the quest to sequence the human genome to its completion, has lasted less than 15 years, the post-genomic era can be expected to last much longer, probably extending over several generations.

At the basis of the scientific grand challenge in computational biology there are problems in computational biology such as identifying genes in DNA sequences and determining the three-dimensional structure of proteins given the protein sequence (the famed protein folding problem). Other unsolved mysteries include the computational estimation of free energies of biomolecules and molecular complexes in aqueous solution as well as the modeling and simulation of molecular interaction networks inside the cell and between cells. Solving these problems is essential for an accurate and effective analysis of disease processes by computer.

Besides these more 'timeless' scientific problems, there is a significant part of computational biology that is driven by new experimental data provided through the dramatic progress in molecular biology techniques. Starting with genomic sequences, the past few years have provided gene expression data on the basis of ESTs (expressed sequence tags) and DNA microarrays (DNA chips). These data have given rise to a very active new subfield of computational biology called expression data analysis. These data go beyond a generic view on the genome and are able to distiniguish between gene populations in different tissues of the same organism and in different states of cells belonging to the same tissue. For the first time, this affords a cell-wide view of the metabolic and regulatory processes under different conditions. Therefore these data are believed to be an effective basis for new diagnoses and therapies of diseases.

Eventually genes are transformed into proteins inside the cell, and it is mostly the proteins that govern cellular processes. Often proteins are modified after their synthesis. Therefore, a cell-wide analysis of the population of mature proteins is expected to correlate much more closely with cellular processes than the expressed

genes that are measured today. The emerging field of proteomics addresses the analysis of the protein population inside the cell. Technologies such as 2D gels and mass spectrometry offer glimpses into the world of mature proteins and their molecular interactions.

Finally, we are stepping beyond analyzing generic genomes and are asking what genetic differences between individuals of a species are the key for predisposition to certain diseases and effectivity of special drugs. These questions join the fields of molecular biology, genetics, and pharmacy in what is commonly named pharmacogenomics.

Pharmaceutical industry was the first branch of the economy to strongly engage in the new technology combining high-throughput experimentation with bioinformatics analysis. Medicine is following closely. Medical applications step beyond trying to find new drugs on the basis of genomic data. The aim here is to develop more effective diagnostic techniques and to optimize therapies. The first steps to engage computational biology in this quest have already been taken.

While driven by the biological and medical demand, computational biology will also exert a strong impact onto information technology. Since, due to their complexity, we are not able to simulate biological processes on the basis of first principles, we resort to statistical learning and data mining techniques, methods that are at the heart of modern information technology. The mysterious encoding that Nature has afforded for biological signals as well as the enormous data volume present large challenges and are continuing to have large impact on the processes of information technology themselves.

In this theme section, we present 15 scientific progress reports on various aspects of computational biology. We begin with two review papers, one from the biological and one from the pharmaceutical perspective. In three further articles we present progress on solving classical grand challenge problems in computational biology. A section of five papers deals with projects addressing computational biology problems pertaining to current problems in the field. In a section with three papers we discuss medical applications. The last two papers concentrate on the role of information technology contributions, specifically, algorithms and visualization.

This theme section witnesses the activity and dynamics that the field of computational biology and bioinformatics enjoys not only among biologists but also among computer scientists. It is the intensive interdisciplinary cooperation between these two scientific communities that is the motor of progress in this key-technology for the 21st century.

Please contact:
Thomas Lengauer – GMD
Tel: +49 2241 14 2777
E-mail: Thomas.Lengauer@gmd.de

Articles in this section:

Computational Genomics: Making Sense of Complete Genomes

by Anton Enright, Sophia Tsoka and Christos Ouzounis

The current goal of bioinformatics is to take the raw genetic information produced by sequencing projects and make sense of it. The entire genome sequence should reflect the inheritable properties of a given species. At the Computational Genomics Group of the European Bioinformatics Institute (an EMBL outstation) in Cambridge, work is underway to tackle this vast flood of data using both existing and novel technologies for biological discovery.

The recent sequencing of the complete genomes of many species (including a 'draft' human genome) has emphasised the importance of bioinformatics research. Once the DNA sequence of an organism is known, proteins encoded by this sequence are predicted. While some of these proteins are highly similar to well-studied proteins whose functions are known, many will only have similarity to another poorly annotated protein from another genome or worse still, no similarity at all. A major goal of computational genomics is to accurately predict the function of all proteins encoded by a genome, and if possible determine how each of these proteins interacts with other proteins in that organism. Using a combination of sequence analysis, novel algorithm development and data-mining techniques the Compuational Genomics Group (CGG) is targeting research on the following fields.

Automatic Genome Annotation

Accurately annotating the proteins encoded by complete genomes in a comprehensive and reproducible manner is important. Large scale sequence analysis necessitates the use of rapid computational methods for functional characterisation of molecular components. GeneQuiz is an integrated system for the automated analysis of complete genomes that is used to derive protein function for each gene from raw sequence information in a manner comparable to a human expert. It employs a variety of similarity search and analysis methods that entail the use of up-to-date protein and DNA databases and creates a compact summary of findings that can be accessed through a Web-based browser. The system applies an 'expert system' module to assess the quality of the results and assign function to each gene.

Assigning Proteins into Families

Clustering protein sequences by similarity into families is another important aspect of bioinformatics research. Many available clustering techniques fail to accurately cluster proteins with multiple domains into families. Multi-domain proteins generally perform at least two functions that are not necessarily related, and so ideally should belong in multiple families. To this end we have developed a novel algorithm called GeneRAGE. The GeneRAGE algorithm employs a fast sequence similarity search algorithm such as BLAST and represents similarity information between proteins as a binary matrix. This matrix is then processed and passed through successive rounds of the Smith-Waterman dynamic programming algorithm, to detect inconsistencies which

Figure 1: The GeneQuiz entry page for the S.cerevisiae genome.

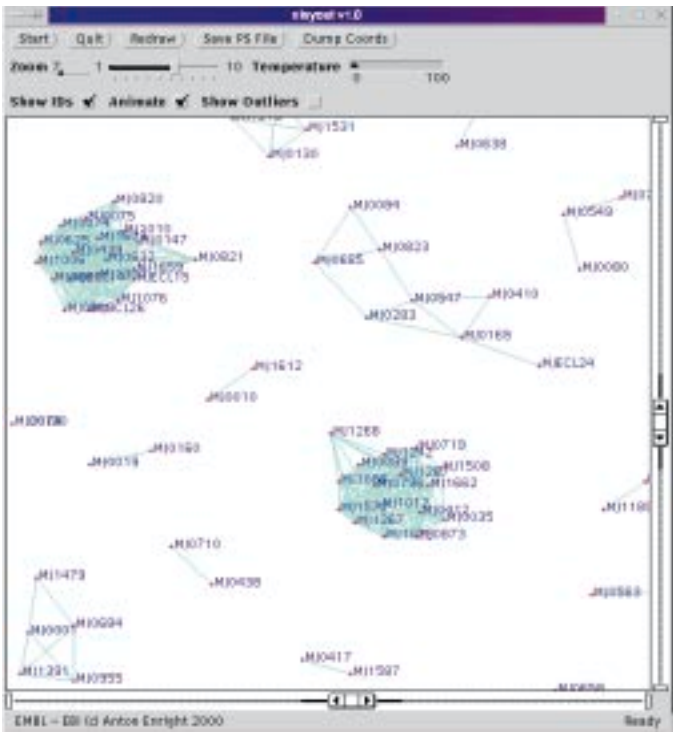


Figure 2: Protein families in the Methanococcus jannashii genome displayed using the X-layout algorithm.

represent false-positive or false-negative similarity assignments. The resulting clusters represent protein families accurately and also contain information regarding the domain structure of multi-domain proteins. A visualization program called xlayout based on the Fruchterman-Rheingold graph-layout optimisation algorithm has also been developed for displaying these complex similarity relationships.

Prediction of Protein Interaction

Another novel algorithm developed in the CGG group is the DifFuse algorithm. This algorithm is based on the hypothesis that there is a selective advantage for proteins performing related functions to fuse together during the course of evolution (eg different steps in the same metabolic pathway). The DifFuse algorithm can detect a fused protein in one genome based its similarity to complementary pair of unfused proteins in another genome. The detection of these fused proteins allows one to predict either functional association or direct physical interaction of the un-fused proteins. This algorithm is related to GeneRAGE in the sense that the fusion detection process is similar to the multi-domain detection step described above. This algorithm can be applied to many genomes for large-scale detection of protein interactions.

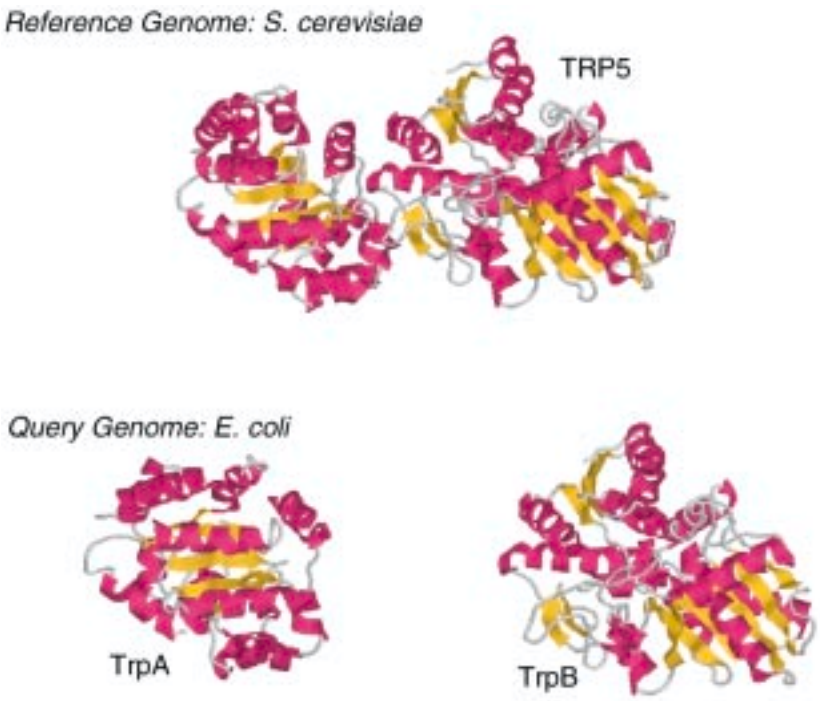


Figure 3: Gene Fusion – The TRP5 tryptophan synthase protein in S.cerevisiae is a fusion of two single domains, such as TrpA and TrpB in E. coli.

Knowledge-Base Development

Databases in molecular biology and bioinformatics are generally poorly structured, many existing as flat text files. In order to get the most out of complex biological databases these data need to be represented in a format suitable for complex information extraction through a simple querying system and also ensure data integrity. An ontology is an exact specification of a data model that can be used to generate such a 'knowledge' base. We have developed an ontology for representation of genomic data which is used to built a database called GenePOOL incorporating these concepts. This system stores computationally-derived information such as functional classifications, protein families and reaction information. Database analysis is performed through flexible and complex queries using LISP that are simply not possible through any other public molecular biology database. Similarly, we have also developed a standard for genome annotation called GATOS (Genome AnoTattiOn System) which is used as a data exchange format. Work is under development to incorporate an XML-based standard called XOL (XML Ontology Language).

Text-Analysis and Data-Mining

There is already a vast amount of data available in the abstracts of published biological text. The MEDLINE database contains abstracts for over 9 million biological papers published worldwide since 1966. However, these data are not represented in a format suitable for large-scale information extraction. We have developed an algorithm called TextQuest which can perform document clustering of MEDLINE abstracts. TextQuest uses an approach that restructures these biological abstracts and obtains the optimal number of terms that can associate large numbers of abstracts into meaningful groups. Using a term-weighting system based on the TF.IDF family of metrics and term frequency data from the British National Corpus, we select words that are biologically significant from abstracts and add them to a so-called go-list. Abstracts are clustered using an unsupervised machine learning approach, according to their sharing of words contained in the go-list. The xlayout algorithm (see above) is then used to display the clustering results. The resulting document clusters accurately represent sets of abstracts referring to the same biological process or pathway. TextQuest has been applied to the development of the dorsal-ventral axis in the fruit-fly Drosophila melanogaster and has produced meaningful clusters relating to different aspects of this developmental process.

Links:
http://www.ebi.ac.uk/research/cgg/

Please Contact:
Christos A. Ouzounis – European Molecular Biology Laboratory, European Bioinformatics Institute
Tel: +44 1223 49 46 53
E-mail: ouzounis@ebi.ac.uk

Searching for New Drugs in Virtual Molecule Databases

by Matthias Rarey and Thomas Lengauer

The rapid progress in sequencing the human genome opens the possibility for the near future to understand many diseases better on molecular level and to obtain so-called target proteins for pharmaceutical research. If such a target protein is identified, the search for those molecules begins which influence the protein's activity specifically and which are therefore considered to be potential drugs against the disease. At GMD, approaches to the computer-based search for new drugs are being developed (virtual screening) which have already been used by industry in parts.

Searching for New Lead Structures

The development process of a new medicine can be divided into three phases. In the first phase, the search for target proteins, the disease must be understood on molecular-biological level as far as to know individual proteins and their importance to the symptoms. Proteins are the essential functional units in our organism and can perform various tasks ranging from the chemical transformation of materials up to the transportation of information. The function is always linked with the specific binding of other molecules. As early as 100 years ago, Emil Fischer recognised the lock-and-key principle: Molecules that bind to each other are complementary to each other both spatially and chemically, just as only a specific key fits a given lock (see Figure 1). If a relationship between the suppression (or reinforcement) of a protein function and the symptoms is recognised, the protein is declared to be a target protein. In the second phase, the actual drug is developed. The aim is to detect a molecule that binds to the target protein, on the one hand, thus hindering its function and that, on the other, has got further properties that are demanded for drugs, for example, that it is well tolerated and accumulates in high concentration at the place of action. The first step is the search for a lead structure - a molecule that binds well to the target protein and serves as a first proposal for the drug. Ideally, the lead structure binds very well to the target protein and can be modified such that the resulting molecule is suitable as a drug. In the third phase, the drug is transformed into a medicine and is tested in several steps to see if it is well tolerated and efficient. The present paper is to discuss the first step, ie the computer-based methods of searching for new lead structures.

New Approaches to Screening Molecule Databases

The methods of searching for drug molecules can be classified according to two criteria: the existence of a three-dimensional structural model of the target protein and the size of the data set to be searched. If a structural model of the protein is available, it can be used directly to search for suitable drugs (structure-based virtual screening); ie we search for a key fitting a given lock. If a structural model is missing, the similarity to molecules that bind to the target protein is used as a measure for the suitability as a drug (similarity-based virtual screening). Here we use a given key to search for fitting keys without knowing the lock. In the end, the size of the data set to be searched decides on the amount of time to be put into the analysis of an individual molecule. The size ranges from a few hundred already preselected molecules via large databases of several millions of molecules to virtual combinatorial molecule libraries theoretically allowing to synthese of up to billions of molecules from some hundred molecule building blocks.

The key problem in structure-based virtual screening is the prediction of the relative orientation of the target protein and a potential drug molecule, the so-called docking problem. For solving this problem we have developed the software tool FlexX [1] in co-operation with Merck KGaA, Darmstadt, and BASF AG, Ludwigshafen. On the one hand, the difficulty of the docking problem arises from the estimation of the free energy of a molecular complex in aqueous solution and, on the other, from the flexibility of the molecules involved. While a sufficient description of the flexibility of the protein presumably will not be possible even in the near future, the more important flexibility of the ligand is considered during a FlexX prediction. In a set of benchmarks tests, FlexX is able to predict about 70 percent of the protein-ligand complexes sufficiently similar to the experimental structure. With about 90 seconds computing time per prediction, the software belongs to the fastest docking tools currently available. FlexX has been marketed since 1998 and is currently being used by about 100 pharmaceutical companies, universities and research institutes.

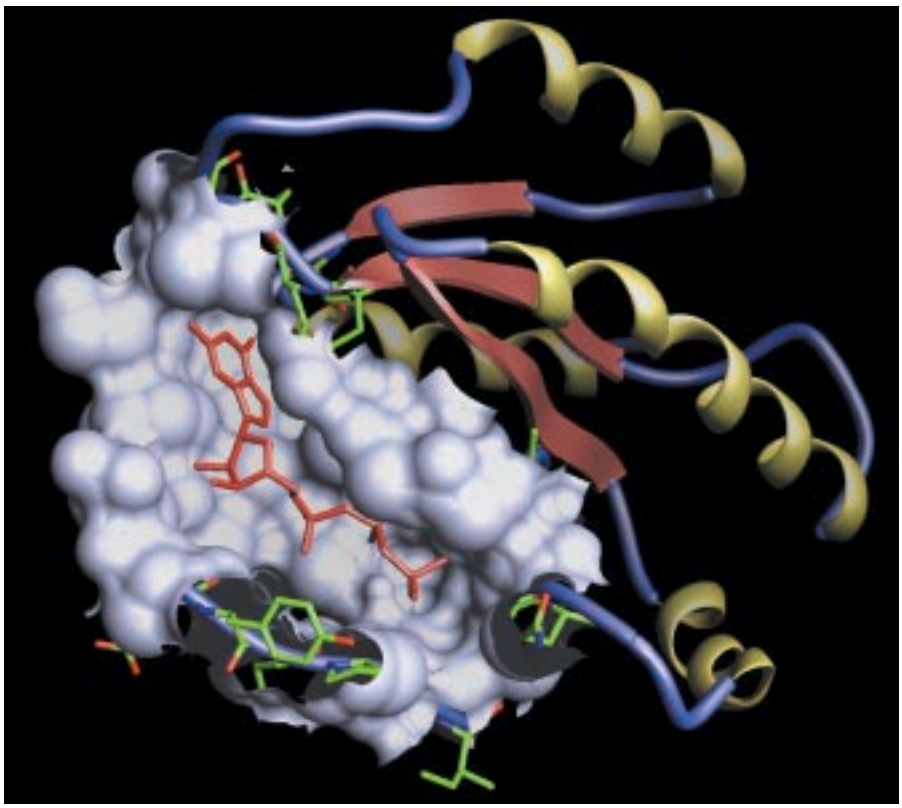
If the three-dimensional structure of the target protein is not available, similarity-based virtual screening methods are applied to molecules with known binding properties, called the reference molecule. The main problem here is the structural alignment problem which is closely related to the docking problem described above. Here, we have to superimpose a potential drug molecule with the reference molecule so that a maximum of functional groups are oriented such that they can form the same interactions with the protein. Along the lines of FlexX, we have developed the software tool FlexS [2,3] for the prediction of structural alignments with approximately the same performance with respect to computing time and prediction quality.

If very large data sets are to be searched for similar molecules, the speed of the alignment-based screening does not suffice yet. The aim is to have comparison operations whose computation takes by far less than one second. Today linear descriptors (bit strings or integral vectors) are usually applied to solve this problem. They store the occurrence or absence of characteristic properties of the molecules such as specific chemical fragments or short paths in the molecule. Once such a descriptor has been determined, the linear

structure enables a very fast comparison. A considerable disadvantage is, however, that the overall structure of the molecule is represented only inadequately and an exact match between the fragments is frequently required for the recognition of similarities. As an alternative, we have developed a new descriptor, the feature tree [4], in co-operation with SmithKline Beecham Pharmaceuticals, King of Prussia (USA) . Unlike the representation using linear descriptors, in this approach a molecule is represented by a tree structure representing the major building blocks of the molecules. If two molecules are to be compared with each other, the task is to find first an assignment of the building blocks of the molecules which might be able to occupy the same regions of the active site upon binding. With the aid of a time-efficient implementation, average comparison times of less than a tenth second can be achieved. This allows 400.000 molecule comparisons to be carried out overnight within a single computation on a single processor. Applying the new descriptor to benchmark data sets, we could show that, in many cases, an increase of the rate of active molecules in a selected subset of the data set is achieved if compared with the standard method.

Designing New Medicines with the Computer?

Unlike many design problems from the field of engineering, for example, the design of complex plants, machines or microchips, the underlying models are still very inaccurate for solving problems in the biochemical environment. Important quantities such as the binding energy of protein-ligand complexes can be predicted only with high error rates. In addition, not only on the interaction of the drug with the target protein is of importance to the development of medicines. The influence of the drug on the whole organism rather is to be examined. Even in the near future, it will not be possible to answer many questions arising in this context accurately by means of the computer due to their complexity: Is the drug absorbed in the organism? Does it produce the desired effect? Which side effects are experienced or is it possibly even toxic? Therefore a medicine originating directly from the computer will not be available neither in the near nor in the distant future. Nevertheless, the importance of the computer increases in drug research. The reason is the very great number of potential molecules which come into consideration as a drug for a specific application. The computer allows a reasonable pre-selection and molecule libraries can be optimised for specific characteristics before synthesising. On the basis of experiments, the computer is able to generate hypotheses which enable in turn better planning of further experiments. In these domains, the computer has already proven to be a tool without which pharmaceutical research cannot be imagined anymore.



Complex between the protein HIV-protease (shown as blue ribbon) and a known inhibitor (shown in red). HIV-protease plays a major role in the reproduction cycle of the HIV virus. Inhibitors like the one shown here are used in the treatment of AIDS.

Links:
http://www.gmd.de/SCAI/

Please contact:
Matthias Rarey – GMD
Tel: +49 2241 14 2476
E-mail: matthias.rarey@gmd.de

A High Performance Computing Network for Protein Conformation Simulations

by Marco Pellegrini

The Institute for Computational Mathematics (IMC-CNR), Pisa, with the collaboration of Prof. Alberto Segre of University of Iowa, is now combining several core areas of expertise (sequential, distributed and parallel algorithms for numerical liner algebra and continuous optimization, methods for efficient computation of electrostatic fields) in a project on Protein Conformation Simulation.

The need for better methodologies to simulate biological molecular systems has not been satisfied by the increased computing power of workstations available nowadays. Instead such increased speed has pushed research into new areas and increasingly complex simulations. A clear example is the problem of protein folding. This problem is at the core of the technology of rational drug design and its impact on future society cannot be underestimated. In a nutshell, the problem is that of determining the 3-dimensional structure of a protein (and especially its biologically active sites) starting from its known DNA encoding. There are favourable opportunities for approaches employing a spectrum of competences, from strictly biological and biochemical to mathematical modeling and computer science, with special openings for the design of efficient algorithms.

Protein Folding requires searching an optimal configuration (minimizing the energy) among exponentially large spaces of possible configurations with a number of degrees of freedom ranging from hundreds to a few thousands. To cope with this challenge, a double action is required. On one hand, the use of High Performance Computing Networks permits the exploitation of the intrinsic parallelism which is present in many aspects of the problem. On the other hand, sophisticated and efficient algorithms are needed to exploit fully the deep mathematical properties of the physics involved. Brute force methods are at a loss here and the way is open for effective sampling methodologies, algorithms for combinatorial and continuous optimization, and strategies for searching over (hyper)-surfaces with multiple local minima.

An image obtained with RASMOL 2.6 of Bovine HYDROLASE (SERINE PROTEINASE).

Innovative techniques will be used in order to push forward the state of the art: a new distributed paradigm (Nagging and Partitioning), techniques from Computational Geometry (hierarchical representations) and innovative algorithms for long range interactions.

Nagging and Partitioning

As observed above, the protein folding problem is reduced to a searching problem in a vast space of conformations. A classical paradigm for finding the optimum over a complex search space is that of 'partitioning'. A master process splits the search space among several slave processes and selects the best one from the solutions returned by the slaves. This approach is valid when it is relatively easy to split the workload evenly among the slave processes and the searching strategy of each slave is already optimized. The total execution time is determined by the slowest of the slaves and, when any slave is faulty, the computation is either blocked or returns a sub-optimal solution. A different and complementary approach is that of 'nagging'. Here slaves operate on the same search space, however each slave uses a different search strategy. The total time is determined by the most efficient slave and the presence of a faulty slave does not block the computation. Moreover, in a branch and bound overall philosophy, it has been shown that the partial solutions of any processor help to speed up the search of others. Searching for the optimal energy conformation is a complex task where parallelism is present at many different levels, so that neither pure nagging nor pure partitioning is the best choice for all of them. A mixed

strategy that uses both is more adaptive and yields better chances of success.

Hierarchical Representations

Techniques for representing and manipulating hierarchies of 3-dimensional objects have been developed in computational geometry for the purpose of speeding up visibility computations and collision detection and could be adapted to the folding problem. One of the main sub-problems in finding admissible configurations is to determine the existence or absence of steric clashes among single atoms in the model. This is a problem similar to that of collision detection for complex 3d models of macroscopic objects in robotics and geometric modeling. A popular technique is that of inclusion hierarchies. The hierarchy is organized as a tree and the root is associated with a bounding volume (eg an axis parallel box) enclosing the molecule. Recursively, we split the molecule into two groups of atoms and build the corresponding bounding boxes. The process stops at the leaves of the tree, each of which correspond to a single atom. Such a representation speeds up steric testing since it quickly rules out many pairs of atoms that are distant in the tree hierarchy. In general many such trees may be built with the same input and the aim is to obtain good properties such as minimizing the total surface or volume of the bounding boxes. It is interesting to note that such trees are more flexible and use less storage than uniform grid decompositions and even oct-tree data structures.

Electrostatic Long Range Interactions

A line of research at IMC has found interesting connections between electrostatic fields and integral geometric formulae leading to new efficient and robust algorithms for computing electrostatic forces. In particular, for 3-dimensional continuous distributions of charge, there is a representation without analytic singularities for which an adaptive Gaussian quadrature algorithm converges with exponential speed. Such recent techniques benefit from the calculation of molecular energy since they obtain a good approximation without considering all pairs of atoms thus avoiding quadratic complexity growth.

Please contact:
Marco Pellegrini – IMC-CNR
Tel: +39 050 315 2410
E-mail: pellegrini@imc.pi.cnr.it
http://www.imc.pi.cnr.it/~pellegrini

Ab Initio Methods for Protein Structure Prediction: A New Technique based on Ramachandran Plots

by Anna Bernasconi and Alberto M. Segre

A new technique for ab initio protein structure prediction, based on Ramachandran plots, is currently being studied at the University of Iowa, under the guidance of Prof. Alberto M. Segre, in collaboration with the Institute for Computational Mathematics, IMC-CNR, Pisa.

One of the most important open problems in molecular biology is the prediction of the spatial conformation of a protein from its primary structure, ie from its sequence of amino acids. The classical methods for structure analysis of proteins are X-ray crystallography and nuclear magnetic resonance (NMR). Unfortunately, these techniques are expensive and can take a long time (sometimes more than a year). On the other hand, the sequencing of proteins is relatively fast, simple, and inexpensive. As a result, there is a large gap between the number of known protein sequences and the number of known three-dimensional protein structures. This gap has grown over the past decade (and is expected to keep growing) as a result of the various genome projects worldwide. Thus, computational methods which may give some indication of structure and/or function of proteins are becoming increasingly important. Unfortunately, since it was discovered that proteins are capable of folding into their unique native state without any additional genetic mechanisms, over 25 years of effort has been expended on the determination of the three-dimensional structure from the sequence alone, without further experimental data. Despite the amount of effort, the protein folding problem remains largely unsolved and is therefore one of the most fundamental unsolved problems in computational molecular biology today.

How can the native state of a protein be predicted (either the exact or the approximate overall fold)? There are three major approaches to this problem: 'comparative modelling', 'threading', and 'ab initio prediction'. Comparative modelling exploits the fact that evolutionarily related proteins with similar sequences, as measured by the percentage of identical residues at each position based on an optimal structural superposition, often have similar structures. For example, two sequences that have just 25% sequence identity usually have the same overall fold. Threading methods compare a target sequence against a library of structural templates, producing a list of scores. The scores are then ranked and the fold with the best score is assumed to be the one adopted by the sequence. Finally, the ab initio prediction methods consist in modelling all the energetics involved in the process of folding, and then in finding the structure with lowest free energy. This approach is based on the 'thermodynamic hypothesis', which states that the native structure of a protein is the one for which the free energy achieves the global

minimum. While ab initio prediction is clearly the most difficult, it is arguably the most useful approach.

There are two components to ab initio prediction: devising a scoring (ie, energy) function that can distinguish between correct (native or native-like) structures from incorrect ones, and a search method to explore the conformational space. In many methods, the two components are coupled together such that a search function drives, and is driven by, the scoring function to find native-like structures. Unfortunately, this direct approach is not really useful in practice, both due to the difficulty of formulating an adequate scoring function and to the formidable computational effort required to solve it. To see why this is so, note that any fully-descriptive energy function must consider interactions between all pairs of atoms in the polypeptide chain, and the number of such pairs grows exponentially with the number of amino acids in the protein. To make matters worse, a full model would also have to contend with vitally important interactions between the protein's atoms and the environment, the so-called 'hydrophobic effect'. Thus, in order to make the computation practical, simplifying assumptions must necessarily be made.

Different computational approaches to the problem differ as to which assumptions are made. A possible approach, based on the discretization of the conformational space, is that of deriving a protein-centric lattice, by allowing the backbone torsion angles, phi, psi, and omega, to take only a discrete set of values for each different residue type. Under biological conditions, the bond lengths and bond angles are fairly rigid. Therefore, the internal torsion angles along the protein backbone determine the main features of the final geometric shape of the folded protein. Furthermore, one can assume that each of the torsion angles is restricted to a small, finite set of values for each different residue type. As a matter of fact, not all torsion angles are created equally. While they may feasibly take any value from -180 to 180 degrees, in nature all these values do not occur with uniform probability. This is due to the geometric constraints from neighboring atoms, which dramatically restrict the commonly occurring legal values for the torsion angles. In particular, the peptide bond is rigid (and shorter than expected) because of the partial double-bond character of the CO-NH bond. Hence the torsion angle omega around this bond generally occurs in only two conformations: 'cis' (omega about 0 degrees) and 'trans' (omega about 180 degrees), with the trans conformation being by far the more common. Moreover the other two torsion angles, phi and psi, are highly constrained, as noted by G. N. Ramachandran (1968). The (paired) values allowed for them can be selected using clustering algorithms operating in a Ramachandran plot space constructed from the protein database (http://www.rcsb.org/pdb), while discrete values for omega can be set to {0,180}. This defines a space of $(2k)^n$ possible conformations for a protein with n amino acids (assuming each phi and psi pair is allowed to assume k distinct values). In this way, the conformational space of proteins is discretized in a protein-centric fashion. An approximate folding is then found by searching this reduced, discrete space, guided by some scoring function. Of course, this discrete search process is an exponential one, meaning that in its most naive form it is impractical for all but the smallest proteins. Thus, the search should be made more palatable by incorporating a number of search pruning and reduction techniques, and/or by exploring the discrete space in parallel. From this naive folding, appropriate constraints (based on atomic bonds present, their bond lengths, and any Van der Waals or sulfide-sulfide interactions in the naive folding) are derived for an interior-point optimization process, which adjusts atomic positions and computes an energy value for this conformation. The computed energy value is then passed back to the discrete space and used to tune the scoring function for further additional parallel pruning.

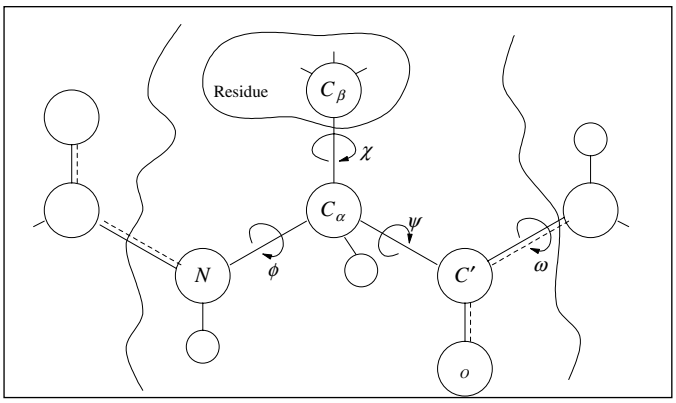


Figure 1: Backbone torsion angles of a protein.

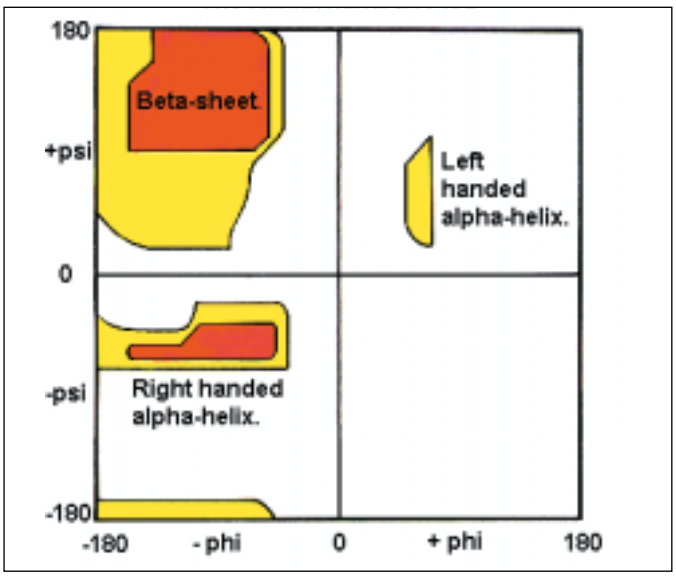


Figure 2: The Ramachandran Plot.

Please contact:
Anna Bernasconi – IMC-CNR
Tel. +39 050 3152411
E-mail: bernasconi@imc.pi.cnr
or Alberto M. Segre – University of Iowa
E-mail: alberto-segre@uiowa.edu

Phylogenetic Tree Reconciliation: the Species/Gene Tree Problem

by Jean-François Dufayard, Laurent Duret and François Rechenmann

An algorithm to find gene duplications in phylogenetic trees in order to improve gene function inferences has been developed in a collaboration between the the Helix team from INRIA Rhône-Alpes and the 'biométrie moléculaire, évolution et structure des génomes' team from the UMR 'biométrie, biologie évolutive de Lyon'. The algorithm and its software is applicable to realistic data, especially n-ary species tree and unrooted phylogenetic tree. The algorithm also takes branch lengths into account.

With appropriate algorithms, it is possible to deduce species history studying genes sequences. Genes are indeed subject to mutations during the evolution process, and hence the corresponding (homologous) sequences in different species differ from each other. A tree can be built from the sequences comparison, relating genes and species history: a phylogenetic tree. Sometimes a phylogenetic tree disagrees with the species tree (constructed for example from anatomical and paleontological considerations). These differences can be explained by a gene being duplicated in a genome, and each copy having its own history. Consequently, a node in a phylogenetic tree can be the division of an ancestral species into two others, as well as a gene duplication. More precisely, in a family of homologous genes, paralogous genes have to be distinguished from orthologous genes. Two genes are orthologous if the divergence from their last common ancestor results from a speciation event, while they are paralogous if the divergence results from a duplication event.

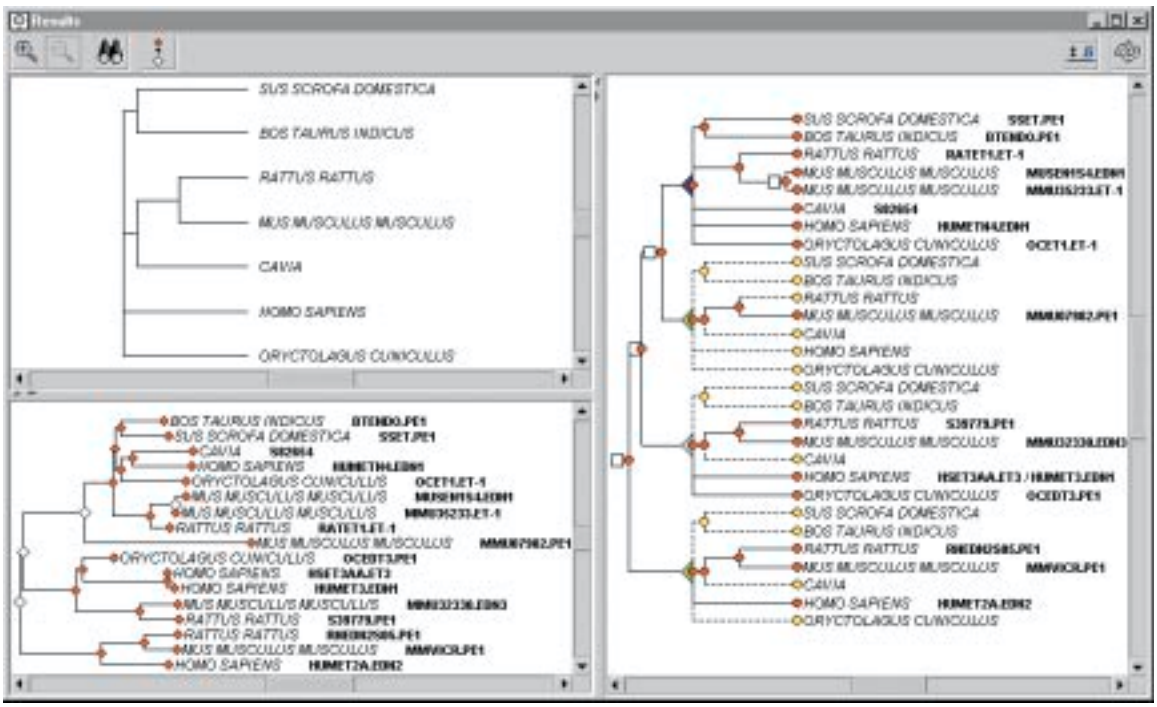
It is essential to make the distinction because two paralogous genes are less likely to have preserved the same function than two orthologues. Therefore, if one wants to predict gene function by homology between different species, it is necessary to check whether genes are orthologous or paralogous to increase the accuracy of the prediction.

An algorithm has been developed which can deduce this information by comparing gene trees with the taxonomy of different species. Currently, the algorithm is applicable to gene families issued of vertebrates. It can be applied to realistic data: species trees may not necessarily be binary, and the tree structures are compared as well as their branch lengths. Finally, phylogenetic trees can be unrooted: the number of duplications is a good criterion to make a choice, and with this method the algorithm is able to root phylogenetic trees.

Software has been developed to use the algorithm. It has been written in JAVA 1.2, and the graphical interface permits an easy application to realistic data. An exhaustive species tree can be easily seen and edited (tested with more than 10,000 leaves). Results can be modified and saved.

Links:
Action Helix:
http://www.inrialpes.fr/helix

Please contact:
Jean-François Dufayard – INRIA Rhône-Alpes
Tel: +33 4 76 61 53 72
E-mail: Jean-Francois.Dufayard@inrialpes.fr



Reconciliation of a phylogenetic tree (lower-left corner) with the corresponding species tree (upper-left corner). Reconciliation produces a reconciled tree (right) which describes both genes and species history. It allows to deduce the location of gene duplications (white squares) which are the only information needed to distinguish orthologous from paralogous genes.

Identification of Drug Target Proteins

by Alexander Zien, Robert Küffner, Theo Mevissen, Ralf Zimmer and Thomas Lengauer

As ever more knowledge is accumulated on various aspects of the molecular machinery underlying the biological processes within organisms, including the human, the question how to exploit this knowledge to combat diseases becomes increasingly urgent. At GMD scientists work to utilize protein structures, expression profiles as well as metabolic and regulatory networks in the search for target proteins for pharmaceutical applications.

Huge amounts of heterogeneous data are pouring out of the biological labs into molecular biology databases. Most popular are the sequencing projects that decipher complete genomes and uncover their protein complements. Many more projects are under way, aiming e.g. at resolving the yet unknown protein folds or at collecting human single nucleotide polymorphisms. In a less coordinated way, many labs measure gene expression levels inside cells in numerous cell states. Last but not least there is an enormous amount of unstructured information hidden in the plethora of scientific publications, as is documented in PubMed, for instance. Each of these sources of data provides valuable clues to researchers that are interested in molecular biological problems. The project TargId at GMD SCAI focuses on methods to address the arguably most urgent problem: the elucidation of the origins and mechanisms of human diseases, culminating in the identification of potential drug target proteins.

TargId responds to the need for bioinformatics support for this task. The goal of the project is to develop methods that extract useful knowledge from the raw data and help to focus on the relevant items of data. The most sophisticated aspect is the generation of new insights through the combination of information from different sources. Currently, our TargId methodology builds on three main pillars: protein structure prediction, expression data analysis and metabolic/regulatory pathway modeling.

Protein Structure Prediction

Knowledge on the three-dimensional structure (fold) of a protein provides clues on its function and aids in the search for inhibitors and other drugs. Threading is an approach to structure prediction which essentially assesses the compatibility of the given protein sequence to a known fold by aligning the sequence onto the known protein structure. Thus, threading utilizes the available knowledge directly in the form of the structures that are already experimentally resolved. Another advantage of threading in comparison to ab initio methods is the low demand of computing time. This is especially true for 123D, a threader developed at GMD that models pairwise amino acid contacts in a way that allows alignment computation by fast dynamic programming. The objective function for the optimization includes potentials accounting for chemical environments within the protein structure, and membership in secondary structure elements as well as amino acid substitution scores and gap penalties, all carefully balanced by a systematic procedure. A second threading program programmed at GMD, called RDP, utilizes more computation time than 123D in order to optimize full pair interaction contact potentials to yield refined alignments.

Expression Data Analysis

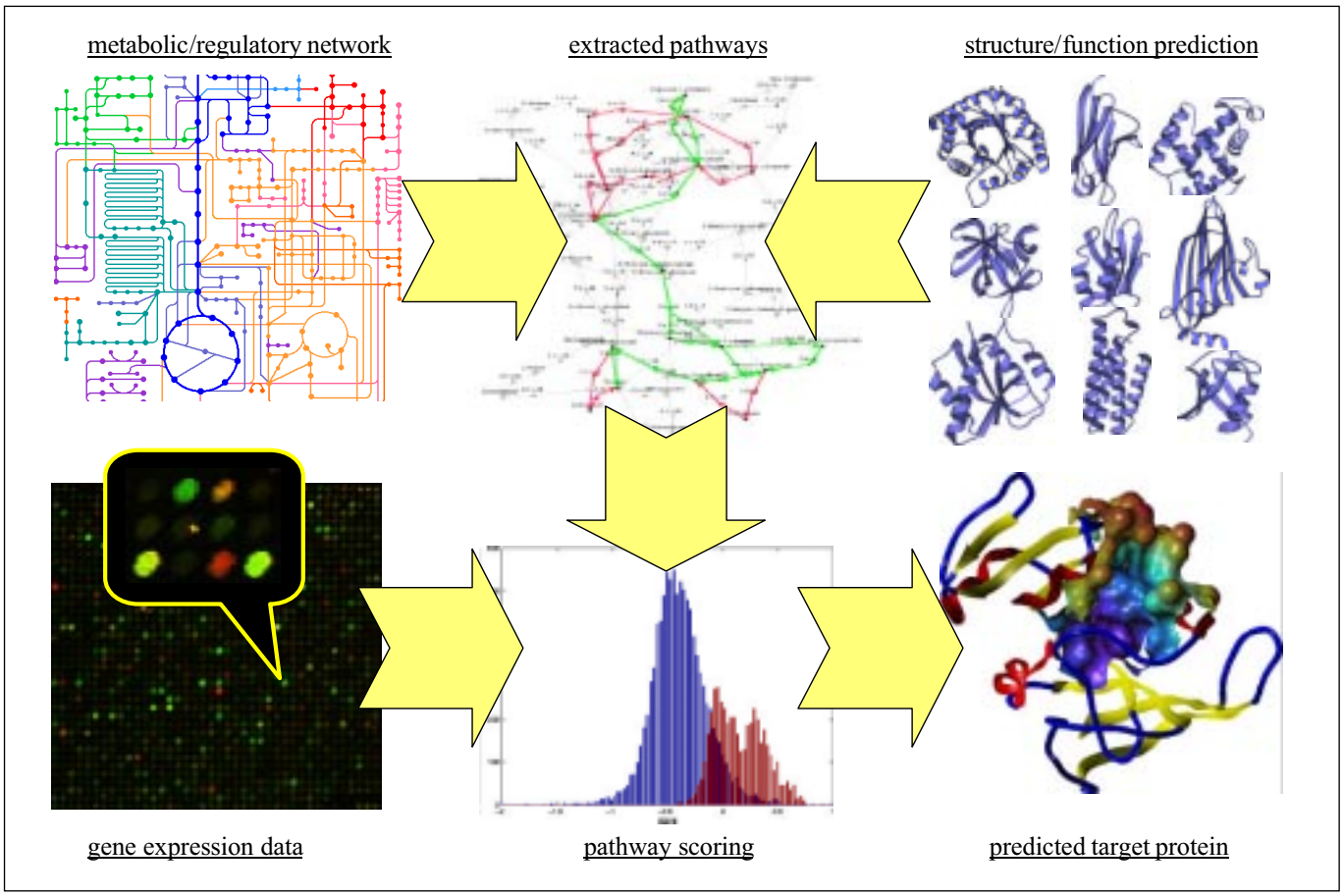
Data on expressed genes comes in several different flavors. The historically first method is the generation of ESTs (expressed sequence tags), i.e. low-quality sequence segments of mRNA, either proportional to its cellular abundance or enriched for rare messages. While ESTs are superseded by more modern methods for the purpose of expression level measurements, they are still valuable for finding new genes and resolving gene structures in the genomic DNA. Thus, we have implemented a variant of 123D that threads ESTs directly onto protein structures, thereby translating nucleotide codons into amino acids on the fly. This program, called EST123D, is useful for proteins that are not yet characterized other than by ESTs.

Nowadays, gene expression levels are most frequently obtained by DNA chips, micro-arrays or SAGE (serial analysis of gene expression). These technologies are designed for high throughput and are already capable of monitoring the complete gene inventory of small organisms (or a large part of all human genes). We apply statistics and machine learning techniques in order to normalize the resulting raw measurement data, to identify differentially regulated genes and to clusters of cell states. Subsequently, we apply statistics and machine learning techniques in order to identify differentially regulated genes, clusters of cell states etc.

Pathway Modeling

While a large part of the current effort is focused on inferring high-level structures from gene expression data, much is already known on the underlying genetic networks. Several databases that are available on the internet document metabolic relations in machine-readable form. The situation is worse for regulatory pathways; most of this knowledge is still hidden in the literature. Consequently, we have implemented methods that extract additional protein relations from article abstracts and model them as Petri nets. The resulting graphs can be restricted to species, tissues, diseases or other areas of interest. Tools are under development for viewing and editing using standard graph and Petri net packages. The generated networks can provide overviews that cross the boundaries of competence fields of human experts.

Further means are necessary to allow for more detailed analyses. We can automatically extract pathways from networks that are far too large and complicated to lend themselves to easy interpretation. Pathways are biologically meaningful subgraphs, e.g. signaling



In the TargId project, new bioinformatics methods combine heterogeneous information in the search for drug target proteins.

cascades or metabolic pathways that account for supply and consumption of any intermediate metabolites. Another method conceived in TargId, called DMD (differential metabolic display), allows for comparing different systems (organisms, tissues, etc.) on the level of complete pathways rather than mere interactions.

Bringing it All Together ...

Each of the methods described above can provide valuable clues pointing to target proteins. But the crux lies in their clever combination, interconnecting data from different sources. In recent work, we have shown that in real life situations clustering alone may not be able to reconstruct pathways from gene expression data. Instead of searching for meaning in clusters, we invented an approach that proceeds inversely: First, a set of pathways is extracted from a protein/gene network, using the methods described above. Then, these pathways are scored with respect to gene expression data. The restriction to pathways prevents us from considering unreasonable groupings of proteins, while it still allows for incorporating and testing hypotheses. E.g., pathways can be constructed from interactions that are observed in different tissues or species. The expression data provide an orthogonal view on these interactions and can thus be used to validate the hypotheses.

Structure prediction can aid in this process at several stages. First, uncharacterized proteins can tentatively be embedded into known networks based on predicted structure and function. Second, structural information can be integrated into the pathway scoring function. Finally, when a target protein is identified, its structure will be of utmost interest for further investigations.

It can be imagined that target finding can gain from broadening the basis for the search to also include, e.g., phylogenetic profiles, post-translational modifications, genome organization or polymorphisms. As these fields are still young and in need of further progress, it is clear that holistic target finding is only in its infancy.

Links:
http://cartan.gmd.de/TargId/

Please contact:
Alexander Zien or Ralf Zimmer – GMD
Tel: + 49 2241 14-2563 or -2818
E-mail: Alexander.Zien@gmd.de or
Ralf.Zimmer@gmd.de

Modeling and Simulation of Genetic Regulatory Networks

by Hidde de Jong, Michel Page, Céline Hernandez, Hans Geiselmann and Sébastien Maza

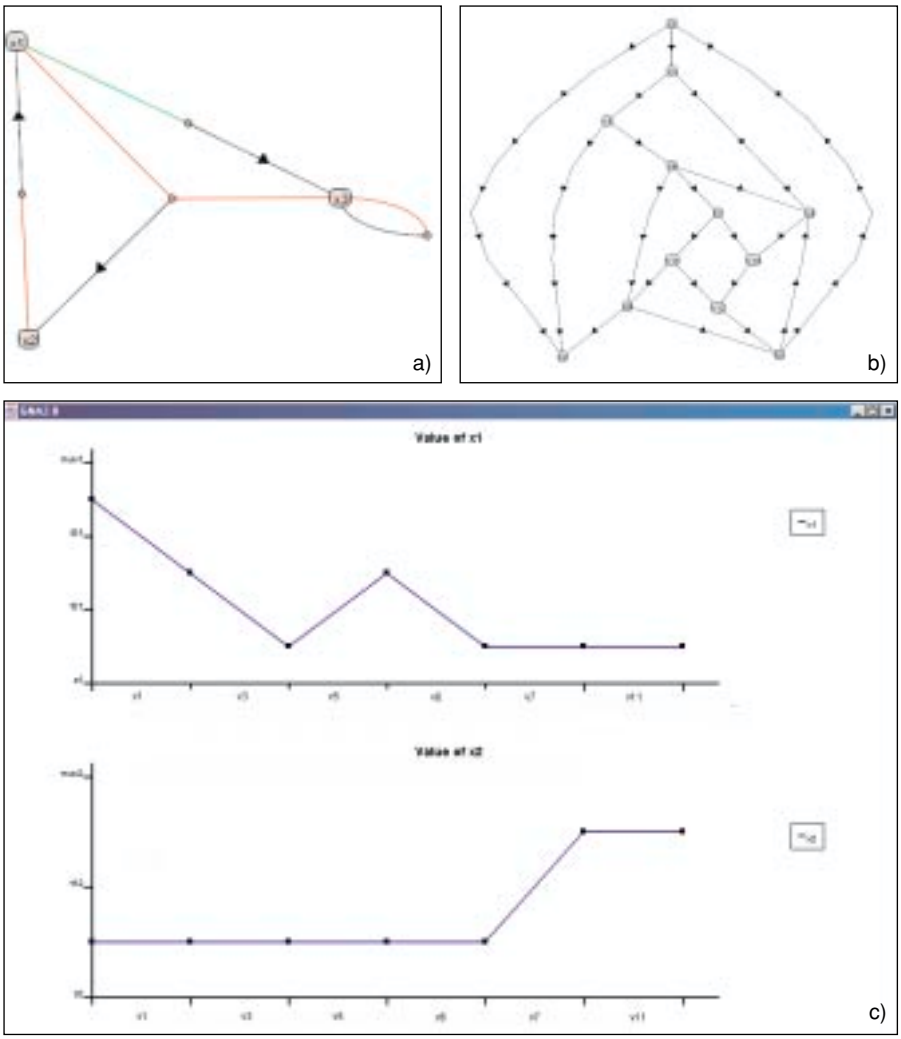
In order to understand the functioning of an organism, the network of interactions between genes, mRNAs, proteins, and other molecules needs to be elucidated. Since 1999, researchers in the bioinformatics group at INRIA Rhône-Alpes have been developing a computer tool for the modeling and simulation of genetic regulatory networks in collaboration with molecular biologists.

The sequencing of the entire genome of prokaryotic and eukaryotic organisms has been completed in the past few years, culminating in the presentation of a working draft of the human genome last June. The analysis of these huge amounts of data involves such tasks as the prediction of folding structures of proteins and the identification of genes and regulatory signals. It is clear, however, that the structural analysis of sequence data needs to be complemented with a functional analysis to elucidate the role of genes in controlling fundamental biological processes.

One of the central problems to be addressed is the analysis of genetic regulatory systems controlling the spatiotemporal expression of genes in an organism. The structure of these regulatory systems can be represented as a network of interactions between genes, proteins, metabolites, and other small molecules. The study of genetic regulatory networks will contribute to our understanding of complex processes like the development of a multicellular organism.

In addition to new experimental tools permitting the expression level to be rapidly measured in a massively parallel way, computer tools for the modeling, visualization, and simulation of genetic regulatory systems will be indispensable. Most systems of interest involve many genes connected through cascades and positive and negative feedback loops, so that an intuitive understanding of their dynamics is hard to obtain. As a consequence of the lack of quantitative information on regulatory interactions, traditional modeling and simulation techniques are usually difficult to apply. To counter this problem, we have developed a method for the qualitative simulation of regulatory systems based on ideas from mathematical biology and artificial intelligence.

The method describes genetic regulatory systems by piecewise-linear differential equations with favourable mathematical properties. The phase space is subdivided into volumes in which the equations reduce to simple, linear and orthogonal differential equations imposing strong constraints on the local behavior of the system. By analyzing the possible transitions between volumes, an indication of the global behavior of the system can be obtained. In particular, the method determines steady-state volumes and volume cycles that are reachable from an initial volume. The steady-state volumes and volume cycles correspond to functional states of the regulatory



Three stages in the simulation process as seen through the GNA user interface. (a) The network of genes and regulatory interactions that is transformed into a mathematical model. (b) The volume transition graph resulting from the simulation. (c) A closer look at the path in the volume transition graph selected in (b). The graph shows the qualitative temporal evolution of two protein concentrations.

system, for instance a response to a physiological perturbation of the organism (a change in temperature or nutrient level).

The above method has been implemented in Java 1.2 in a program called GNA (Genetic Network Analyzer). GNA reads and parses input files with the equations and inequalities specifying the model of the system as well as the initial volume. An inequality reasoner iteratively generates the volumes that are reachable from the initial volume through one or more transitions. The output of the program consists of the graph of all reachable volumes connected by transitions. A graphical interface facilitating the interaction of the user with the program is under development. At present, a visualization module has been realized by which a network of interactions between genes can be displayed, as well as the volume transition graph resulting from the simulation. In addition, the user can focus upon particular paths in the graphs to study the qualitative temporal evolution of gene product concentrations in more detail (see figures).

GNA has been tested using genetic regulatory networks described in the literature, such the example of lambda phage growth control in the bacterium *Escherichia coli*. Simulation experiments with random regulatory networks have shown that, with the current implementation, our method remains tractable for systems of up to 18 genes involved in complex feedback loops.

We plan GNA to evolve into an environment for the computer-supported analysis of genetic regulatory networks, covering a range of activities in the design and testing of models. These activities, such as the validation of hypothesized models of regulatory networks by means of experimental data, will be accessible through a user-friendly graphical interface. In parallel, we will apply the method to the analysis of bacterial regulatory systems in collaboration with biologists at the Université Joseph Fourier in Grenoble.

Links:
HELIX project: http://www.inrialpes.fr/helix/

Please contact:
Hidde de Jong – INRIA Rhône-Alpes
Tel: +33 4 76 61 53 35
E-mail: Hidde.de-Jong@inrialpes.fr

Bioinformatics for Genome Analysis in Farm Animals

by Andy S. Law and Alan L. Archibald

The Bioinformatics Group at the Roslin Institute develops tools and resources for farm animal genome analysis. These encompass the databases, analytical and display tools required for mapping complex genomes. The World Wide Web is used to deliver the resources to users.

The Bioinformatics Group at the Roslin Institute aims to provide access to appropriate bioinformatics tools and resources for farm animal genome analysis. Genome research in farm animals is largely concerned with mapping genes that influence economically important traits. As yet, there are no large-scale genome sequencing activities. The requirements are for systems to support genetic (linkage), quantitative trait locus (QTL), radiation hybrid and physical mapping and to allow data sharing between research groups distributed world-wide.

resSpecies – a Resource for Linkage and QTL Mapping

Genetic linkage maps are constructed by following the co-segregation of marker alleles through multi-generation pedigrees. In quantitative trait locus (QTL) mapping, the performance of the animals is also recorded. Both QTL and linkage mapping studies require databases to store and share the experimental observations. Data sharing between research groups is particularly valuable in linkage mapping. Only by pooling data from the collaborating groups can comprehensive maps be built.

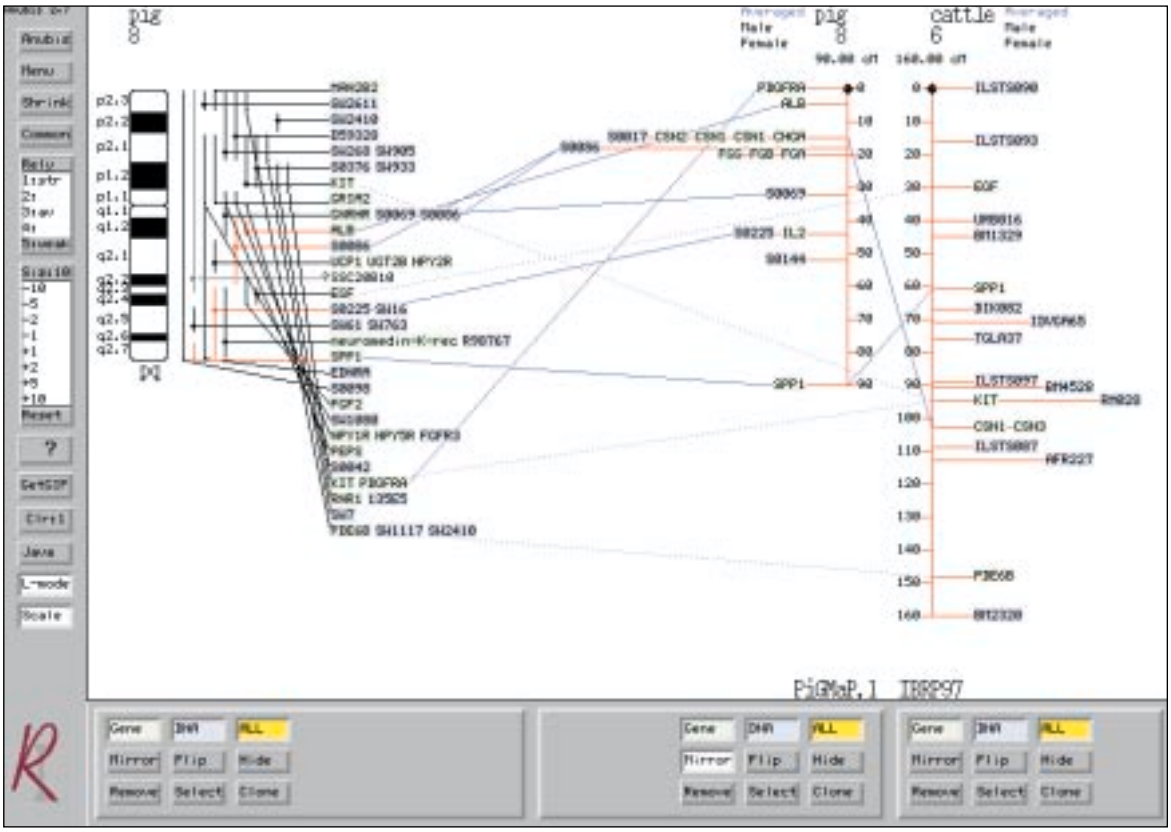
We developed resSpecies to meet this need. It uses a relational database management system (RDBMS - INGRES) with a web-based interface implemented using Perl and Webintool (Hu et al. 1996. WebinTool: A generic Web to database interface building tool. Proceedings of the 7th International Conference and Workshop on Database and Expert Systems (DEXA 96), Zurich, September 9-13, 1996 pp 285-290). This makes international collaborations simple to effect.

The relational design ensures that complicated pedigrees can be represented relatively simply. Populations are defined as groups of individuals. Within resSpecies, access is granted to individual contributors/collaborators on each population separately. The database stores details of markers and alleles. Genotypes may be submitted through a simple web interface that infers missing genotypes, checks for Mendelian inheritance and rejects data that contains inheritance errors. Using a series of simple query forms, data can be extracted in the correct format expected by a number of popular genetic analysis algorithms (eg crimap). This eliminates the possibility of cryptic typographical errors occurring and ensures that the most up-to-date data is available at all times.

resSpecies is used to support Roslin's internal programmes and several international collaborative linkage and QTL mapping projects.

ARKdb – a Generic Genome Database

Scientists engaged in genome mapping research also need access to contemporary summaries of maps and other genome-related data.



A comparison of physical map and genetic maps of pig chromosome 8 with a genetic map of cattle chromosome 6. The maps are drawn 'on-the-fly' by the Anubis map viewer using data held in the ARKdb pig and cattle genome databases.

We have developed a relational (INGRES) genome database model (ARKdb) to handle these data, along with web-based tools for data entry and display. The information stored in the ARKdb databases includes linkage and cytogenetic map assignments, polymorphic marker details, PCR primers, and two point linkage data. Each observation is attributed to a reference source. Hot links are provided to other data sources eg sequence databases and Medline (Pubmed).

The ARKdb database model has been implemented for data from pigs, chickens, sheep, cattle, horses, deer, turkeys, cats, salmon and tilapia. The full cluster of ARKdb databases are mounted on the genome server at Roslin with subsets at Texas A+M and Iowa State Universities. We have also developed The Comparative Animal Genome database (TCAGdb) to capture evidence that specific pairs of genes are homologous. We are developing automated Artificial Intelligence methods to evaluate homology data.

The Anubis Map Viewer

Visualisation is the key to understanding complex data and tools that transform raw data into graphical displays are invaluable. The Anubis map viewer was the first genome browser to be operable as a fully-fledged GUI (Graphical User Interface) over the WWW (URL http://www.ri.bbsrc.ac.uk/anubis). It is used as the map viewer for ARKdb databases and the INRA BOVMAP database. We have recently launched a prototype java version of Anubis - Anubis4 (http://www.ri.bbsrc.ac.uk/arkdb/newanubis/).

Future Activities

We are developing systems to handle the data from radiation hybrid, physical (contig) mapping, expression profiling (microarray) and expressed sequence tag (EST) experiments. Exploitation of the wealth of information from the genomes of human and model organism is critical to farm animal genome research. Therefore, we are exploring ways of improving the links and interoperability with other information systems. Our current tools and resources primarily address the requirements for data storage, retrieval and display. In the future we need to fully integrate analytical tools with the databases and display tools.

The Roslin Bioinformatics Group has grown to eleven including software developers, programmers and database curators. In the past we have received support from the European Commission and Medical Research Council. The group is currently funded by grants from the UK's Biotechnology and Biological Sciences Research Council.

Links:
http://www.roslin.ac.uk/bioinformatics/

Please contact:
Alan L. Archibald – Roslin Institute
Tel: +44 131 527 4200
E-mail: alan.archibald@bbsrc.ac.uk

Modelling Metabolism Knowledge using Objects and Associations

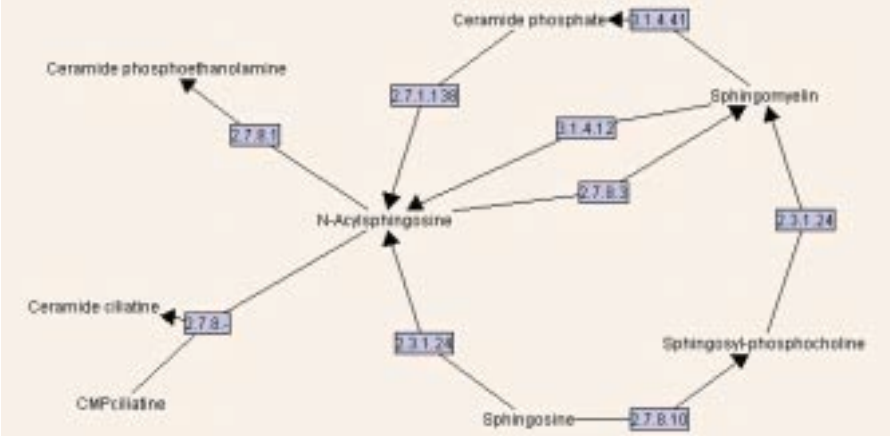
by Hélène Rivière-Rolland, Loïc Taloc, Danielle Ziébelin, François Rechenmann and Alain Viari

A knowledge base to represent metabolism data has been developed by the Helix research team at INRIA Rhône-Alpes. This base provides access to information on chemical compounds, biochemical reactions, enzymes, genes and metabolic pathways from fully sequenced micro-organisms. The model has been implemented by using an object/association technology developed at INRIA. Beside its use as a general repository, the base may have applications in metabolic simulations and pathway reconstruction in newly sequenced genomes.

The cellular metabolism can be defined as the panel of all biochemical reactions occurring in the cell. It consists of molecular synthesis (anabolism) and degradation (catabolism) necessary for cell growth and division. These reactions drive the energetic processes, the synthesis of structural and catalytic components of the cell and the elimination of cellular wastes.

A fairly large amount of metabolic data is readily available, either in the literature or in public data banks (eg the KEGG project: http://star.scl.genome.ad.jp/kegg) and this information will probably grow in the near future due to the development of new 'large scale' experimental technologies like DNA-arrays. Therefore, there is a need to organise this data in a rational and formalised way, ie to model our knowledge of metabolic data. The first goal is of course the storage and recovery of pertinent information. The complexity of this kind of data and in particular the fact that some information is held in the relationship between the biological entities rather than in the entities themselves, makes their selection and recovery difficult. Moreover, our knowledge in this area is often incomplete (elements are missing or pathways may be totally unknown in a newly sequenced organism). A challenge is therefore to cope with this partial information and to develop databases that could provide some inference mechanisms to assist the discovery process. Finally, another challenge is to link these data to other relevant genomic and biochemical information like protein structure, regulation of gene expression, whole genome organisation (eg syntheny) and evolution.

Following the pioneering work of P. Karp and M. Riley with the Eco-Cyc system (http://ecocyc.PangeaSystems.com/ecocyc) we attempted to develop a knowledge base of metabolic data. We wanted to experiment a different representation model in which associations are explicitly represented as entities. To this purpose, we used the AROM system developed at INRIA (http://www.inrialpes.fr/ romans/ pub/arom). The main originality of AROM is the presence of two complementary entities of representation: classes and associations. As in any object-oriented system, a class represents a set of objects described by slots; but, in AROM, such a slot cannot refer to another object. This connection is done by means of associations which therefore denotes a set of tuples (not necessarily only two) of objects (associations are therefore n-ary). As objects, tuples have their own slots and as classes, associations can be organised in hierarchies therefore allowing for usual inheritance and specialisation mechanisms. The explicit representation of n-ary associations turned out to be very useful for representing biological data. For instance, it makes the representation of alternative substrates of a metabolic reaction a much easier task.



Graphical representation of a simple metabolic pathway (sphingophospholipid biosynthesis): the nodes represent chemical compounds (eg N-acylsphingosine) and the edges represent the biochemical reactions which transform these compounds. Each edge is labelled by a number (E.C number) which identifies the enzyme that catalyses the reaction. The figure has been automatically generated using the data stored in the database.

After implementing the data model in AROM, we extracted the metabolic data from public sources (mostly KEGG) by using parsers and Unix shell scripts. Coherence of sequence data between data banks has been checked by using home-made sequences alignment programs and/or Blast. At the present time we are developing several graphical interfaces to this base. One will be devoted to querying the knowledge base. Another interface will be devoted to the automatic graphical representation of pathways which are complex non-planar directed graphs (see Figure). At the present time all the system (AROM and the inter faces) is implemented in JAVA and we plan to put it into play through a web applet-server in a near future.

Links:
Action Helix:
http://www.inrialpes.fr/helix.html

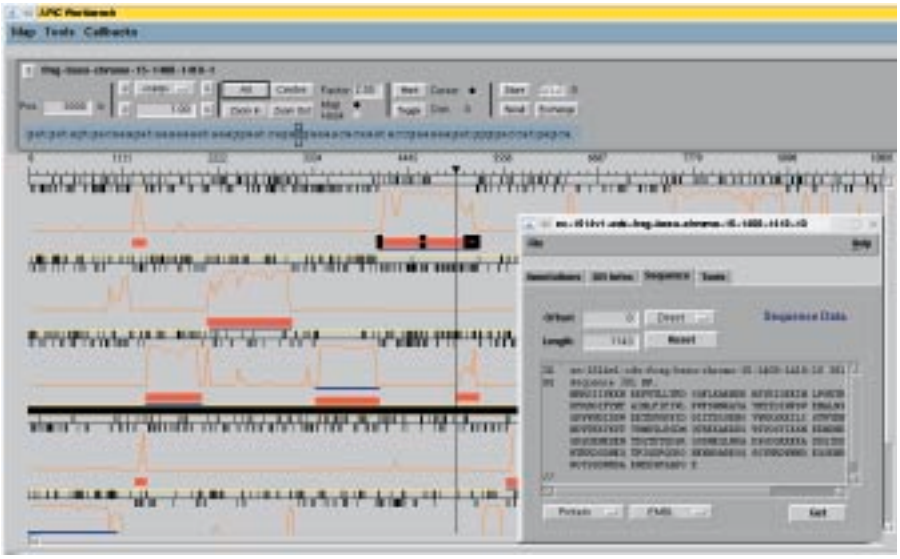
Please contact:
Alain Viari – INRIA
Tel: +33 4 76 61 54 74
E-mail: alain.viari@inrialpes.fr

Co-operative Environments for Genomes Annotation: from Imagene to Geno-Annot

by Claudine Médigue, Yves Vandenbrouke, François Rechenmann and Alain Viari

'Imagene' is a a co-operative computer environment for the annotation and analysis of genomic sequences developed in collaboration between INRIA, Université Paris 6, Institut Pasteur and the ILOG company. The first version of this software was dedicated to bacterial chromosomes. Its capabilities are currently extended to handle both prokaryotic and eukaryotic data and to link pure genomic data to 'post-genomic' data, particularly metabolic and gene expression data.

In the context of large-scale genomic sequencing projects the need is growing for integration of specific sequence analysis tools within data management systems. With this aim in view, we have developed the Imagene co-operative computer environment dedicated to automatic sequence annotation and analysis (http://abraxa.snv.jussieu.fr/imagene). In this system, biological knowledge produced in the course of a genome sequencing project (putative genes, regulatory signals, etc) together with the methodological knowledge, represented by an extensible set of sequence analysis methods, are uniformly represented in an object oriented model.



Imagene view of a fragment of the B. subtilis chromosome: The display superimposes the output of several methods. Red boxes represent putative protein coding region (gene); the blue boxes represent the result of a data bank similarity scan (here the Blastx program); the yellow curve represents the coding probability as evaluated by using a Markov chain. The translated protein sequence of the currently selected gene is shown in the insert.

Imagene is the result of a five years collaboration between INRIA, Université Paris 6, the Institut Pasteur and the ILOG company. The system has been implemented by using an object oriented model and a co-operative solving engine provided by ILOG. In Imagene, a global problem (task) is solved by successive decompositions into smaller sub-tasks. During the execution, the various sub-tasks are graphically displayed to the user. In that sense, Imagene is more transparent to the user than a traditional menu-driven package for sequence analysis since all the steps in the resolution are clearly identified. Moreover, once a task has been solved, the user can restart it at any point; the system then keeps track of the different versions of the execution. This allows to maintain several hypothesis in parallel during the analysis. Imagene also provides a user interface to display, on the same picture, the results produced by one or several strategies (see Figure). Due to the homogeneity of the whole software, this display is fully interactive and the graphical objects are directly connected to their database counterpart.

Imagene has been used within several bacterial genome sequencing projects (Bacillus subtilis and Mycoplasma pulmonis) and has proved to be particularly useful to pinpoint sequencing errors and atypical genes. However this first version suffers several drawbacks. First it was limited to the representation of prokaryotic data only, second the development tools were commercial thus giving rise to difficulties in its diffusion, last, it was designed to handle pure sequence data from a single genome. In order to overcome these limitations, we undertook a new project (Geno-Annot) through a collaboration between INRIA, the Institut Pasteur and the Genome-Express biotech compagny. As a first step, the data model was extended to eukaryotes and completely re- implemented using the AROM system developed at INRIA (http://www.inrialpes.fr/romans/pub/arom). We are now in the process of re-designing the task-engine and the graphical user interfaces in JAVA. Finally, our ultimate goal will be to integrate Geno-Annot within a more general environment (called Geno-*) in order to fully link all the pieces of genomic information together (ie sequence data, metabolism, gene expression etc). Geno-Annot is a two years project that started in September 1999.

Links:
Action Helix:
http://www.inrialpes.fr/helix.html
Imagene:
http://abraxa.snv.jussieu.fr/imagene

Please contact:
Alain Viari – INRIA
Tel: +33 4 76 61 54 74
E-mail: alain.viari@inrialpes.fr

Arevir: Analysis of HIV Resistance Mutations

by Niko Beerenwinkel, Joachim Selbig, Rolf Kaiser and Daniel Hoffmann

To develop tools that assist medical practitioners in finding an optimal therapy for HIV-infected patients – this is the aim of a collaboration funded by Deutsche Forschungsgemeinschaft that has been started this year by researchers at GMD, the University of Cologne, CAESAR, the Center of Advanced European Studies and Research, Bonn, and a number of cooperating university hospitals in Germany.

The Human Immunodeficiency Virus (HIV) causes the Acquired Immunodeficiency Syndrome (AIDS). Currently, there are two types of drugs in the fight against HIV, namely inhibitors of the two viral enzymes protease (PR) and reverse transcriptase (RT).

Since HIV shows a very high genomic variability, even under the usual combination therapy (HAART – highly active antiretroviral therapy) consisting of several drugs, mutations occur, that confer resistance to the prescribed drugs and even to drugs not yet prescribed (cross resistance). Therefore, the treating physician is faced with the problem of finding a new therapy rather frequently.

Clinical trials have shown that therapy changes based on a genotypic resistance test, ie sequencing of the viral gene coding for PR and RT and looking for mutations known to cause resistance, result in a significantly better therapy success. However, not all patients benefit from therapy changes after resistance testing. There are several possible reasons for therapy failure in this situation: the occurrence of an HIV-strain resistant to all available antiretroviral drugs, no sufficient drug-level is reached in the patient, or the chosen drug-combination was not able to suppress the virus sufficiently. The latter occurs because the relations between observed mutations, phenotypic resistance and therapy success are poorly understood.

While the PR inhibitors, for example, all bind in the catalytic center of this enzyme, mutations associated with resistance occur at many different locations spread all over the three-dimensional structure of the protein (see Figure 1). The relation between point mutations and drug resistance remains unclear in many cases, not to speak of the interpretation of more complex mutation patterns.

The goal of the Arevir project is to develop bioinformatics methods that help to understand these connections and that contribute directly to therapy optimization. In a first step a database is set up in collaboration with project partners from university hospitals and virological institutes, in which clinical data, sequence data and phenotypic resistance data are collected.

These correlated data are used to learn about the outcome of a therapy as a function of the drugs making up the components of this therapy and the genotype of the two relevant enzymes PR and RT. A successful outcome of a therapy can be measured as a substantial reduction in virus load (ie the number of virus particles measured in the patients' blood plasma; see Figure 2). We can formulate a classification problem on the set of all pairs consisting of the therapy's drug components and the amino acid sequence of PR and RT assigning to each such pair either the class 'successful' or 'unsuccessful'.

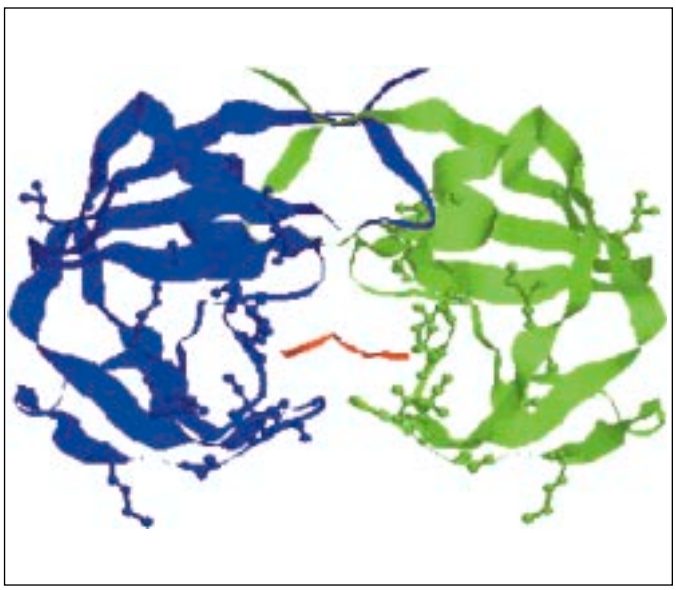


Figure 1: Ribbon representation of the HIV protease homodimer (blue and green) complexed with an inhibitor (red), some resistance associated mutations (codon positions 10, 20, 36, 46, 48, 50, 54, 63, 71, 82, 84 and 90) are indicated in the ball-and-stick mode.

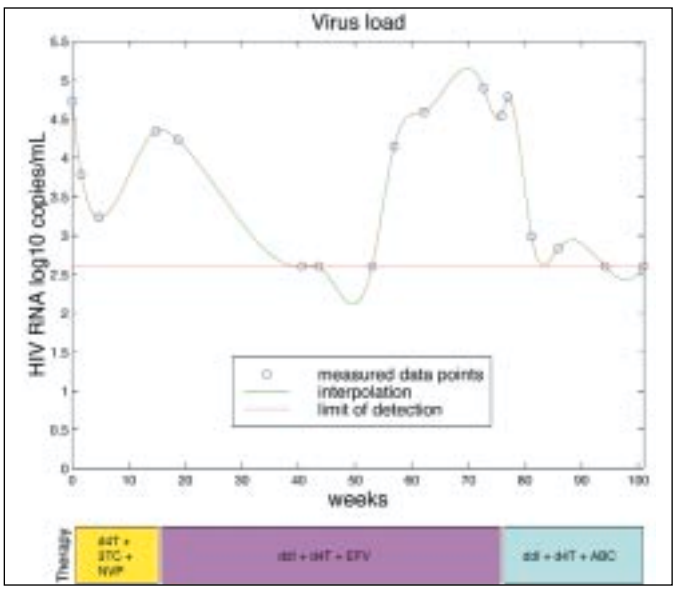


Figure 2: Virus load and therapies of a patient in a time period of more than 100 weeks (the measured data points have been interpolated, values under the limit of detection have been normalized to this limit (400 copies/ml), drugs are encoded in a three letter code).

We use decision tree building, a well known machine learning technique, to derive a model that can make an educated guess on the therapy outcome. The treating physician faced with the problem of finding a new therapy can consult the decision tree with a therapy proposal, provided a genotypic test result is available. In this way the automated interpretation of sequence data can directly improve patient care. Furthermore, the learned rules derived from the decision tree can be interpreted by virologists and clinicians, and help to better understand and manage drug resistance.

In future work the mechanisms of drug resistance will be studied at the molecular level. To this end we will carry out force field based calculations on enzyme-inhibitor complexes.

While traditionally bioinformatics methods are used in the course of developing new drugs, we develop and apply methods that can directly improve patient care by contributing to therapy optimization. Thus, in treatment of HIV infected patients techniques of interpreting genotypic data are of immediate relevance to therapy decision.

Links:
Bioinformatics at http://www.gmd.de/SCAI

Please contact:
Niko Beerenwinkel – GMD
Tel.: +49 2241 14 2786
E-mail: niko.beerenwinkel@gmd.de

Human Brain Informatics – Understanding Causes of Mental Illness

by Stefan Arnborg, Ingrid Agartz, Mikael Nordström, Håkan Hall and Göran Sedvall

The Human Brain Informatics (HUBIN) project aims at developing and using methodology for cross-domain investigations of mental illness, particularly schizophrenia. A comprehensive data base with standardized information on individual patients and healthy control persons is being built and is analyzed using modern data mining and statistics technology.

Mental disease is a complex phenomenon whose causes are not known, and only symptomatic therapy exists today. Moreover, mental disease affects maybe one third of all humans at least once during their life time. Approximately 1% of all humans suffer from schizophrenia. The more severe cases lead to tragic life-long disability, and the annual costs can be measured in many billions of Euros even in moderate size countries like Sweden. On the other hand, some of our greatest artists and scientists suffered from schizophrenia. Large research projects have mapped mental diseases so that their typical manifestations are well known separately in several specialized medical domains such as genetics, physiology, psychiatry, neurology and brain morphology. But little is known about the relationships between the different domains.

The Project

The HUBIN project started in 1998, and is carried out at the Clinical Neurosciences Department of the Karolinska Institute, with participation by Swedish Institute of Computer Science (SICS), IBM and several medical schools: Lund and Umeå in Sweden, Cardiff in the UK, and Iowa City, USA. It is financed by a grant from the Swedish Wallenberg foundation, with supplementary funding from SICS, NUTEK, Teknikbrostiftelsen and from some private foundations.

Project Goals

The aim of the project is, in the long term, to find causes and effective therapies for mental illness. The short term aims of the project are to develop methodology in intra-domain and cross-domain investigations of schizophrenia by building a comprehensive data base with information about individual patients and control persons and using modern statistical and data mining technology.

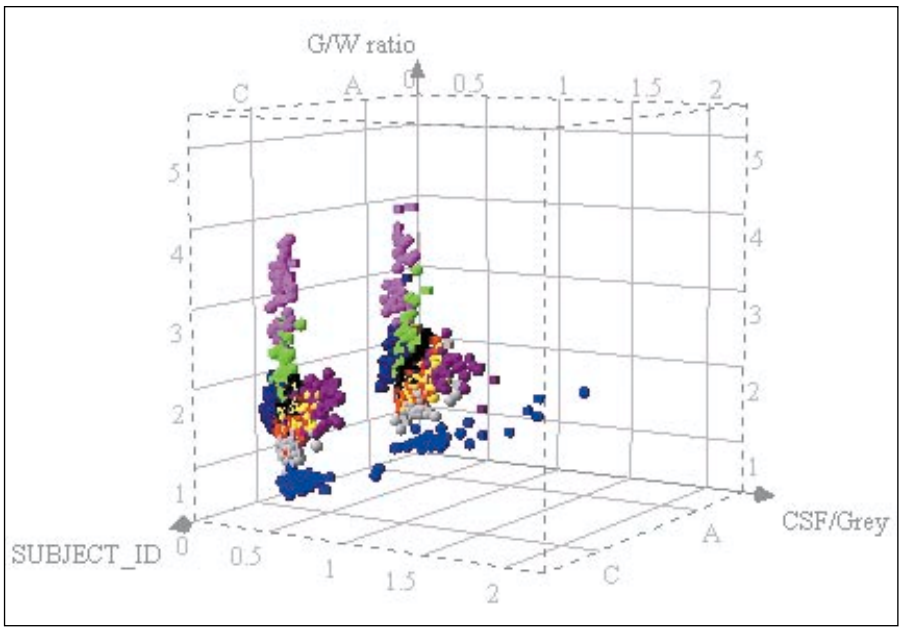
Possible Causes of Schizophrenia

Schizophrenia is believed to develop as a result of disturbed signalling within the human brain. Causes of these disturbances can be anomalous patterns of connections between neurons of the functional regions, loss or anomalous distribution of neurons, or disturbances in the biochemical signalling complex. It is known that it is significantly influenced by genetic factors, although in a complex way which has not yet been finally attributed to individual genes. It is also related to disturbances in early (pre-natal) brain development, in ways not yet known but statistically confirmed by investigations of birth and maternity journals. The shape of the brain is also affected, but it is difficult to trace the complex interactions between disease, medication and brain morphology.

Complex Sets of Medical Information

Modern medical investigations create enormous and complex data sets, whose analysis and cross-analysis poses significant statistical and computing challenges. Some of the data used in HUBIN are:

- Genetic micro-array technology can measure the activity of many thousand genes from a single measurement on a microscopic (post mortem) brain sample. These data are noisy and extremely high-dimensional, so standard regression methods fail miserably. Support vector and mixture modelling techniques are being investigated. It is necessary to combine the hard data with subjective information in the form of hypotheses on the role of different genes. Heredity investigations based on disease manifestation and marker maps of relatives also yield large amounts of data with typically (for multiple gene hunting) extremely weak statistical signals.
- Medical imaging methods can give precise estimates of the in vivo anatomy of the brain and the large individual variations in shape and size of white (axon, glia), gray (neurons) and wet



Tissue ratios in different brain regions: A multidimensional view of measured anatomical volumes in diseased and controls reveals a characteristic but irregular increase of ventricle sizes (blue dots) for schizophrenia patients. The data were obtained from MR scans performed in the brain morphology project, using volumetry methods developed at University of Iowa. Measurements were obtained by Gaku Okugawa.

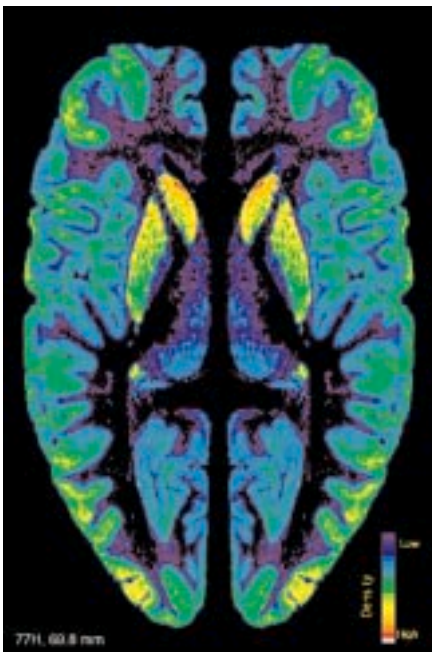


Image: Håkan Hall, Karolinska Institutet.

Dopamine-D1 receptors in the human brain post-mortem visualized using whole hemisphere autoradiography and [^{3}H]NNC-112.

(CSF in ventricles and outside the brain) matter in a large number of anatomical regions of the brain.

- Diffusion tomography gives approximate measures of the diffusion tensor in small cubes (ca 3 mm side) which indicates number of and direction of axons that define the long-distance signalling connections of the brain. This tensor can thus be used to get approximate measures of the signalling connectivity of the brain.
- Functional MRI measures the metabolism (blood oxygenation) that approximates neural activity with high resolution. These investigations give extremely weak signals, and for inference it is usually necessary to pool several investigations.

Post-mortem whole brain investigations can give extremely high resolution maps of the biochemical signalling system of the brain, and of gene activity in the brain. More than 50% of the active human genome is believed to be related to brain development, and very little is known about the mechanisms involved.

The patients mental state is measured by a psychiatrist using standardized questionaries. It is vital that the subjective information entered is standardized and quality assured. Obviously, it is a difficult problem to get high-quality answers to 500 questions from a patient, so it is critical to balance the number of questions asked to patients.

As a first sifting of this large information set, summary indices are computed and entered in a relational data base, where standard data mining and statistical visualization techniques give a first set of promising lines of investigations. Information regarding the mental state of individuals is extremely sensitive and its collection and use is regulated by ethics councils of participating universities and hospitals. Identities of patients and controls are not stored in the data base, but it must still be possible to correlate individuals across domains. This is accomplished with cryptographic methods.

Conclusions

The current activities are aimed at showing the feasibility of our interdisciplinary approach when coupled with standardized recording of clinical information.

Presently, the main efforts go into collection of standardized clinical measurements and evaluation of statistical and visual analysis methods in image and genetic information analysis. Current activities aim at relating brain morphology (size, orientation and shape of various regions and tissue types of the brain) to physiological and psychiatric conditions and to investigate relationships between the different domains.

Needless to say, the current project is only one and as yet a small one of many projects worldwide aimed at improving conditions for persons affected by mental illness. In order to connect and improve communications between these many groups, our efforts also involve intelligent text mining for rapidly scanning the literature.

We are building a web site, http://hubin.org/about/index_en.html, for communication between researchers worldwide and between medical experts and relatives of affected persons on national (native language) basis.

Links:
HUBIN: http://hubin.org/about/index_en.html

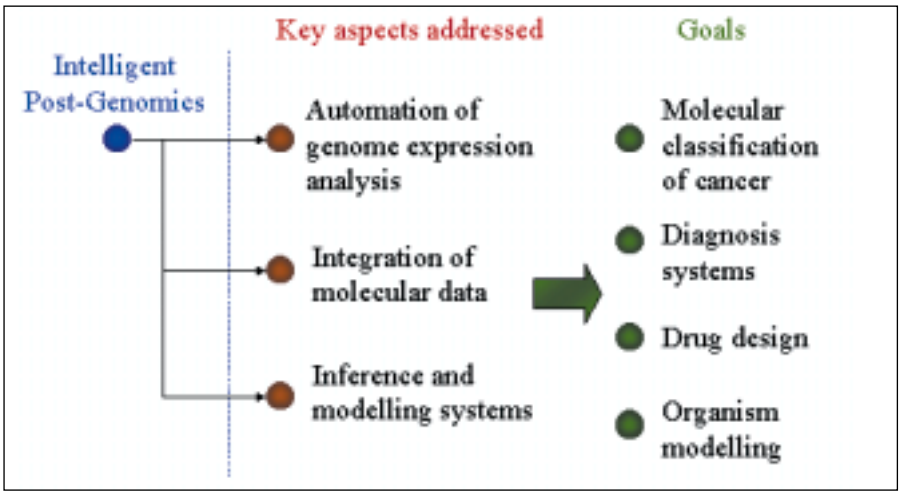
Please contact:
Stefan Arnborg – Nada, KTH
Tel +46 8 790 71 94
E-mail: stefan@nada.kth.se
http://www.nada.kth.se/~stefan/

Intelligent Post-Genomics

by Francisco Azuaje

The Intelligent Post-Genomics initiative aims to support a number of knowledge discovery tasks with crucial applications in medicine and biology, such as diagnostic systems, therapy design and organism modelling. This research is mainly focused on the development of a new generation of genomic data interpretation systems based on Artificial Intelligence and data mining techniques.

The union of theoretical sciences such as mathematics, physics and computer science with biology is allowing scientists to model, predict and understand crucial mechanisms of life and the cure of diseases. The success of this scientific synergy will depend not only on the application of advanced information processing methods, but also on the development of a new multidisciplinary language. Researchers from the Artificial Intelligence Group and Centre for Health Informatics together with the Departments of Microbiology, Biochemistry and Genetics of the University of Dublin are convinced that this joint action will yield significant benefit in the understanding of key biological problems. The Figure illustrates the main aspects addressed in this project as well as some of its possible applications.



Intelligent Post-Genomics: research issues and applications.

The need for higher levels of reliability, emphasising at the same time theoretical frameworks to perform complex inferences about phenomena, makes Artificial Intelligence (AI) particularly attractive for the development of the post-genomic era. The incorporation of techniques from AI and relating research fields may change the way in which biological experiments and medical diagnoses are implemented. For example, the analysis of gene expression patterns allows us to combine advanced data mining approaches in order to relate genotype and phenotype.

Within the past few years, technologies have emerged to record the expression pattern of not one, but tens of thousands of genes in parallel. Recognising the opportunities that these technologies provide, the research groups mentioned above are launching a long-term initiative to apply AI, data mining and visualisation methods to important biological processes and diseases in order to interpret their molecular patterns. Initial efforts have already allowed us to explore and confirm the potential of these technologies for the development of tumour classification systems, detection of new classes of cancer and the discovery of associations between expression patterns.

Thus, one of our major goals is the automation of the genome expression interpretation process. This task should provide user-friendly, effective and efficient tools capable of organising complex volumes of expression data. It also aims to allow users to discover associations between apparently unrelated classes or expression patterns. This may represent not only a powerful approach to understand genetic mechanisms in the development of a specific disease, but also to support the search for fundamental processes that differentiate multiple types of diseases. A number of intelligent hybrid frameworks based on neural networks, fuzzy systems and evolutionary computation have been shown to be both effective and efficient for the achievement of these decision support systems. Furthermore it may support the identification of drug targets by making inferences about functions that are associated to sequences and genome patterns.

Other crucial research goals involve the combination of gene expression data with other sources of molecular data (such as drug activity patterns) and morphological features data. Similarly, we recognise the need to develop approaches to filter, interconnect and organise dispersed sources of genomic data.

Collaboration links with other European research institutions, such as the German Cancer Research Centre (Intelligent Bioinformatics Systems Group) and Max Planck Institute for Astrophysics (Molecular Physics Group), are actually being developed as part of our efforts to develop advanced data interpretation technologies for life sciences.

Links:
A collection of representative links to genomics and bioinformatics resources: Genomes & Machines: http://www.cs.tcd.ie/Francisco.Azuaje/genomes&machines.html

Please contact:
Francisco Azuaje – Trinity College Dublin
Tel: +353 1 608 2459
E-mail: Francisco.Azuaje@cs.tcd.ie

Combinatorial Algorithms in Computational Biology

by Marie-France Sagot

It is almost a 'cliché' to say that the sequencing at an everincreasing speed of whole genomes, including that of man, is generating a huge amount of information for which the use of computers becomes essential. This is true but is just a small part of the truth if one means by that the computer's capacity to physically store and quickly sift through big quantities of data. Computer science may strongly influence at least some areas of molecular biology in other, much deeper ways.

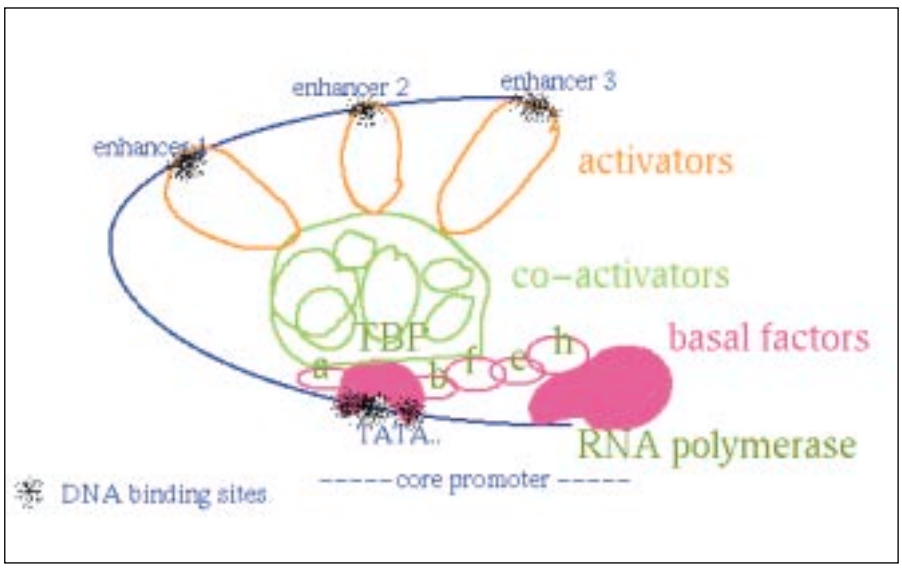
A lot of the problems such areas have to address, in the new post-sequencing era but not only, requires exploring the inner structure of objects, trying to discover regularities that may lead to rules, building general models which attempt to weave and exploit relations among objects and, finally, obtaining the means to represent, test, question and revise the elaborated models and theories. These are all activities that are at the essence of most currently elaborated computer algorithms for analysing, often in quite subtle ways, biological data.

Among the algorithmical approaches possible, some try to exhaustively explore all the ways a molecular process could happen given certain (partial) beliefs we have about it. This leads to hard combinatorial problems for which efficient algorithms are required. Such algorithms must make use of complex data representations and techniques.

For instance, quite a few processes imply the recognition by the cell machinery of a number of sites along a DNA sequence. The sites represent fragments of the DNA which other macromolecular complexes must be able to recognize and bind to in order for a given biological event to happen. Such an event could be the transcription of DNA into a so-called messenger RNA which will then be partially translated into a protein. Because their function is important, sites tend to be conserved (though the conservation is not strict) while the surrounding DNA suffers random changes. Other possible sequential and/or spatial characteristics of sites having an equivalent function in the DNA of an organism may exist but are often unknown. Hypotheses must then be elaborated on how recognition works. For example, one probable scenario is that sites are cooperative as suggested in Figure 1. Each one is not recognized alone but in conjunction with others at close intervals in space and time. A model based upon such a hypothesis would try to infer not one site (corresponding to inferring the presence of a conserved word at a number of places along a string representing a DNA sequence), but various sites simultaneously, each one located at an often precise (but unknown) distance from its immediate neighbours (a problem one may see as equivalent to identifying a maximal common sequence of conserved and spatially constrained motifs). The model may be further sophisticated as the analysis progresses. What is important to observe is that the revision of a model may be effected only if the algorithm based upon it is exhaustive.

Exact algorithms may therefore be powerful exploratory tools even though they greatly simplify biological processes (all approaches including experimental ones, do). In particular, exact algorithms allow to 'say something', derive some conclusions even from what is not identified because we know that, modulo some parameters they may take as input, they exhaustively enumerate all objects they were built to find. Whether they indeed find something (for instance, a site initially thought to exist with a certain characteristic) or not, we may thus, by understanding their inner workings, emit biological hypotheses on 'how things function', or do not function. This is possible also with non exact approaches such as some of the statistical ones, but is often more difficult to realize.

A first algorithm using a model which allows to simultaneously infer cooperative sites in a DNA sequence has been elaborated (in C) by the algorithmics group at the Pasteur Institute (composed of one researcher and PhD students from the Gaspard Monge Institute). In



A much simplified illustration of the possible cooperativeness of sites implicated in the process of transcription from DNA to messenger RNA. Objects in orange, green and pink represent protein or RNA-protein complexes. The sites on the DNA where they are supposed to bind are depicted as clouds of black points. Promoter sequences are recognized by the pink objects. Each is a protein called RNA polymerase. Enhancers may reinforce or weaken the binding strength of a promoter. The example sketched in this figure concerns a eukaryotic organism. The figure was taken from URL: http://linkage.rockefeller.edu/wli/gene/right.html

collaboration with biology researchers from the Biology and Physico-Chemistry and the Pasteur Institutes in Paris, this algorithm has started been applied to the analysis of the sequenced genomes of three bacteria: *Escherichia coli*, *Bacillus subtilis* and *Helicobacter pylori*. The objective was to identify the main sequence elements involved in the process of transcription of DNA into messenger RNA (what is called the promoter sequences). For various reasons, it was suspected that such elements would be different in *Helicobacter pylori* as compared to the other two bacteria. A blind, exhaustive inference of single conserved words in the non coding regions of the genome of *Helicobacter pylori* permitted to identify some elements but not to put together with enough certainty those implicated in the transcriptional process. The development of more complex models was called upon and, indeed, allowed to reveal that the sequence elements are not the same in *Helicobacter pylori*. It has also enabled to suggest a consensus for what the main so-called promoter elements should be in the bacterium. It was later verified that the consensus proposed strongly resembles one experimentally determined for a bacterium which is closely related in terms of evolution to *Helicobacter pylori*.

This particular algorithm is been continously ameliorated, expanded and extended to take into account the increasingly more sophisticated models which further analyses of other genomes and processes reveal are necessary to get closer to the way recognition happens in diverse situations. Other combinatorial algorithms, addressing a variety of subjects – molecular structure matching, comparison and inference; regularities detection; gene finding; identification of recombination points; genome comparison based upon rearrangement distances – are currently been studied within the Pasteur Institute algorithmics group, often in collaboration with outside PhD students and researchers.

It is worth observing that many computer scientists come to biology seduced by the huge range of nice problems that this offers to them without necessarily wanting to get implicated into the intricacies of biology itself. Although the two, computer science and biology, remain quite distinct fields of research, it appears difficult to gain maturity and any real insight into the domain without getting further involved into biology. This is particularly, but not exclusively, true if one adopts a combinatorial approach as the operations of building models that try to adhere to the way biological processes happen, and algorithms for exploring and testing such models, become intimately linked in a quasi-symbiotic relationship. It would, anyway, be a pity not to get so involved as the problems become much more interesting, even from a purely computational and combinatorial point of view, when they are not abstracted from what gave origin to them: a biological question. The case of identifying sites is just one example among many.

Links:
Papers by the group on this and other combinatorial problems:
http://www-igm.univ-mlv.fr/~sagot
Series of seminars on this and other topics:
http://www.pasteur.fr/infosci/conf/AlgoBio

Please contact:
Marie-France Sagot – Institut Pasteur
Tel: + 33 1 40 61 34 61
E-mail: sagot@pasteur.fr

Crossroads of Mathematics, Informatics and Life Sciences

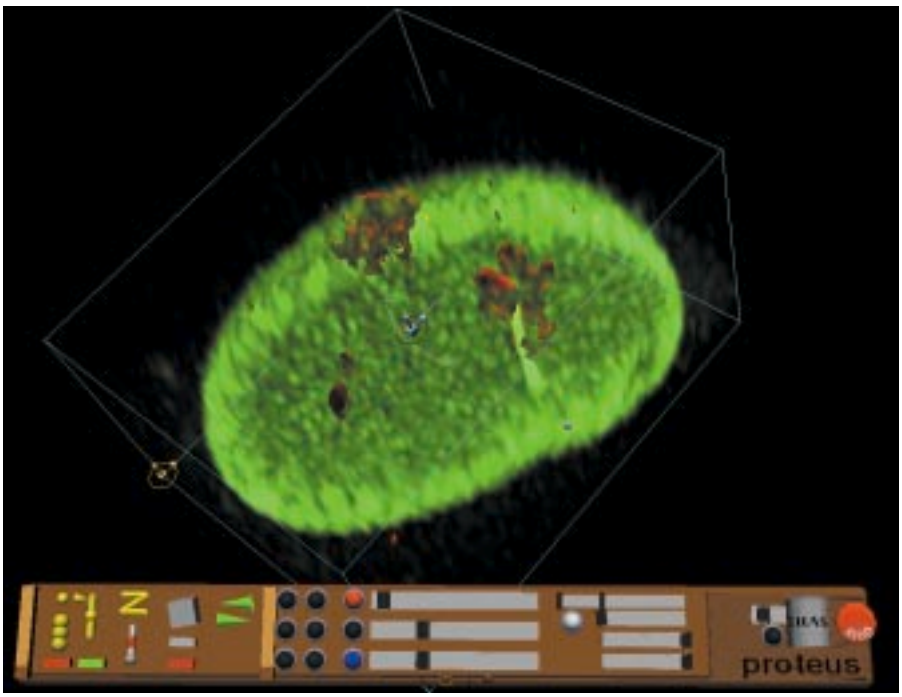
by Jan Verwer and Annette Kik

The 20th century has been a period of dramatic technological and scientific breakthroughs and discoveries. A general feeling amongst scientists in all areas is that developments in science and technology will go on, perhaps most notably in the life and computer sciences as testified by the current rapid progress in bio- and information technology. New subdisciplines have been created such as computational biology and bioinformatics, which are at the crossroads of biology, physics, chemistry, informatics and mathematics. CWI is active in a number of projects with life science applications while other projects are forthcoming.

One of our current projects is 'Macro-Molecular Crowding in Cell Biology', in co-operation with several cell biologists. Living cells are relatively densely packed with macromolecules (proteins), implying that diffusion of larger molecules can be much slower than diffusion of smaller ones. This crowding effect is important for the uptake of molecules entering a cell (glucose). The influence of this crowding effect is as yet largely unknown. The aim is to study metabolic control in pathways by means of numerical and analytical studies of systems of reaction-diffusion equations which take into account effects of molecular crowding.

Another project is 'Analysis of Biological Structures by Virtual Reality Techniques'. This aims at exploring the potentials of virtual reality visualization techniques in order to obtain detailed insight into complex biological structures. Such structures include proteins and cellular structures, as wel as organs and complete organisms. Visualization of biological structures is a key step in understanding normal and pathological biological processes and systems, such as biocatalysis, signal transduction, gene expression and embryonal development.

Our forthcoming projects have different applications. CWI will study models for the development of phytoplankton blooms. Phytoplankton lies at the beginning of the food chain in oceans and takes up large amounts of carbon dioxide from the atmosphere. Hence there is a



A three-dimensional visualization of a cell laminea (green) and chromosomes (red). The panel on the bottom is part of interactive visualization and measuring software being developed by Wim de Leeuw at CWI (Data provided by Roel van Driel, SILS Amsterdam).

Result of a simulation of axon development in the presence of three diffusible fields. Courtesy A. van Ooyen, Netherlands Institute for Brain Research.

clear link with climate research. The complicated models require efficient numerical solution techniques.

The project 'Distributed Visualization of Biological Structures' aims at the development of an experimental platform to study advanced image processing techniques for cell structures and interaction schemes for remote steering of a laboratory apparatus, in the context of high speed networks. Interactive distributed visualization coupled with virtual environments allows for collaborative analysis, interrogation of large repositories of biological structures and distance learning.

Since about a decade, scientists realise that many biological and physical systems seem to drive themselves, without precise external tuning, towards a state which resembles a typical equilibrium statistical mechanics system at the critical point. Examples are certain types of epidemics, forest fires, and processes in the brain. Most studies of this this so-called Self-Organized Critical behaviour are of a very heuristic nature, whereas we will use a more rigorous mathematical approach.

A hot question in neuroscience is how the hand-shaped ends of axons growing out of neurons explore the territory they traverse on their way to a target tissue. This growth process is partly guided by concentration gradients of biochemical molecules in the extracellular space, arising from diffusion and various chemical interactions. This leads to models consisting of parabolic equations coupled with gradient-type equations for the positions of the axons. We will study these models and their numerical solution and implement efficient numerical algorithms in software that is user-friendly for neuroscientists, as a joint activity with the Netherlands Institute for Brain Research.

Furthermore, there is a whole range of problems linking to Combinatorial Optimization. For instance, the primary access to the functionality of molecules is through laboratory experiments. However, experimental data typically provide only partial information about structures and properties under investigation. This prompts the problem of recovering the actual molecular, chemical properties and structures from the incomplete data. This often leads to so-called combinatorial optimization problems. In biology the complexity is very large, so that there will be a need to improve or redesign existing optimization methods or even come up with totally new ideas.

Finally, CWI plans to do research in the field of Control and System Theory for the Life Sciences. The aim of this research is to gain understanding of the feedback control in cells, plants, living beings, and in ecological communities. This understanding may be used to develop treatments for diseases, and for ecological communities.

Links:
http://www.cwi.nl/~janv

Please contact:
Jan Verwer – CWI
Tel: +31 20 592 4095
E-mail: Jan.Verwer@cwi.nl
Mark Peletier – CWI
Tel: +31 20 592 4226
E-mail Mark.Peletier@cwi.nl

Biomolecular Computing

by John McCaskill

Biomolecular computing, 'computations performed by biomolecules', is challenging traditional approaches to computation both theoretically and technologically. Often placed within the wider context of 'natural' or even 'unconventional' computing, the study of natural and artificial molecular computations is adding to our understanding both of biology and computer science well beyond the framework of neuroscience. The papers in this special theme document only a part of an increasing involvement of Europe in this far reaching undertaking. In this introduction, I wish to outline the current scope of the field and assemble some basic arguments that biomolecular computation is of central importance to both computer science and biology. Readers will also find arguments for not dismissing DNA Computing as limited to exhaustive search and for a qualitatively distinctive advantage over all other types of computation including quantum computing.

The idea that molecular systems can perform computations is not new and was indeed more natural in the pre-transistor age. Most computer scientists know of von Neumann's discussions of self-reproducing automata in the late 1940s, some of which were framed in molecular terms. Here the basic issue was that of bootstrapping: can a machine construct a machine more complex than itself?

Important was the idea, appearing less natural in the current age of dichotomy between hardware and software, that the computations of a device can alter the device itself. This vision is natural at the scale of molecular reactions, although it may appear utopic to those running huge chip production facilities. Alan Turing also looked beyond purely symbolic processing to natural bootstrapping mechanisms in his work on self-structuring in molecular and biological systems. Purely chemical computers have been proposed by Ross and Hjelmfelt extending this approach. In biology, the idea of molecular information processing took hold starting from the unraveling of the genetic code and translation machinery and extended to genetic regulation, cellular signaling, protein trafficking, morphogenesis and evolution - all of this independently of the development in the neurosciences. For example, because of the fundamental role of information processing in evolution, and the ability to address these issues on laboratory time scales at the molecular level, I founded the first multi-disciplinary Department of Molecular Information Processing in 1992. In 1994 came Adleman's key experiment demonstrating that the tools of laboratory molecular biology could be used to program computations with DNA *in vitro*. The huge information storage capacity of DNA and the low energy dissipation of DNA processing lead to an explosion of interest in massively parallel DNA Computing. For serious proponents of the field however, there really never was a question of brute search with DNA solving the problem of an exponential growth in the number of alternative solutions indefinitely. In a new field, one starts with the simplest algorithms and proceeds from there: as a number of contributions and patents have shown, DNA Computing is not limited to simple algorithms or even, as we argue here, to a fixed hardware configuration.

After 1994, universal computation and complexity results for DNA Computing rapidly ensued (recent examples of ongoing projects here are reported in this collection by Rozenberg, and Csuhaj-Varju). The laboratory procedures for manipulating populations of DNA were formalized and new sets of primitive operations proposed: the connection with recombination and so called splicing systems was particularly interesting as it strengthened the view of evolution as a computational process. Essentially, three classes of DNA Computing are now apparent: intramolecular, intermolecular and supramolecular. Cutting across this classification, DNA Computing approaches can be distinguished as either homogeneous (ie well stirred) or spatially structured (including multi-compartment or membrane systems, cellular DNA computing and dataflow like architectures using microstructured flow systems) and as either *in vitro* (purely chemical) or *in vivo* (ie inside cellular life forms). Approaches differ in the level of programmability, automation, generality and parallelism (eg SIMD vs MIMD) and whether the emphasis is on achieving new basic operations, new architectures, error tolerance, evolvability or scalability. The Japanese Project lead by Hagiya focuses on intramolecular DNA Computing, constructing programmable state machines in single DNA molecules which operate by means of intramolecular conformational transitions. Intermolecular DNA Computing, of which Adleman's experiment is an example, is still the dominant form, focusing on the hybridization between different DNA molecules as a basic step of computations

and this is common to the three projects reported here having an experimental component (McCaskill, Rozenberg and Amos). Beyond Europe, the group of Wisconsin are prominent in exploiting a surface based approach to intermolecular DNA Computing using DNA Chips. Finally, supramolecular DNA Computing, as pioneered by Eric Winfree, harnesses the process of self-assembly of rigid DNA molecules with different sequences to perform computations. The connection with nanomachines and nanosystems is then clear and will become more pervasive in the near future.

In my view, DNA Computation is exciting and should be more substantially funded in Europe for the following reasons:

- it opens the possibility of a simultaneous bootstrapping solution of future computer design, construction and efficient computation
- it provides programmable access to nanosystems and the world of molecular biology, extending the reach of computation
- it admits complex, efficient and universal algorithms running on dynamically constructed dedicated molecular hardware
- it can contribute to our understanding of information flow in evolution and biological construction
- it is opening up new formal models of computation, extending our understanding of the limits of computation.

The difference with Quantum Computing is dramatic. Quantum Computing involves high physical technology for the isolation of mixed quantum states necessary to implement (if this is scalable) efficient computations solving combinatorially complex problems such as factorization. DNA Computing operates in natural noisy environments, such as a glass of water. It involves an evolvable platform for computation in which the computer construction machinery itself is embedded. Embedded computing is possible without electrical power in microscopic, error prone and real time environments, using mechanisms and technology compatible with our own make up. Because DNA Computing is linked to molecular construction, the computations may eventually also be employed to build three dimensional self-organizing partially electronic or more remotely even quantum computers. Moreover, DNA Computing opens computers to a wealth of applications in intelligent manufacturing systems, complex molecular diagnostics and molecular process control.

The papers in this section primarily deal with Biomolecular Computing. The first contribution outlines the European initiative in coordinating Molecular Computing (EMCC). Three groups present their multidisciplinary projects involving joint theoretical and experimental work. Two papers are devoted to extending the range of formal models of computation. The collection concludes with a small sampler from the more established approach to biologically inspired computation using neural network models. It is interesting that one of these contributions addresses the application of neural modelling to symbolic information processing. However, the extent to which informational biomolecules play a specific role in long term memory and the structuring of the brain, uniting neural and molecular computation, still awaits clarification.

Please contact:
John McCaskill – GMD
Tel: +49 2241 14 1526
E-mail: mccaskill@gmd.de

Articles in this section:

The European Molecular Computing Consortium

by Grzegorz Rozenberg

The rapidly expanding research on DNA Computing in the US and in Japan has already been channelled there into national projects with very substantial financial support. The growth of DNA Computing in Europe has been somewhat slower, although a number of European researchers have participated in the development of DNA Computing from the very beginning of this research area. In 1998 a number of research groups in Europe took the initiative to create the European Molecular Computing Consortium (EMCC) officially established in July 1998 during the DNA Computing Days organized by the Leiden Center for Natural Computing in Leiden.

The EMCC is a scientific organisation of researchers in DNA Computing and is composed of national groups from 11 European countries. The EMCC activities are coordinated by the EMCC Board consisting of: Grzegorz Rozenberg (The Netherlands) - director, Martyn Amos (United Kingdom) - deputy director, Giancarlo Mauri (Italy) - secretary, and Marloes Boon-van der Nat (The Netherlands) - administrative secretary. The purpose of the EMCC is best expressed in its official document 'Aims and Visions' which now follows.

Countries currently participating in the European Molecular Computing Consortium.

Molecular computing is a novel and exciting development at the interface of Computer Science and Molecular Biology. Computation using DNA or proteins, for example, has the potential for massive parallelism, allowing trillions of operations per second. Such a parallel molecular computer will have huge implications, both for theoreticians and practitioners. Traditional definitions of computation are being re-defined in the light of recent theoretical and experimental developments. Although thriving, the field of molecular computing is still at an early stage in its development, and a huge and concerted effort is required to assess and exploit its real potential.

The European Molecular Computing Consortium has been created in order to co-ordinate, foster and expand research in this exciting new field, especially in Europe. The EMCC is the result of discussions between different research groups in nine different European countries. A key function of the consortium is to foster co-operation between scientific, technological and industrial partners. A particular effort will be made to create genuinely multi-disciplinary co-operation between Computer Science, Molecular Biology, and other relevant scientific areas. The EMCC will organize various research-enhancing activities such as conferences, workshops, schools, and mutual visits that will provide forums for the exchange of results, and for establishing or strengthening existing co-operations in the field of molecular computing. The EMCC will also actively seek to promote the field, both within the scientific community and to the public at large, via scientific publications, seminars, workshops and public lectures. All participating sites will make the utmost effort to develop their theoretical and laboratory resources. It is hoped that all of these combined efforts will allow the field of molecular computing to thrive in Europe.

The EMCC will also strive to establish and maintain fruitful cooperation with researchers in the area of molecular computing from outside Europe, and in particular with the project 'Consortium for Biomolecular Computing' in the US and with the Japanese 'Molecular Computer Project'.

Links:
EMCC web page:
http://www.csc.liv.ac.uk/~emcc/

Please contact:
Grzegorz Rozenberg – Leiden Center for Natural Computing, Leiden University
Tel: +31 71 5277061/67
E-mail: rozenber@liacs.nl

Configurable DNA Computing

by John McCaskill

The DNA Computing Project carried out at the GMD Biomip Institute, aims at making molecular systems more programmable. Computer scientists, chemists, molecular biologists, physicists and microsystem engineers are working together to produce both a technological platform and theoretical framework for feasible and evolvable molecular computation.

Although the massive parallelism of DNA in solution is impressive (more than 1020 bytes of active memory per liter) and the energy consumption is very low, the ultimate attraction of DNA-Computers is their potential to design new hardware solutions to problems. Unlike conventional computers, DNA computers can construct new hardware during operation. Thus, the closest point of contact to electronic computing involves hardware design, in particular reconfigurable hardware design, rather than conventional parallel algorithms or languages. Molecular computers can be constructed reversibly in flow systems, where an exchange of DNA populations is possible. Rapid hardware redesign opens the door to evolving computer systems, so that configurable DNA Computing also aims at harnessing evolution for design and problem solving. Because of the huge information storage potential of aqueous solutions containing DNA, comparatively low flow rates suffice for massively parallel processing so that synthetic DNA can be treated as an affordable, easily degradable resource. Sequence complexity on the other hand is expensive to purchase but is generated within the DNA Computer, starting from simple sequence modules.

The GMD DNA Computing Project is multidisciplinary in scope, aiming at making molecular systems more programmable. Computer scientists, chemists, molecular biologists, physicists and microsystem engineers are working together to produce both a technological platform and theoretical framework for an effective use of molecular computation. The initial barrier is the issue of complexity: just how scalable and programmable are the basic hybridisation processes underlying DNA Computing? The group has devised an optically programmable, scalable concept for DNA Computing in microflow reactors,

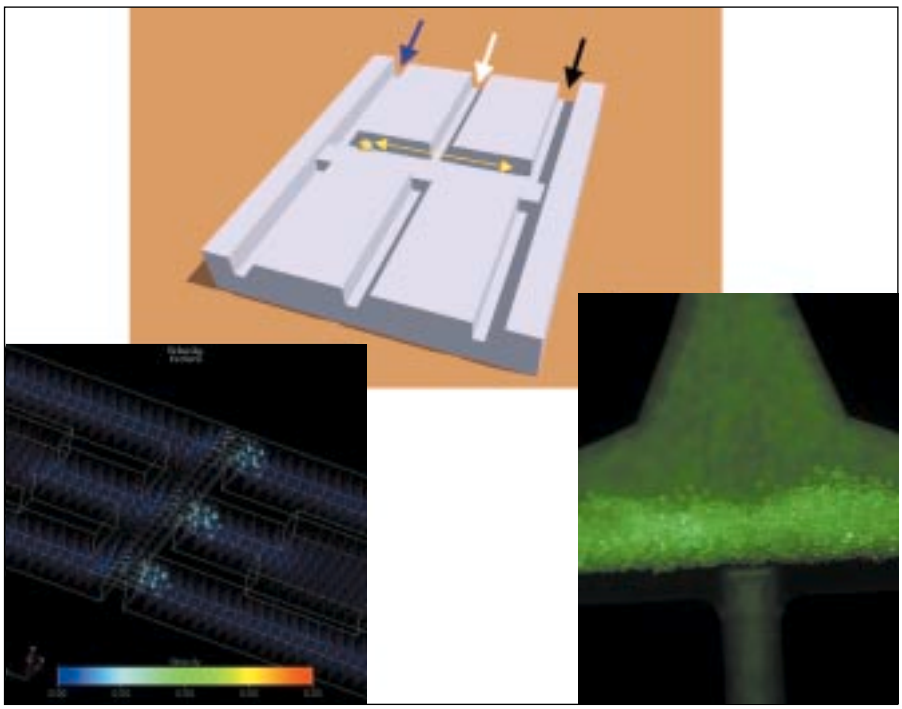


Figure 1: Massively parallel subset selection module for DNA Computing: The three images show a schematic, hydrodynamic simulation and fluorescence image of a microstructured selection module made at the GMD. The concept is that a subset of a mixed population of DNA flowing through the left hand side of the reactor binds to complementary DNA sequence labels attached to magnetic beads. When the beads are transferred (synchronously for all such modules) to the right hand side, they enter a denaturing solution which causes the bound DNA subset to be released. The released DNA subset solution is neutralised before being subjected to further modules. The hydrodynamic flow in one such module is seen in the color coded image on the left. Continuous flow preserves the integrity of the two different chemical environments in close proximity in each module. The fluorescence image shows DNA hybridized to the DNA labels on magnetic beads in such a microreactor, allowing the DNA processing to be monitored.

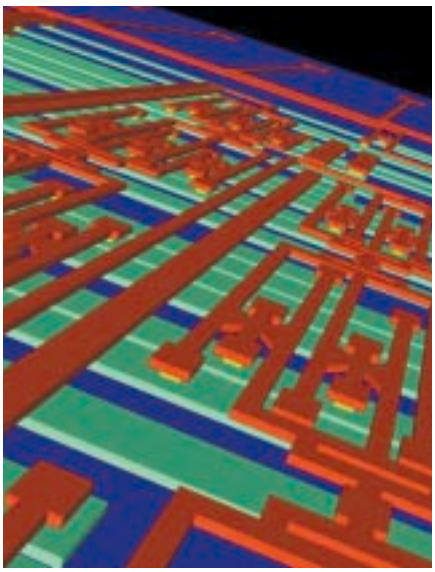
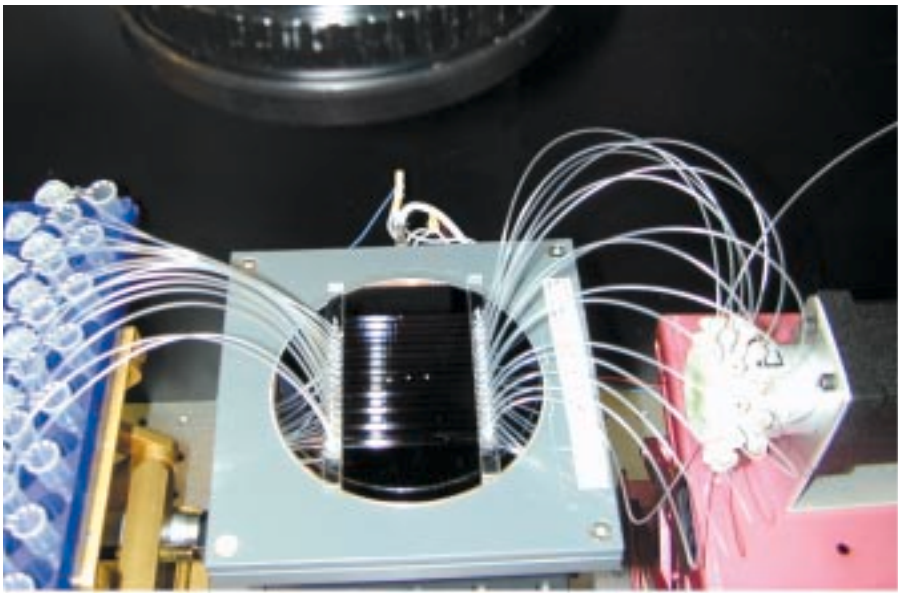
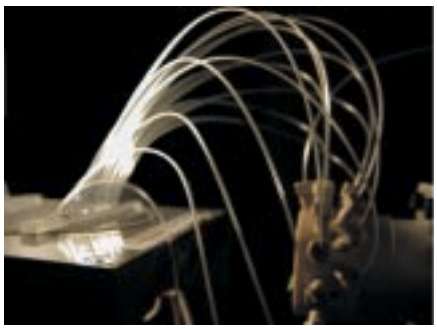
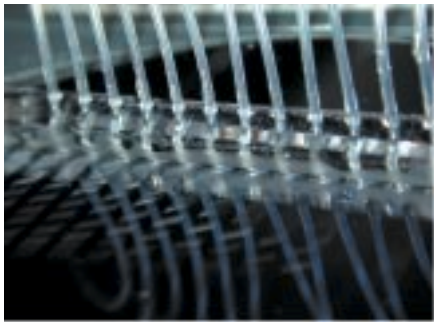


Figure 2: Medium Scaleup Programmable Microflow System for Benchmark Problem: Portion of the two layer mask design structure for integrating many selection modules (see Fig. 1) to solve a combinatorial optimization problem - in this case maximal clique. The structures are etched into a silicon substrate on the top and bottom sides (green and red), with through connections at the sites of squares. The microflow reactor is sealed at top and bottom with anodically bonded pyrex wafers and fitted with connecting tubing as shown in Fig. 3. This particular microreactor design can solve any intermediate scaled instance of the clique problem up to N=20. Which instance is programmed optically by directing the attachment of DNA to beads.







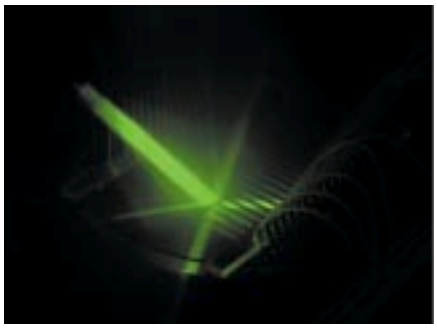


Figure 3: Experimental DNA Computing Apparatus: The left photo shows a microreactor with attached tubing connected to a multivalve port on the right and liquid handling system (not shown). The microreactor is imaged under green laser light to detect fluorescence at different locations stemming from DNA molecules, monitoring the time course of the computation. A microscope setup, not shown, is also used for reading and writing. The images on the right show close ups of the fluidic connections to the microreactor and of the laser illumination of the wafer.

designed to be extensible to allow full evolutionary search (see below). No flow switching is required (since this is not currently scalable) in this dataflow like architecture. Currently the microreactor geometry is fixed to evaluate scalability on a benchmark problem: Maximum Clique. Specific problem instances are reconfigured optically.

Effective DNA Computing is dependent on the construction of a powerful interface to the molecular world. In this project, the interface involves configurable microreactors with photochemical input and fluorescence readout down to the single molecule level. Fluorescence detection is the most sensitive spectroscopic technique for detecting the presence of specific molecules in solution and it can be employed as an imaging tool to gain vast amounts of information in parallel about the status of computations or the final answer. Consistent with our perspective on DNA Computing as hardware design, we employ photolithographic techniques to program the attachment of DNA to specific mobile elements within the microreactor. At the spatial resolution and time scale required, photolithographic projection can be made completely dynamically programmable.

Directed molecular evolution provides a second stepping stone to exploiting more powerful algorithms in DNA Computing. The programming problem shifts to defining selection conditions which match the given problem. The strategy employed in this project is to program molecular survival by employing flow networks involving sequence-dependent molecular transfers in series and parallel. Feedback loops and amplification modules have been designed and will be introduced into the computer in due course to complete the integration with molecular evolution. Reconfiguration and evolution of the compartmentation and flow network is planned at a second phase in the technology development to increase the general programmability of the computer.

DNA-Computers can produce their calculated output in functional molecular form for direct use. Application areas include the pharmaceutical and diagnostics industry (where molecular complexity requires increasingly sophisticated algorithms for combinatorial libraries construction and readout implemented in molecular hardware), nanotechnology and high performance computing for coding, associative retrieval and combinatorial optimisation.

Status of the Project

A scalable architecture for configurable DNA Computing has been developed for fully optically programmable solution of the NP-Complete test problem "Maximal Clique". Individual modules (strand transfer and amplification) have been implemented in microreactors and the optical programmability has been developed and demonstrated. Spatially resolved single molecule fluorescence detection has been developed to allow an on-line readout of information in complex DNA populations. Microsystem integration has proceeded up to the level of 20x20x3 selection modules, but current evaluation work is of a 6x6x3 DNA microprocessor. A DNA library for

medium scale combinatorial problems (N=32) has been designed and the lower portion of it constructed. The potential of iterated modular evolution in DNA Computing has been evaluated on a massively parallel reconfigurable electronic computer (NGEN). All of these developments have alternative technological applications beyond the immediate range of DNA Computing.

Cooperation

Collaborations are being fostered with the European Molecular Computing Consortium (EMCC), in particular University of Leiden, Holland (Prof. G. Rozenberg, Prof. H. Spaink); North Rhine Westfalia initiative in Programmable Molecular Systems: including University of Cologne (Prof. Howard), University Dortmund (Prof. Banzhaf) and University Bochum (Prof. Kiedrowski).

Links:
BioMIP website:
http://www.gmd.de/BIOMIP

Please contact:
John S. McCaskill – GMD
Tel: +49 2241 14 1526
E-mail: McCaskill@gmd.de

Molecular Computing Research at Leiden Center for Natural Computing

by Grzegorz Rozenberg

Molecular computing is an exciting and fast growing research area. It is concerned with the use of (bio)molecules and biochemical processes for the purpose of computing. Although it is centered around computer science, molecular computing is a very interdisciplinary area with researchers from computer science, mathematics, molecular biology, crystallography, biochemistry, physics, etc participating in it.

Molecular computing has the potential to resolve two well recognized obstacles of silicon based computer technology: miniaturization and massive parallelism. Through molecular computing one 'descends' to the nano-scale computing which solves the miniaturization problem. Since, eg, a single drop of solution can contain trillions of DNA molecules, and when an operations is performed on a tube containing DNA molecules then it is performed on every molecule in the tube, massive parallellism is obtained on a grand scale.

DNA computing (the area of molecular computing where one considers DNA molecules) offers a number of other features which make it an attractive alternative (or supplementary) technology to modern silicon computing. These features include very impressive energy efficiency and information density.

One may say that the main thrust of the current research in molecular computing is the assessment of its full potential. The results obtained to date are cautiously optimistic. In particular, the conceptual understanding and experimental testing of basic principles achieved is already quite impressive.

The research on molecular computing at Leiden University takes place at the Leiden Center for Natural Computing (LCNC), an interdepartmental institute of the Faculty of Mathematics and Natural Sciences. Molecular Computing is one of the main research programs of LCNC and it is multidisciplinary, with three participating groups: Leiden Institute for Advanced Computer Science (Prof. G. Rozenberg), Institute of Molecular Plant Biology (Prof. H. Spaink), and Department of Biophysics (Prof. T. Schmidt). The research on molecular computing also involves the Evolutionary Algorithms research program of LCNC (Prof. J. Kok and Prof. T. Baeck).

The two main research lines on molecular computing within LCNC are:
(1) models and paradigms for molecular computing where the following theoretical topics are currently under investigation:

- splicing systems
- forbidding-enforcing systems
- molecular landscapes
- membrane systems
- linear self-assembly of complex DNA tiles
- models for evolutionary DNA computing
- models for gene assembly in ciliates (DNA computing *in vivo*).

(2) design of laboratory experiments testing models for molecular computing. Current laboratory experiments include:

- the use of plasmids as data registers for DNA computing
- the use of molecules other than DNA for molecular computing
- design of molecules for evolutionary DNA computing
- DNA computing methods based on single molecule detection systems
- experimental confirmation of gene assembly operations in ciliates.

In our research, both theoretical and experimental, we cooperate with a number of research centers around the world - in particular with: University of Colorado at Boulder (USA), Princeton University (USA), California Institute of Technology (USA), State University of New York at Binghamton (USA), Turku Center for Computer Science (Finland), Romanian Academy (Romania) and Waseda University (Japan).

Here at LCNC we feel that the interdisciplinary research on molecular computing has significantly deepened our understanding of computation taking place all around us: in computer science, biology, physics, etc. We certainly look forward to many years of exciting and challenging research.

Links:
Leiden Center for Natural Computing:
http://www.wi.leidenuniv.nl/~lcnc/
Grzegorz Rozenberg's homepage:
http://www.liacs.nl/~rozenber

Please contact:
Grzegorz Rozenberg – Leiden Center for Natural Computing, Leiden University
Tel: +31 71 527 70 61/67
E-mail: rozenber@liacs.nl

Cellular Computing

by Martyn Amos and Gerald G. Owenson

The recent completion of the first draft of the human genome has led to an explosion of interest in genetics and molecular biology. The view of the genome as a network of interacting computational components is well-established, but researchers are now trying to reverse the analogy, by using living organisms to construct logic circuits.

The cellular computing project is the result of collaboration between teams at the University of Liverpool (Alan Gibbons, Martyn Amos and Paul Sant) and the University of Warwick (David Hodgson and Gerald Owenson). The field emerged in 1994 with the publication of Adleman's seminal article, in which he demonstrated for the first time how a computation may be performed at a molecular level. Our group contributed to the development of the area by describing a generalization of Adleman's approach, proposing methods for assessing the complexity of molecular algorithms, and carrying out experimental investigations into error-resistant laboratory methods. This work quickly confirmed that the massively-parallel random search employed by Adleman would greatly restrict the scalability of that approach. We therefore proposed an alternative method, by demonstrating how Boolean logic circuits may be simulated using operations on strands of DNA.

Our original intention was to implement the Boolean circuit in vitro (ie in a laboratory 'test tube'). However, after much consideration we decided to attempt a rather more ambitious approach, harnessing genetic regulatory mechanisms *in vivo*, within the living *E. coli.* bacterium.

The central dogma of molecular biology is that DNA (information storage) is copied, producing an RNA message (information transfer). This RNA then acts as the template for protein synthesis. The basic 'building blocks' of genetic information are known as genes. Each gene codes for a specific protein which may be turned on (expressed) or off (repressed) when required. In order for the DNA sequence to be converted into a protein molecule, it must be read (transcribed) and the transcript converted (translated) into a protein. Each step of the conversion from stored information (DNA) to protein synthesis (effector) is itself affected or catalyzed by other molecules. These molecules may be enzymes or other compounds (for example, sugars) that are required for a process to continue. Consequently, a loop is formed, where products of one gene are required to produce further gene products, and may even influence that gene's own expression.

The interaction of various components of the E. coli genome during development may be described in terms of a logic circuit. For example, a gene's expression may require the presence of two particular sugars. Thus, we may view this gene in terms of the Boolean AND function, where the presence or absence of the two sugars represent the two inputs to the function, and the expression or repression of the gene corresponds to the function's output. That gene's product may then be required for the expression (or repression) of another different gene, so we can see how gene products act as 'wires', carrying signals between different 'gates' (genes).

In order to implement our chosen logic circuit, we select a set of genes to represent gates, ensuring that the inter-gene dependencies correctly reflect the connectivity of the circuit. We then insert these genes into a bacterium, using standard tools of molecular biology. These insertions form the most time-consuming component of the entire experimental process, as the insertion of even a single gene can be problematic. However, once the entire set of genes is present in a single

PCR equipment, used to amplify DNA samples.

Postdoctoral researcher Gerald Owenson at the lab bench.

colony of bacteria we have an unlimited supply of 'biological hardware' at our disposal. The inputs to the circuit are set by enforcing in the cell's environment the presence or absence of various compounds that affect the expression of the genes representing the first level gates. Then, essentially, the development of the cell and the complex regulatory processes involved, simulate the circuit, without any additional human intervention. This last point is crucial, as most existing proposals for molecular computing require a series of manipulations to be performed by a laboratory technician. Each manipulation reduces the probability of success for the experiment, so the ideal situation is a 'one-pot' reaction, such as the one we propose.

We have recently begun work on simulating a small (3 gate) circuit of NAND gates in vivo. We believe that, within the next three years, the introduction of human-defined logic circuits into living bacteria will be a reality. Of course, such implementations will never rival existing silicon-based computers in terms of speed or efficiency. However, our goal differs from that of a lot of groups in the community, who insist that DNA-based computers may eventually rival existing machines in domains like encryption. Rather, we see the potential applications of introducing logic into cells as lying in fields such as medicine, agriculture and nano-technology. The currenct 'state of the art' in this area has resulted in the reprogramming of E. coli genetic expression to generate simple oscillators. This work, although 'blue sky' in nature, will advance the field to a stage where cells may be reprogrammed to give them simple 'decision making' capabilities.

Our group is a member of the European Molecular Computing Consortium (see article on page 32). We acknowledge the support of the BBSRC/EPSRC Bioinformatics Programme.

Links:
Publications:
http://www.csc.liv.ac.uk/~martyn/pubs.html
The Warwick group:
http://www.bio.warwick.ac.uk/hodgson/index.html
EMCC: http://www.csc.liv.ac.uk/~emcc

Please contact:
Martyn Amos – School of Biological Sciences and Department of Computer Science, University of Liverpool
Tel: +44 1 51 794 5125
E-mail: mamos@liv.ac.uk
or Gerald G. Owenson
Department of Biological Sciences, University of Warwick
Tel: +44 2 47 652 2572
E-mail: G.Owenson@warwick.ac.uk

Research in Theoretical Foundations of DNA Computing

by Erzsébet Csuhaj-Varjú and György Vaszil

DNA computing is a recent challenging area at the interface of computer science and molecular biology, providing unconventional approach to computation. The Research Group on Modelling Multi-Agent Systems at SZTAKI develops computational paradigms for DNA computing: theoretical models for test tube systems and language theoretical frameworks motivated by the phenomenon of Watson-Crick complementarity.

The famous experiment of Leonard Adleman in 1994, when he solved a small instance of the Hamiltonian path problem in a graph by DNA manipulation in a laboratory, seeded his ideas on how to construct a molecular computer. Following on from this, Richard J. Lipton proposed a kind of programming language for writing algorithms dealing with test tubes. The basic primitive of the proposed formalism is the test tube, a set or a multiset (the elements are with multiplicities) of strings of an alphabet, and the basic operations correspond to operations with test tubes, merge (put together the contents of two test tubes), separate (produce two test tubes from one tube, according to a certain criteria), amplify (duplicate a tube), and check whether a tube is empty or not.

Test tube systems based on splicing and test tube systems based on cutting and recombination operations are theoretical constructs realizing the above idea. These are distributed communicating systems built up from computing devices (test tubes) based on operations motivated by the recombinant behaviour of DNS strands. These operations are applied to the objects in the test tubes (sets of strings) in a parallel manner and then the results of the computations are redistributed according to certain specified criteria (input/output filters associated with the components) which allow only specific parts of the contents of the test tube to be transferred to the other tubes. Test tube systems based on splicing were presented in 1996 by Erzsébet Csuhaj-Varjú, Lila Kari and Gheorghe Paun, an another model in 1997 by Rudolf Freund, Erzsébet Csuhaj-Varjú and Franz Wachtler. Both models proved to be as powerful as Turing machines, giving a theoretical proof of the possibility to design universal programmable computers with such architectures of biological computers based on DNA molecules. Since that time, the theory of theoretical test tube systems has been extensively investigated by teams and authors from various countries.

The main questions of research into models of test tube system are – among other things – comparisons of the variants with different basic operations motivated by the behaviour of DNA strands (or DNA-related structures) according to their computational power, programmability, simplicity of the filters, topology of the test tubes and approachability of intractable problems. Our investigations follow this line.

At present we explore test tube systems with multisets of objects (symbols, strings, data structures) and operations modelling biochemical reactions. In addition to the computational power, the emphasis is put on elaborating complexity

notions for these devices. As a further development, we extended the concept of the test tube system to a pipeline architecture allowing objects to move in the system.

The other important direction of our research is to study language theoretical models motivated by Watson-Crick complementarity, a fundamental concept in DNA Computing. According to this phenomenon, when bonding takes place (supposing that there are ideal conditions) between two DNA strands, the bases opposite each other are complementary.

A paradigm in which Watson-Crick complementarity is viewed in the operational sense was proposed for further consideration by Valeria Mihalache and Arto Salomaa in 1997. According to the proposed model, a 'bad' string (a string satisfying a specific condition, a trigger) produced by a generative device induces a complementary string either randomly or guided by a control device. The concept can also be interpreted as follows: in the course of a developmental or computational process, things can go wrong to such an extent that it is advisable to switch to the complementary string, which is always available.

In close cooperation with Prof. Arto Salomaa (Turku Centre for Computer Science), we study properties of Watson-Crick DOL systems, a variant where the underlying generative device is a parallel string rewriting mechanism modelling developmental systems. Recently, we have demonstrated the universal power of some important variants of these systems. In the frame-work of our joint research, we have introduced and now explore networks of Watson-Crick DOL systems, where the communication is controlled by the trigger for conversion to the complementary form: whenever a "bad" string appears at a node, the other nodes receive a copy of its corrected version. In this way, the nodes inform each other about the correction of the emerging failures. In addition to the computational power of these constructs (which proved to be computationally complete in some cases), we have achieved interesting results in describing the dynamics of the size of string collections at the nodes. We have been dealing with challenging stability issues, for example, detecting so-called black holes in the network (nodes which never emit any string). Our future plans aim at comparisons of these devices with different underlying mechanisms and different qualitative/ quantitative conditions employed as triggers, according to computational power, stability and complexity issues, including complexity measures different from the customary ones.

We have been in contact and cooperation with several leading persons from the European Molecular Computing Consortium. The group is open for any further cooperation.

Links:
http://www.sztaki.hu/mms

Please contact:
Erzsébet Csuhaj-Varjú – SZTAKI
Tel: +36 1 4665 644
E-mail: csuhaj@sztaki.hu
or György Vaszil – SZTAKI
Tel: +36 1 4665 644
E-mail: vaszil@sztaki.hu
or Arto Salomaa – Turku Centre
for Computer Science
Tel: +358 2 333 8790
E-mail: asalomaa@utu.fi

Representing Structured Symbolic Data with Self-organizing Maps

by Igor Farkas

Artificial self-organizing neural networks - especially self-organizing maps (SOMs) - have been studied for several years at the Institute of Measurement Science of the Slovak Academy of Sciences in Bratislava. The SOM research oriented to structured symbolic data is the most recent and is being supported by the Slovak Grant Agency for Science (VEGA) within a project of the Neural Network Laboratory. The goal of this research is to assess the potential of SOM to represent structured symbolic data, which is associated with various high-level cognitive tasks.

Various information processing tasks can be successfully handled by neural network models. Predominantly, neural nets have proven to be useful in tasks which deal with real-valued data, such as pattern recognition, classification, feature extraction or signal filtering. Neural networks have been criticised as being unsuitable for tasks involving symbolic and/or structured data, as occur in some cognitive tasks (language processing, reasoning etc.). These tasks were previously tackled almost exclusively by classical, symbolic artificial-intelligence methods. However, facing this criticism, during the last decade a number of neural architectures and algorithms were designed to demonstrate that neural networks also possess the potential to process symbolic structured data. The best known examples of such models include recursive auto-associative memory (RAAM), tensor product based algorithms or models that incorporate synchrony of neurons' activation. For example, the RAAM, being a three-layer feed-forward neural network, can be trained to encode various data structures presented to the network (such as 'A(AB)', where each of the two symbols is encoded as a binary vector). These encoding (representations) are gradually formed in the layer of hidden units during training. However, this is often a time-consuming process (moving target problem). Despite this drawback, RAAM has become one of the standard neural approaches to represent structures.

For the learning method to work, the representations of structures generated must be structured themselves to enable structure-sensitive operations. In other words, the emerged representations must be systematic (compositional), requiring

(in some sense) that similar structures should yield similar encoding.

Objectives

We focused on an alternative model that incorporates a neural network having the capability to generate structured representations. As opposed to RAAM, the main advantage of our model consists in fast training, because it is based on a rather different concept, employing unsupervised learning (self-organisation). The core of the model are the self-organising maps (SOMs) which are well-known as a standard neural network learning algorithm used for clustering and visualisation of high-dimensional data patterns. The visualisation capability of the SOM is maintained by its unique feature – topographic (non-linear) transformation from input space to output units. The latter are arranged in a regular grid – which implies that data patterns originally close to each other are mapped onto nearby units in the grid. We exploited this topographic property analogously in representing sequences.

Results

In our approach, each symbolic structure is first decomposed onto a hierarchy of sequences (by some external parsing module) which are then mapped onto a hierarchy of SOMs. For instance, a structure 'A(AB)' which can be viewed as a tree of depth two, is decomposed into two sequences 'Ao' and 'AB', the latter belonging to the first level in the hierarchy and the former to the second higher level. In this notation, symbol 'o' is simply treated as another symbol and represents a subtree. In general, structures with maximum depth 'n' require 'n' SOMs to be used and stacked in a hierarchy. Hence, the representation of a structure consists in simultaneous representations of associated sequences emerging after decomposition, distributed across the SOMs at all levels.

How is a sequence represented in the SOM? The idea comes from Barnsley's iterated function systems (IFS) which can be used for encoding (and decoding) symbolic sequences as points in a unit hypercube. In the hypercube, every symbol is associated with a particular vertex (the dimension of the hypercube must be sufficiently large to contain enough vertices for all symbols in alphabet). During the encoding process (going through the sequence symbol by symbol), a simple linear transformation is recursively applied resulting in 'jumps' within the hypercube, until the final point of this series (when the end of sequence is reached) becomes an IFS-representing point of the sequence. As a result, we get as many points as there are sequences to be encoded. The important feature of these IFS-based sequence representations is the temporal analogue of the topographic property of a SOM. This means that the more similar two sequences are in terms of their suffices, the more closely are their representations placed in the hypercube. The encoding scheme described has been incorporated in training the SOM. The number of units employed has to be selected according to the length of sequences being used: the longer the sequences, the finer the resolution needed and the more units are required. The training resulted in a topographically ordered SOM, in which a single sequence can afterwards be represented by the winning unit in the map. If we do not use winner co-ordinates as output (as commonly done), but take the global activity of all units in the map (as a Gaussian-like profile, with its peak centred at the winner's position in the grid), then we can simultaneously represent multiple sequences in one map.

We have applied the above scheme on simple two- and four-symbol structures of depth less than four, and the preliminary results (evaluted using a hierarchical cluster diagram of representations obtained) show that the generated representations display the property of systematic order. However, to validate this approach, we shall need to test how it scales with larger symbol sets as well as with more complex structures (both in terms of 'height' and 'width').

Please contact:
Igor Farkaso – Institute of Measurement Science, Slovak Academy of Sciences
Tel: +421 7 5477 5938
E-mail: farkas@neuro.savba.sk

Neurobiology keeps Inspiring New Neural Network Models

by Lubica Benuskova

Biologically inspired recurrent neural networks are investigated at the Slovak Technical University in Bratislava. This project, supported by the Slovak Scientific Grant Agency VEGA, builds on the results of the recently accomplished Slovak-US project 'Theory of neocortical plasticity' in which theory and computer simulations were combined with neurobiological experiments in order to gain deeper insights into how real brain neurons learn.

The human brain contains several hundred thousand millions of specialised cells called neurons. Human and animal neurons share common properties, thus researchers often use animal brains to study specific questions about processing information in neural networks. The aim is to extrapolate these findings to ideas about how our own brains work. This understanding can be crucial not only for the development of medicine but also for computer science. Of course, the validity of extrapolating findings in animal studies depends on many aspects of the problem studied. We studied a certain category of plastic changes occuring in neurons when an animal was exposed to a novel sensory experience. We work with the evolutionarily youngest part of the brain, ie neocortex, which is involved mainly in the so-called cognitive functions (eg, perception, association, generalisation, learning, memory, etc).

Neurons emit and process electric signals. They communicate these signals through specialized connections called synapses. Each neuron can receive information from as well as send it to thousands of synapses. Each piece of information is transmitted via a specific and well defined set of synapses. This set of synapses has its anatomical origin and target as well as specific properties of signal transmission. At present, it is widely accepted that origins and targets of connecting synapses are determined genetically as well as most of the properties of signal transmission. However, the efficacy of signal transfer at synapses can change throughout life as a consequence of learning.

In other words, whenever we learn something, somewhere in our neocortex, changes in the signal transfer functions of many synapses occur. These synaptic changes are then reflected as an increase or decrease in the neuron's response to a given stimulus. All present theories of synaptic learning refer to the general rule introduced by the Canadian psychologist Donald Hebb in 1949. He postulated that repeated activation of one neuron by another, across a particular synapse, increases its strength. We can record changes in neurons' responses and then make inferences about which synapses have changed their strengths and in which direction, whether up or down. For this inference, we need to introduce some reasonable theory, that is a set of assumptions and rules which can be put together into a model which simulates a given neural network and which, in simulation, reproduces the evolution of its activity. If the model works, it can give us deeper insights into what is going on in the real neural networks.

Objectives and Research Description

Experience-dependent neocortical plasticity refers to the modification of synaptic strengths produced by the use and disuse of neocortical synapses. We would like to contribute to an intense scientific effort to reveal the detailed rules which hold for modifications of synaptic connections during learning. Further, we want to investigate self-organising neural networks with time-delayed recurrent connections for processing time series of inputs.

Cortical neuron. Courtesy of Teng Wu, Vanderbilt University, USA.

Research Results

Experience-dependent neocortical plasticity was evoked in freely moving adult rats in the neocortical representation of their main tactile sense, i.e. whiskers. We developed a neural network model of the corresponding neuroanatomical circuitry. Based on computer simulations we have proposed which synapses are modified, how they are modified and why. For the simulation of learning, we used the theory of Bienenstock, Cooper and Munro (BCM). Originally, the BCM theory was introduced for the developing (immature) visual neocortex. They modelled experiments done on monkeys and cats. We have shown that the BCM rules apply also for the mature stage of the brain development, for a different part of the neocortex, and for the different animal species (rats). The main distinguishing feature of the BCM theory against other Hebbian theories of synaptic plasticity is that it postulates the existence of a shifting synaptic potentiation threshold, the value of which determines the sign of synaptic changes. The shifting threshold for synaptic potentiation is proportional to the average of a neuron's activity over some recent past. Prof. Ebner's team at Vanderbilt University in Nashville, TN, USA used an animal model of mental retardation (produced by exposure of the prenatal rat brain to ethanol) to show a certain specific impairment of experience-evoked neocortical plasticity. From our model, we have derived an explanation of this impaired plasticity in terms of an unattainably high potentiation threshold. Based on a comparison between computational results and experimental data, revealing a specific biochemical deficit in these faulty cortices, we have proposed that the value of the potentiation threshold depends also on a specific biochemical state of the neuron.

The properties of the self-organising BCM learning rule have inspired us to investigate the state space organisation of recurrent BCM networks which process time series of inputs. Activation patterns across recurrent units in recurrent neural networks (RNNs) can be thought of as sequences involving error back propagation. To perform the next-symbol prediction, RNNs tend to organise their state space so that 'close' recurrent activation vectors correspond to histories of symbols yielding similar next-symbol distributions. This leads to simple finite-context predictive models, built on top of recurrent activation, grouping close activation patterns via vector quantization. We have used the recurrent version of the BCM network with lateral inhibition to map histories of symbols into activation patterns of the recurrent layer. We compared the finite-context models built on top of BCM recurrent activation with those constructed on top of RNN recurrent activation vectors. As a test bed, we used complex symbolic sequences with rather deep memory structures. Surprisingly, the BCM-based model has a comparable or better performance than its RNN-based counterpart.

Please contact:
Lubica Benuskova – Slovak Technical University
Tel: +421 7 602 91 696
E-mail: benus@elf.stuba.sk,
http://www.dcs.elf.stuba.sk/~benus

Contribution to Quantitative Evaluation of Lymphoscintigraphy of Upper Limbs

by Petr Gebousky, Miroslav Kárny and Hana Křížová

Lymphoscintigraphy is an emerging technique for judging of the state of lymphatic system. It serves well for diagnosis of lymphedema. Its early and correct detection prevents a range of complications leading often to a full disability. A wider use of this technique is inhibited by a lack of a reliable quantitative evaluation of its results especially in difficult case of upper limbs in early stage of the disease. The modeling and subsequent identification can be exploited in order to improve the accuracy of diagnosis.

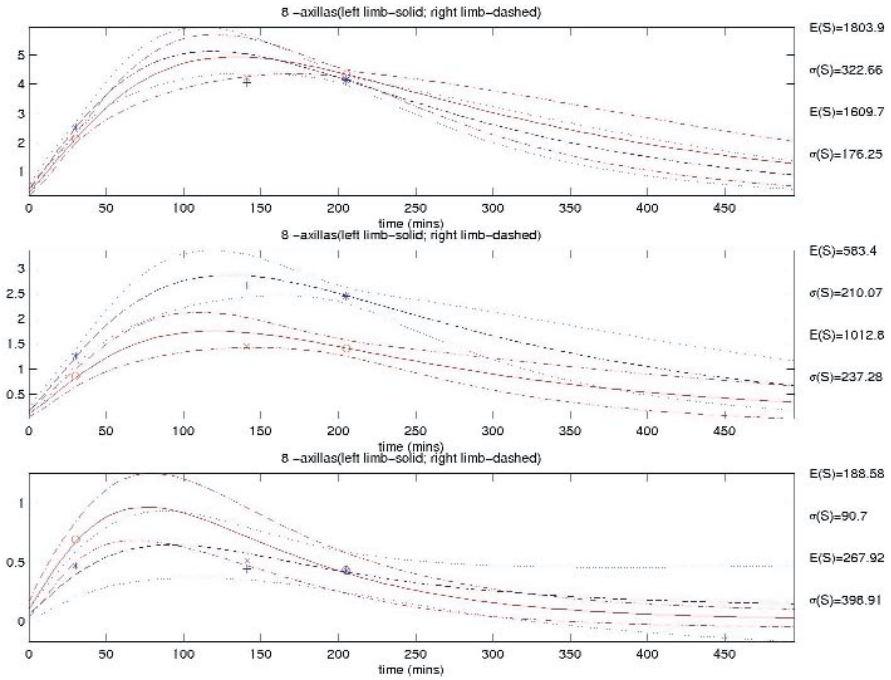
Lymphedema is a progressive condition characterized by four pathologic features: excess tissue protein, edema, chronic inflammation and fibrosis. Upper limb lymphedema is a frequent complication of breast cancer.

A variety of imaging modalities have been employed in the evaluation of the lymphatic system. Lymphoscintigraphy is a noninvasive technique in which radionuclides are used to image regional lymph drainage systems.

It provides information about lymph transportation, filtration and reticuloendothelial function in extremities.

Standard inspection method consists of administration of 25 MBq of Tc-99m labeled colloid administered to the 1st interstitial space of both hands. Static 1 min images are taken immediately, 30 & 180 min after administration, morphologically evaluated and arm results semi-quantitatively compared. No standard is available for healthy population.

The system identification is employed here for description of individual response of the patient. We tried to model the accumulation of the radioactive tracer in local regions of upper limb during lymphoscintigraphic procedure by a chain of simple identical compartmental models with a common time constant, unknown gain and unknown number of constituents. The chosen 3-parameters model was motivated by a standard modeling of distributed parameter systems. Their Bayesian estimator was designed that converts integral counts taken over regions of interests (ROI : forearm, upper arm, axilla) into posterior distribution of discretised common time constant, model order and gain. Each triple of parameters determines uniquely time-activity curve in individual ROIs together with the important parameter for diagnosis - area under curve (residence time). This allows us to compute posterior distribution of residence time in each ROI.



Time-activity curves (with shapes) estimated from 2 measurements (the 1st & 3rd) in various ROIs where responses in left and right limb are compared. Red colour refers to left blue to right limb. The estimated residence times are written on the right, the first one for left limb.

Having in these distributions in hands, we can (i) judge (in)significance of differences between both arms (ii) study quantitative relationships between disease staging and measurements (iii) optimize measurements moments.

The proposed model and data processing were tested on data of 24 patients that were measured more often than usual (up to 5 measurements). Real data was provided by Department of Nuclear Medicine FNM. The evaluation results indicate adequacy of the adopted model and encourage us to follow the direction outlined under (i) to (iii).

Please contact:
Petr Gebousky – CRCIM (UTIA)
Tel: +420 2 66052583
E-mail: gebousky@utia.cas.cz

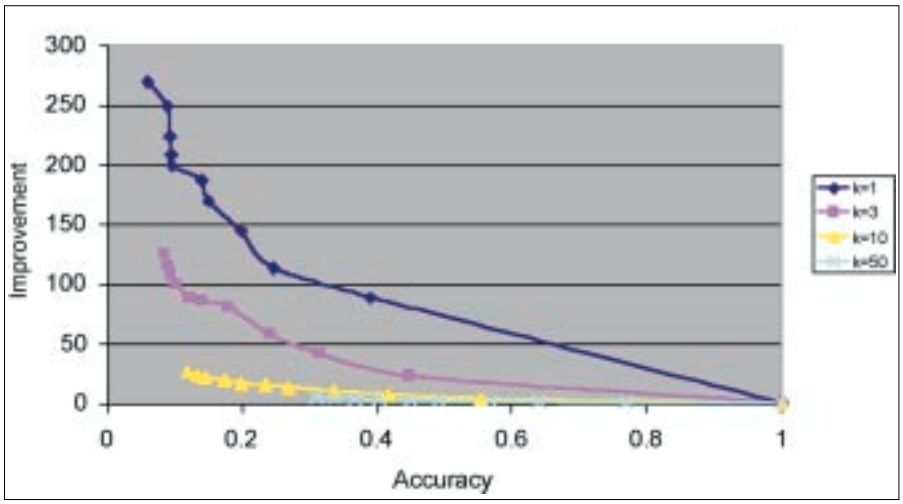
Approximate Similarity Search

by Giuseppe Amato

Similarity searching is fundamental in various application areas. Recently it has attracted much attention in the database community because of the growing need to deal with large volume of data. Consequently, efficiency has become a matter of concern in design. Although much has been done to develop structures able to perform fast similarity search, results are still not satisfactory, and more research is needed. The performance of similarity search for complex features deteriorates and does not scale well to very large object collections. Given the intrinsically interactive nature of the similarity-based search process, the efficient execution of elementary queries has become even more important, and the notion of approximate search has emerged as an important research issue.

Contrary to traditional databases, where simple attribute data are used, the standard approach to searching modern data repositories, such as multimedia databases, is to perform search on characteristic features that are extracted from information objects. Features are typically high dimensional vectors or some other data items, the pairs of which can only be compared by specific functions. In such search environments, exact match has little meaning; thus, concepts of similarity are typically applied. In order to illustrate this, let us consider an image data repository. It is clear that images are not atomic symbols, so equality is not a particularly realistic predicate. Instead, search tends to be based on similarity, because resemblance is more important than perfectly matching bit patterns. On the other hand, all that is similar is not necessarily relevant, so this paradigm tends to entail the retrieval of false positives that must be manually discarded. In other words, the paradigm rejects the idea that queries may be expressed in terms of necessary and sufficient conditions that will determine exactly which images we wish to retrieve. Instead, a query is more like an information filter, defined in terms of specific image properties, which reduce the user's task by providing only a small number of candidates to be examined. It is more important that candidates that are likely to be of interest are not excluded than it is that possibly irrelevant candidates be included.

We have investigated the problem of approximated similarity search for the range and nearest neighbour queries in the environment of generic metric spaces. From a formal point of view, the mathematical notion of metric space provides a useful abstraction of similarity or nearness. We modified existing tree-based similarity search structures to achieve approximate results at substantially lower costs. In our proposal, approximation is controlled by an external parameter of proximity of regions that allows avoiding access to data regions that possibly do not contain relevant objects. When the parameter is zero, precise results are guaranteed, and the higher the proximity threshold, the less accurate the results are and the faster the query is executed.



Evaluation of approximate k nearest neighbours search.

In order to have good quality results an accurate proximity measure is needed. The technique that we use to compute proximity between regions adopts a probabilistic approach: given two data regions, it is able to determine the probability that the intersection of these two regions contains relevant data objects. In fact, note that, even if two regions overlap, there is no guarantee that objects are contained in their intersection. Extensive experimental tests have shown a high reliability of this approach that gave a substantial contribution to the quality of the approximate results and to the efficiency of the approximate similarity search algorithm. We applied this idea for the similarity range and the nearest neighbours queries and verified its validity on real-life data sets. Improvements of two orders of magnitude were achieved for moderately approximated search results.

The main contributions of our approach can be summarised as follows:

- A unique approximation approach has been applied to the similarity range and the nearest neighbours queries in metric data files. Previous designs have only considered the nearest neighbours search, sometimes even restricted to one neighbour.
- The approximation level is parametric, and precise response to similarity

queries is achieved by specifying zero proximity threshold.
- The approach to computation of proximity keeps the approximated results probabilistically bound. Experimental results demonstrate high precision and improvements of efficiency.
- We have experimentally demonstrated the importance of precise proximity measures, the application of which can lead to effective and efficient approximations of similarity search.
- Though implementation is demonstrated by extending the M-tree, the approach is also applicable to other similarity search trees at small implementation costs.

In the future, we plan to properly compare all existing approaches to approximation in uniform environment. We also hope to develop a system-user interface and apply the approach to real image and video archives. Finally, we intend to study the cases of iterative similarity search and complex approximated similarity search.

Other people that have contributed to this research are Pavel Zezula, Pasquale Savino, and Fausto Rabitti.

Please contact:
Giuseppe Amato – IEI-CNR
Tel: +39 050 315 2906
E-mail: G.Amato@iei.pi.cnr.it

Education of 'Information Technology Non-professionals' for the Development and Use of Information Systems

by Peter Mihók, Vladimír Penjak and Jozef Bucko

Information creating and processing in modern information systems requires increasing engagement of all users in the analysis and development stages. The article attempts to investigate how deep a knowledge of object oriented methodologies for the development of information systems is necessary for students of applied mathematics, economics and young business managers. We describe, in the context of the concept of student education developed at the Faculty of Sciences of P.J. Safárik University and Economical Faculty of Technical University in Ko_ice, our experience in this and formulate certain questions and problems which we believe to be relevant in this area.

The ability to use modern information and communication technologies efficiently is one of the fundamental requirements for graduates of all types of university level educational institutions. Future economists, managers, businessmen are facing a reality which encompasses the necessity of working in an environment of global integrated information systems (IIS) which are becoming an integral part of the global information society. Courses which support the use of modern information and communication technologies are designed to provide the students with a certain amount of theoretical fundamental knowledge but their main purpose is to impart practical knowledge and skills. The students are motivated to search the worldwide web for the newest information and make their own survey of the services and products offered in connection with direct (electronic) banking and acquaint themselves with the forms and principles of electronic commerce.

Object Oriented Methodologies and the Training of Information System Users

The development, maintenance and use of efficient integrated information systems is becoming a serious competitive advantage in economic competition. One of the reasons for this is the fact that the efficiency and effectively of these systems is directly dependent on the ability of information system users to provide the best possible specification of their requirements. In spite of the fact that the development and implementation of information systems is the domain of professional information technologists. it is impossible to build an integrated information system without close cooperation with the prospective system users whose numbers are increasing in a dramatic way. This forces us to make a new evaluation of the way the 'information technology non-professionals - IIS users' are trained and the way they should be trained to prepare them for these tasks. Since the last years of our century are marked by a pronounced shift towards object oriented methodologies we decided to plan the education of future IIS users in accordance with the basic ideas of object oriented approaches. We are of the opinion that it is an advantage for an effective cooperation between information technology professionals and users if the users possess at least a basic understanding of the concepts and models used by object oriented methodologies.

The use of objects made possible the utilization of their synergy in that the object model united all the stages to yield one design which could be used without modification up to and including the implementation stage. This was because in all stages the designer used one language - the language of objects and classes. Thus it becomes necessary to explain these fundamental concepts. This poses a non-trivial didactical problem: how detailed an explanation of the fundamental concepts 'object' and 'class' should be provided?

Object oriented methodologies offer a wide range of graphical modeling tools, many kinds of diagrams. It is necessary to choose the most suitable modeling tools for their creation. After careful

consideration of various alternative models which could be used in training future IS users we decided to start with a detailed presentation of the so-called Use Case Model. Again, what level of detail should be selected for training non-specialists in the development of Use Case models? Is it a good idea to teach certain parts of the UML language - and if so, how detailed a description is best?

Object oriented methods which we have briefly mentioned are used in many CASE tools of various types which are indispensable for IS development. Developing extensive information systems without these tools is unthinkable. The use of objects has become an essential approach in the development of modern distributed applications on the Internet. One of the most frequent magic words which emerged quite recently in information technology is CORBA - Common Object Request Broken Architectur, there are other technologies as eg DCOM - Distributed Component Object Model-backed by Microsoft. Is it necessary for future IIS users to have at least an idea of its architecture and infrastructure?

A detailed description of all known requirements which the new system should satisfy is the most important common activity of the future user and the developer at the start of a system development project. Experience gained in many projects shows that it is an advantage if the process of user requirement specification is controlled by experienced analysts. A well prepared document entitled Specification of User Requirements is of fundamental importance for system introduction as well as for resolving eventual misunderstandings between the supplier and contractor. User cooperation is a necessary condition for the preparation of such a document. How much information should the user receive concerning Requirements Engineering methods?

Giving a comprehensive reply to all these questions is undoubtedly a complicated task and their understanding will no doubt be affected in various ways by the evolution in the field of IIS development. However, it seems certain that some understanding of this problem domain is necessary for all future managers. Let us remark that this conclusion is borne out by our experience in developing IIS both for state administration and the university environment. A training method which we have found suitable is based on so-called 'virtual projects'. A student becomes a virtual top manager of a company or institution. Based on a fictional 'present state of the IS' he formulates the goals and specifications for his company's IIS. At a later stage he attempts, using the user requirement specification and Use Case model, to create specification for a selected subsystem of the IIS. In spite of the fact that most of our students have little knowledge of information technology some of the more than 300 virtual projects submitted so far are on a high level and are an encouragement for further and more detailed consideration of the questions raised in this note.

Please contact:
Peter Mihók – P. J. Safarik University/SRCIM
Tel: +421 95 622 1128
E-mail: mihok@kosice.upjs.sk

Searching Documentary Films On-line: the ECHO Project

by Pasquale Savino

Wide access to large information collections is of great potential importance in many aspects of everyday life. However, limitations in information and communication technologies have, so far, prevented the average person from taking much advantage of existing resources. Historical documentaries, held by national audiovisual archives, constitute some of the most precious and less accessible cultural information. The ECHO project intends to contribute to the improvement of the accessibility to this precious information, by developing a Digital Library (DL) service for historical films belonging to large national audiovisual archives.

The ECHO project, funded by the European Community within the V Framework Program, KA III, intends to provide a set of digital library services that will enable a user to search and access documentary film collections. For example, users will be able to investigate how different countries have documented a particular historical period of their life, or view an event which is documented in the country of origin and see how the same event has been documented in other countries, etc. One effect of the emerging digital library environment is that it frees users and collections from geographic constraints.

The Project is co-ordinated by IEI-CNR. It involves a number of European institutions (Istituto Luce, Italy; Institut Nationale de l'Audiovisuel, France; Netherlands Audiovisual Archive, and Memoriav, Switzerland) holding or managing unique collections of documentary films, dating from the beginning of the century until the seventies. Tecmath, EIT, and Mediasite are the industrial partners that will develop and implement the ECHO system. There are two main academic partners (IEI-CNR and Carnegie Mellon University - CMU) and four associate partners (CNRS-LIMSI, IRST, University of Twente, and University of Mannheim).

ECHO System Functionality

The emergence of the networked information system environment allows us to envision digital library systems that transcend the limits of individual collections to embrace collections and services that are independent of both location and format. In such an environment, it is important to support the interoperability of distributed, heterogeneous digital collections and services. Achieving interoperability among digital libraries is facilitated by

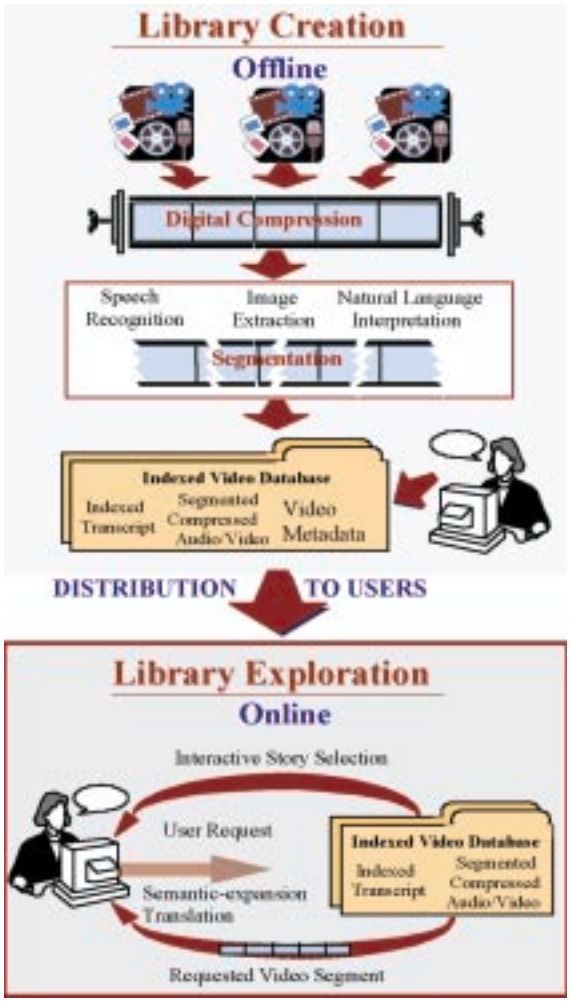
conformance to an open architecture as well as agreement on items such as formats, data types, and metadata conventions.

ECHO aims at developing a long term reusable software infrastructure and new metadata models for films in order to support the development of interoperable audiovisual digital libraries. Through the development of new models for film metadata, intelligent content-based searching and film-sequence retrieval, video abstracting tools, and appropriate user interfaces, the project intends to improve the accessibility, searchability, and usability of large distributed audiovisual collections. Through the implementation of multilingual services and cross language retrieval tools, the project intends to support users when accessing across linguistic, cultural and national boundaries. The ECHO system will be experimented, in the first place, for four national collections of documentary film archives (Dutch, French, Italian, Swiss). Other archives may be added in a later stage.

The ECHO System

The figure provides an overview of the main operations supported by the ECHO system. ECHO assists the population of the digital library through the use of mechanisms for the automatic extraction of content. Using a high-quality speech recogniser, the sound track of each video source is converted to a textual transcript, with varying word error rates. A language understanding system then analyses and organises the transcript and stores it in a full-text information retrieval system. Multiple speech recognition modules for different European languages will be included. Likewise, image understanding techniques are used for segmenting video sequences by automatically locating boundaries of shots, scenes, and conversations. Metadata is then associated with film documentaries in order to complete their classification.

Search and retrieval via desktop computers and wide area networks is performed by expressing queries on the audio transcript, on metadata, or by image similarity retrieval. Retrieved documentries, or their abstracts, are then presented to the user. By the collaborative



System overview.

interaction of image, speech and natural language understanding technology, the system compensates for problems of interpretation and search in the error-full and ambiguous data sets.

Expected Results

The project will follow an incremental approach to system development. Three prototypes will be developed offering an increasing number of functionalities. The starting point of the project will be a software infrastructure resulting from an integration of the Informedia and Media Archive(r) technologies.

The first prototype (Multiple Language Access to Digital Film Collections) will integrate the speech recognition engines with the infrastructure for content-based indexing. It will demonstrate an open system architecture for digital film libraries with automatically indexed film collections and intelligent access to them on a national language basis.

The second prototype (Multilingual Access to Digital Film Collections) will add a metadata editor which will be used to index the film collections according to a common metadata model. Index terms, extracted automatically during the indexing/segmentation of the film material (first prototype), will be integrated with local metadata, extracted manually, in a common description (defined by the common metadata model).The second prototype will support the interoperability of the four collections and content based searching and retrieval.

The third prototype (ECHO Digital Film Library) will add summarization, authentication, privacy and charging functionalities in order to provide the system with full capabilities.

Links:
http://www.iei.pi.cnr.it/echo/

Please contact:
Pasquale Savino – IEI-CNR
Tel: +39 050 315 2898
E-mail: P.Savino@iei.pi.cnr.it

IS4ALL: A New Working Group promoting Universal Design in Information Society Technologies

by Constantine Stephanidis

IS4ALL (Information Society for All) – is a new EC-funded project aiming to advance the principles and practice of Universal Access in Information Society Technologies, by establishing a wide, interdisciplinary and closely collaborating network of experts (Working Group) to provide the European IT&T industry in general, and Health Telematics in particular, with a comprehensive code of practice on how to appropriate the benefits of universal design.

The International Scientific Forum 'Towards an Information Society for All' was launched in 1997, as an international ad hoc group of experts sharing common visions and objectives, namely the advancement of the principles of Universal Access in the emerging Information Society. The Forum held three workshops to establish interdisciplinary discussion, a common vocabulary to facilitate exchange and dissemination of knowledge, and to promote international co-operation. The 1st workshop took place in San Francisco, USA, August 29, 1997, and was sponsored by IBM. The 2nd took place in Crete, Greece, June 15-16, 1998 and the 3rd took place in Munich, Germany, August 22-23, 1999. The latter two events were partially funded by the European Commission. The Forum has produced two White Papers, published in, International Journal of Human-Computer Interaction, Vol. 10(2), 107-134 and Vol. 11(1), 1-28, while a third one is in preparation. You may also visit [http://www.ics.forth.gr/proj/at-hci/files/white_paper_1998.pdf] and [http://www.ics.forth.gr/proj/at-hci/files/white_paper_1999.pdf]. The White Papers report on an evolving international R&D agenda focusing on the development of an Information Society acceptable to all citizens, based on the principles of universal design. The proposed agenda addresses technological and user-oriented issues, application domains, and support measures. The Forum has also elaborated on the proposed agenda by identifying challenges in the field of human-computer interaction, and clusters of concrete recommendations for international collaborative R&D activities. Moreover, the Forum has addressed the concept of accessibility beyond the traditional fields of inquiry (eg, assistive technologies, built environment, etc), in the context of selected mainstream Information Society Technologies, and important application domains with significant impact on society as a whole. Based on the success of its initial activities, the Forum has proposed to the European Commission the establishment, on a formal basis, of a wider, interdisciplinary and closely collaborating network of experts (Working Group), which has been now approved for funding.

Universal Design

Universal Design postulates the design of products or services that are accessible, usable and, therefore, acceptable by potentially everyone, everywhere and at any time. Although the results of early work dedicated to promoting Universal Access to the Information Society (for a review see: http://www.ics.forth.gr/proj/at-hci/files/TDJ_paper.PDF) are slowly finding their way into industrial practices (eg, certain mobile telephones, point-of-sale terminals, public kiosks, user interface development toolkits), a common platform for researchers and practitioners in Europe to collaborate and arrive at applicable solutions is still missing. Collaborative efforts are therefore needed to collect, consolidate and validate the distributed wisdom at the European as well as the international level, and apply it in application areas of critical importance, such as Health Telematics, catering for the population at large, and involving a variety of diverse target user groups (eg, doctors, nurses, administrators, patients). Emerging interaction platforms, such as advanced desktop-oriented environments (eg, advanced GUIs, 3D graphical toolkits, visualisers), and mobile platforms (eg, palmtop devices), enabling ubiquitous access to electronic data from anywhere, and at anytime, are expected to bring about radical improvements in the type and range of Health Telematics services. Accounting for the accessibility, usability and acceptability of these technologies at an early stage of their development life cycle is likely to improve their market impact as well as the actual usefulness of the end products.

IS4ALL

The IS4ALL Working Group aims to provide European industry with appropriate instruments to approach, internalise and exploit the benefits of universal access, with particular emphasis on Health Telematics. Toward this objective, IS4ALL will develop a comprehensive code of practice (eg, enumeration of methods, process guidelines) consolidating existing knowledge on Universal Access in the context of Information Society Technologies, as well as concrete recommendations for emerging technologies (eg, emerging desktop and mobile platforms), with particular emphasis on their deployment in Health Telematics. IS4ALL will also undertake a mix of outreach activities to promote Universal Access principles and practice, including workshops and seminars targeted to mainstream IT&T industry.

IS4ALL is a multidisciplinary Working Group co-ordinated by ICS-FORTH. The membership includes: Microsoft Healthcare Users Group Europe, the European Health Telematics Association, CNR-IROE, GMD, INRIA and Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. - Institut für Arbeitswirtschaft und Organisation, Germany.

Please contact:
Constantine Stephanidis – ICS-FORTH
Tel: +30 81 39 17 41
E-mail: cs@ics.forth.gr

Identifying Vehicles on the Move

by Beate Koch

Research carried out in Germany in the field of information and communications technology is extremely diverse in nature. A major role is played by the Fraunhofer-Gesellschaft with its total of 47 Institutes and branches in the USA and Asia. The following example shows how the preliminary research conducted by the Fraunhofer Institutes - with a constant focus on the longer term - enables scientific findings to be converted into pioneering and marketable products.

Automated image-processing systems constitute a booming market with double-digit growth rates. The breakthrough was achieved in the combination of camera and information technology - thereby implementing mechanized sight, far swifter and more reliable in operation than the human eye.

For many years, Fraunhofer researchers in Berlin have played a leading role in the development of intelligent systems for the identification of signs and symbols. Their latest breakthrough is a compact system capable of recognizing registration plates and loading signs on vehicles in flowing traffic up to a speed of 120 km/h.

Bertram Nickolay of the Fraunhofer Institute and his team have constructed an entire family of modules that allows various components of road traffic to be recorded and automatically analyzed. The individual components can be put together in any combination. "Our ISY family solves many tasks in traffic engineering," reports project leader Nickolay, proudly. ISYCOUNT allows vehicles to be counted in any traffic situation. The data are then used as the basis for telematic systems that control traffic flow and prevent holdups. ISYTRACK combined with ISYPLATE can be used to register vehicles entering and leaving waste disposal sites or gravel pits. Unlike private cars, commercial trucks carry their identification plates in different places, and they have to be located amidst a number of other signs, such as the company logo and signs indicating the laden weight and the nature of the goods being transported. ISYSHAPE registers the shape and dimension of a vehicle and is then able to identify its type from specific features of its design, and make a rough classification of its loading.

Photo: Voss Fotografie

Camera and processor are merged in a single device. This allows registration plates to be detected and identified in flowing traffic or at the entrances and exits to car parks and factory sites. License plates are made clearly visible by the optical trigger.

The latest addition to the family is a software program that enhances the image recorded by the automatic recognition system: ISYFLASH is capable of determining the ideal moment at which to capture an image, even if the vehicle is traveling at high speed. This ensures the highest possible resolution and the maximum degree of certainty in identifying the registration plate. The family of modules is based on the latest methods of intelligent character recognition (ICR) combined with learning processes on the basis of soft computing. They are capable of solving complex problems, even in the presence of out-of-focus or doubtful elements such as mud on the number plate, deformations or unusual colors. Most of the currently available re-cognition systems are only able to identify license plates common in one or a few selected countries. They have problems if the type and color of the background changes, or if a new character

font is introduced. The system developed by the IPK is already capable of recognizing signs on vehicles registered in 25 different countries; it does not matter if the vehicle is registered in Germany, the UK, Poland, Israel, South Africa or even Russia. That is because the system learns fast: Given a few samples, such as actual number plates or video pictures of a new character font, the system trains itself and by the next day it is capable, for example, of reading signs written in Cyrillic script. The IPK is marketing its software through a number of international cooperating partners.

A further advantage of the system is that the camera and the electronic analysis circuits are incorporated in a small compact unit to make the equipment more portable and economical. The camera and the processor are merged into a single 'smart device'. This simplifies the structure, reducing the number of interfaces and hence the number of potential sources of errors. This portable apparatus possesses several features that give it the edge over other recognition systems used to identify and record moving vehicles: Laser scanners are more expensive to maintain, light barriers are more susceptible to faults. Work-intensive earth-moving operations are required in order to lay induction loops in the road surface. And none of these systems are infallible - they can just as easily be triggered by a person or an animal passing through their field of recogni-tion. ISYFLASH uses image-analysis methods to exclude such disturbances. The portable units are ideal for monitoring and regulating the access ramps and exits of high-rise car parks and industrial sites, quickly and smoothly. Police forces in a number of countries have also indicated considerable interest in the system developed by the IPK, which can be used to help track down criminals on the road. Installed in a police vehicle, the mobile ISYFLASH equipment can identify the registration numbers of other road users while on the move or from a parked position alongside a highway or at one of its exits. If a car bearing the suspect number drives past, an alarm lamp starts to flash. Other, innocent drivers are then spared the inconvenience of having to stop at police road blocks during a large-scale police operation.

Videomaut is not exactly a new idea, but it is still not ready for widespread implementation: Drivers slowly steer their car through the video-control barrier, in the hope that the camera will register the number plate correctly as they move past at walking speed. If not, the barrier remains closed and the driver has to back out and join the line waiting at the regular toll booth. The intelligent recognition system developed by the researchers in Berlin is different: Barriers and toll booths are obsolete, and the driver doesn't even need to brake to a slower speed. A number of countries are considering automating their highway toll systems using the software from the IPK. The advantage: No more tailbacks at the toll stations, and useful telemetry data for traffic regulation.

Please contact:
Bertram Nickolay – Fraunhofer Gesellschaft
Tel: +49 30 3 90 06 2 01
E-mail: nickolay@ipk.fhg.de

Virtual Planetarium at Exhibition 'ZeitReise'

by Igor Nikitin and Stanislav Klimenko

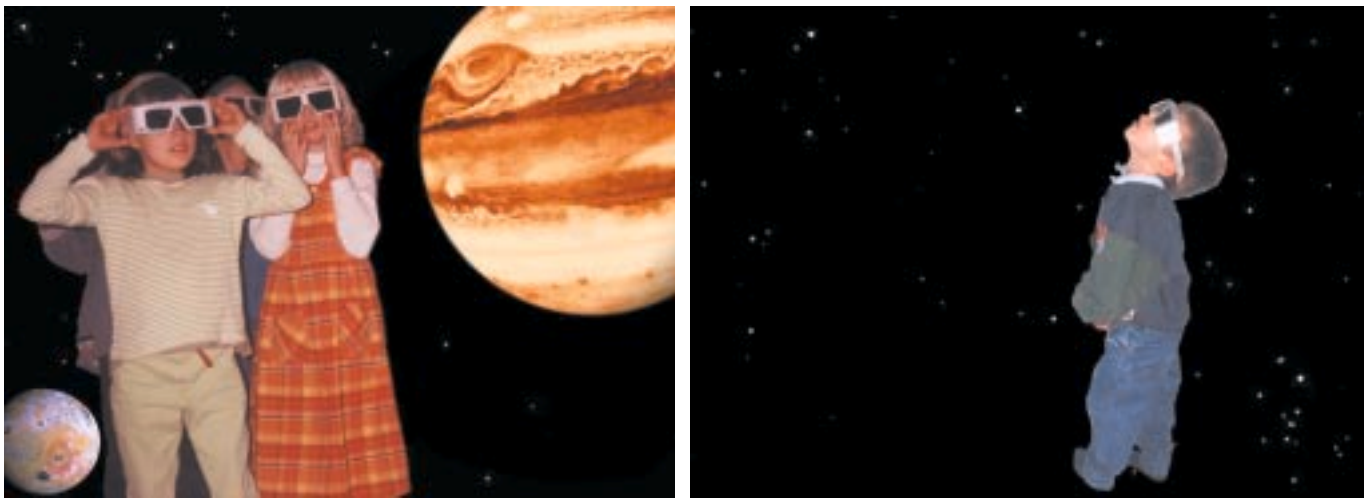
At the Exhibition ZeitReise/TimeJourney from 12 May to 25 June 2000 at the Academy of Arts in Berlin, GMD's Virtual Environments Group presented a 3D-installation 'Virtual Planetarium'.

The Virtual Planetarium is an educational application that uses special methods to display the astronomical objects realistically, as they are visible by astronauts of a real spacecraft, preserving correct visible sizes of all objects for any viewpoint and using 3D models based on real astronomical data and images. The installation includes the 3200 brightest stars, 30 objects in the Solar System, an interactive map of constellations, composed of ancient drawings, a large database, describing astronomical objects textually and vocally in English and German. A stereoscopic projection system is used to create an illusion of open cosmic space. The application is destined primarily for CAVE-like virtual environment systems, giving a perception of complete immersion into the scene, and also works in simple installations using a single wall projection.

Avango Application Development Toolkit

Avango is a programming framework for building distributed, interactive virtual environment applications. It is based on a SGI Performer to achieve the maximum possible performance for an application and addresses the special needs involved in application development for virtual environments. Avango uses the C++ programming language to define the objects and scripting language Scheme to assemble them in a scene graph. Avango also introduces a dataflow graph, conceptually orthogonal to the scene graph, which is used to define interaction between the objects and to import the data from external devices into the application.

A non-linear geometrical model was used to represent the objects in the Solar System, overcoming a problem of 'astronomical scales'. This problem consists in the fact that a size of the Solar System (diameter of Pluto orbit) and a sizes of small planets (Phobos, Deimos) differ by factor more than 109, as a result single precision 4 bytes real numbers, used in standard graphical hardware, are insufficient to represent the coordinates of objects in such a scene. A special non-linear transformation should intermediate the real scale astronomical model and its virtual analog, mapping astronomical double precision sizes ranging from 1 to 1010 km into virtual environment single



Scolar excursion.

precision sizes ranging from 1 to 20 m. Such transformation was chosen, satisfying two requirements: (1) it preserves actual angular sizes of planets for any viewpoint to achieve realism of presentation, (2) it prohibits penetration inside the planets to implement no-collision algorithm. Analogous transformations are also applied to the velocity of the observer, to make the exploration of near-Earth space and distant planets possible in one demo session.

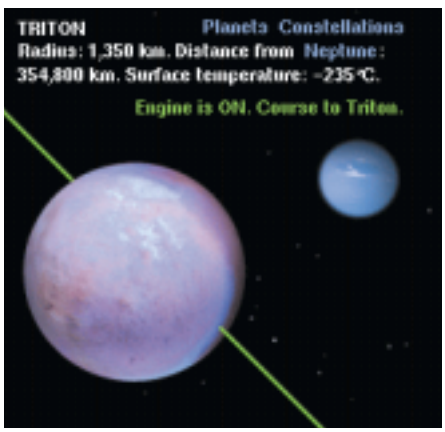
Navigation in CAVE is performed using electromagnetic 3D pointing device (stylus), or joystick/mouse in more simple installations. User can manipulate by a green ray in a virtual model, choosing the direction of motion and objects of interest. A navigation panel, emulating HTML browser, is used to display the information about selected objects and to choose the route of journey.

Sound accompaniment includes two musical themes, representing the rest state and fast motion, which are mixed dependently on the velocity (courtesy to Martin Gerke and Thomas Vogel), and voices, describing surrounding objects (courtesy to Christian Brückner and Ulrike Tzschöckell). Another variant of vocal representation is using the IBM ViaVoice text-to-speech conversion tool. The sound scheme can be selected and reconfigured during runtime.

Sources of planetary images and stellar data are from public Internet archives and databases available at the NASA: http://photojournal.jpl.nasa.gov, http://adc.gsfc.nasa.gov and US Geological Survey: http://wwwflag.wr.usgs.gov/. Images were imposed on a spherical geometry as textures and were enlightened by a composition of bright solar and dim ambient lights. Positions, intensities and colors of stars in the model correspond to astronomical data from catalogue. Additionally, high resolution images of constellations can be displayed in the sky. Source of images is an ancient stellar map: (Hemisphaerium coeli boreale/ Hemisphaerium coeli australe: in quo fixarum loca secundum eclipticae ductum ad an[n]um 1730 completum exhibentur/ a Ioh. Gabriele Doppelmaiero math. prof. publ. Acad. imper. leopoldino-carolinae naturae curiosorum et Acad. scient. regiae prussicae socio ; opera Ioh. Baptistae Homanni sac. caes. maj. geogra. – Norimbergae [Nürnberg]: [Homännische Erben], [erschienen 1742].), courtesy of Kaiserslautern University library.

Links:
http://heimat.de/zeitreise
http://viswiz.gmd.de/~nikitin/stars

Please contact:
Igor Nikitin – GMD
Tel: +49 2241 14 2137
E-mail: Igor.Nikitin@gmd.de



Navigation panel.

Images of constellations.

SPONSORED BY ERCIM

6th Eurographics Workshop on Virtual Environments

by Robert van Liere

From practical calibration procedures for Augmented Reality and effects on group collaboration on presence in a collaborative virtual environment to a virtual planetarium. The 6th Eurographics Workshop on Virtual Environments covered a broad range of topics, reflecting the expanding scope virtual reality. This workshop organized by the Eurographics Association, was held June 1-2 at the CWI in Amsterdam. ERCIM acted as a co-sponsor to the workshop.

The workshop was attended by 56 participants, which was much more than initially expected. In addition, the organizers were very pleased that the participants came from all over the world. This reflects the growing interest in virtual environments as a field of research.

The keynote speaker – Michael Zyda from the Navel Postgraduate School in Monterey, California – gave a very interesting talk on the Modeling, Virtual Environments and Simulation Program. The talk focussed on research directions in the field of entertainment. Zyda gave his views on technologies for immersion (low-cost 3D image generation, spatial tracking, game platform utilization, multimodal sensory presentation), networked simulation (high bandwidth networks, dynamically extensible network software architectures, area of interest management, latency reduction) and computer-generated autonomy (agent-based simulation, adaptability, learning, human behavior representations). Zyda gave examples of how entertainment in virtual environments could take advantage of these technologies and which bottlenecks still have to be overcome.

The workshop ended at the Dutch supercomputer center, SARA, in which a transatlantic collaborative virtual walk-through demonstration was given. Two CAVEs, one at the SARA in Amsterdam and one at the Electronic Visualization Laboratory in Illinois, were linked with a 155 Mbit ATM link. In one demonstration, participants could collaboratively walk through a virtual building which was designed by the Dutch architect Remco Koolhaas.

The atmosphere at workshop was very relaxed and enjoyable; the participants particularly enjoyed the social event at the end of the first day: a traditional Indonesian Rijsttafel in the heart of Amsterdam. (In fact, the social event clearly effected the number of participants at the first session of the second day!)

Links:
Conference website:
http://www.cwi.nl/egve00/

Please contact:
Robert van Liere – CWI
Tel: +31 20 592 4118
E-mail: robertl@cwi.nl

Trinity College Dublin hosted the Sixth European Conference on Computer Vision – ECCV'2000

by David Vernon

Ten years ago, the inaugural European Conference on Computer Vision was held in Antibes, France. Since then, ECCV has been held biennially under the auspices of the European Vision Society at venues around Europe. This year, the privilege of organizing ECCV 2000 fell to Ireland and, from 26 June to 1 July, Trinity College Dublin hosted what has become one of the most important events in the calendar of the Computer Vision community.

The Trinity campus, set in the heart of Dublin, is an oasis of tranquility and its beautiful squares, elegant buildings, and tree-lined playing-fields provided the perfect setting for the conference. Add to this the superb facilities in the Trinity Conference Centre, six days of uninterrupted sunshine, 280 delegates, and the stage was set for a memorable week.

The first day of the week was devoted to a series of tutorials on focussed topics which were given by international experts in that field. These afforded many delegates an ideal opportunity of getting a snap-shot view of the state of the art in a perhaps unfamilar subject.

The next four days comprised the conference proper and was opened by the Vice-Provost of Trinity College, Professor David McConnell, who set the tone for the coming days with his well-chosen words on the nature and relevance of perception and the need for computational models of vision to enable both industrial applications and reseach in other scientific disciplines. Because ECCV is a single-track conference, it severely limits the number of papers that can be accepted and ensures that they are of the highest quality. During the week,

forty-four papers were presented at the podium and seventy-two were presented during daily poster sessions. The proceedings were published by Springer-Verlag as two volumes, each with approximately 900 pages (LNCS Vols. 1842 & 1843).

The final day was devoted to four Workshops on topics ranging from visual surveillance, through 3-dimensional reconstruction of large objects such as buildings, to empirical evaluation of computer vision algorithms.

Whilst the technical excellence of the scientific programme is undoubtedly the most important aspect of ECCV, there are other facets to an enjoyable and productive conference, facets which should engender conviviality, discourse, and interaction - in other words, the social programme! This was extensive and varied and it featured the renowned Trinity Gala Evening with exhibitions of excellent Irish cuisine, music, and dancing; and a reception in the Long Room (Trinity's original library and perhaps one of the most beautiful rooms in the College) where delegates were regaled by the Librarian on the history and future plans of the Trinity College Dublin Library. They also had an opportunity to see the Book of Kells. The conference dinner was set in the splendour of the Royal Hospital Kilmainham and some delegates also enjoyed a trip to Johnny Fox's pub - apparently the highest pub in Ireland – for another evening of Irish music and dance.

The goal in organizing ECCV 2000 at Trinity was to ensure that delegates would depart with great memories, many new friends, and inspirational ideas for future research; that they also now think the sun always shines in Ireland is an additional bonus!

Links:
Conference Website:
http://www.eccv2000.tcd.ie/

Please contact:
David Vernon – Conference Chair, ECCV'2000
National University of Ireland, Maynooth
Tel: +971 6 535 5355
E-mail: dvernon@ece.ac.ae

SPONSORED BY ERCIM

Fifth Workshop of the ERCIM Working Group on Constraints

by Eric Monfroy

The Fifth Workshop of the ERCIM Working Group on Constraints, held in Padova, Italy, 19-21 June 2000, gathered some 35 experts from Europe and the United States to discuss issues related to applications of constraints. The workshop was organized by Krzysztof R. Apt (CWI, Amsterdam), Eric Monfroy (CWI, Amsterdam), and Francesca Rossi (University of Padova).

The ERCIM Working Group on Constraints (coordinated by Krzysztof R. Apt) was founded in the fall of 1996. Currently, it comprises 16 ERCIM institutes and associated organizations. This group brings together ERCIM researchers that are involved in research on the subject of Constraints.

The Padova workshop was the fifth meeting of the group. It took place at the Department of Pure and Applied Mathematics of the University of Padova, Italy on 19-21 June 2000. It was the fourth workshop of the group jointly organized with the CompulogNet (Esprit Network of Excellence on Computational Logic) area on 'Constraint Programming', coordinated by Francesca Rossi.

This workshop covered all aspects of constraints (solving and propagation algorithms, programming languages, new formalisms, novel techniques for modeling, experiments, etc.), with particular emphasis on applications. The authors of theoretical papers were encouraged to investigate and discuss a possible use of their work in practice.

The workshop attracted some 35 researchers and ran for 3 days. The call for presentations attracted a number of submissions out of which, given the time limitations, we could accept only 20. Besides the accepted papers, we had three invited talks on various application areas for constraints:

- Thom Fruehwirth from the Ludwig-Maximilians-University of Munich gave a presentation on 'Applications of Constraint Handling Rules',
- Alan Borning from the University of Washington discussed 'Applying Constraints to Interactive Graphical Applications'
- Michel Scholl from INRIA gave a presentation on 'Constraint databases for spatio-temporal applications'.

The workshop was considered to be very successful and led to a number of interesting lively discussions on the topics raised during the presentations. The workshop provided a useful platform that allowed all of us, practitioners as well as researchers involved in theoretical work, to exchange information, learn about each others work, and discuss possible future cooperation.

The detailed program of the workshop, the proceedings in the form of extended abstracts of the presentations, and more information about the Working Group are electronically available through the Web site of the ERCIM Working Group on Constraints.

Links:
ERCIM Working Group on Constraints:
http://www.cwi.nl/projects/ercim-wg.html.

Please contact:
Eric Monfroy – CWI
Tel: +31 20 592 4195
E-mail: Eric.Monfroy@cwi.nl

CALL FOR PAPERS

International Journal Universal Access in the Information Society

1st Issue: February 2001

Universal Access in the Information Society (UAIS) is an international, interdisciplinary refereed journal that solicits original research contributions addressing the accessibility, usability, and, ultimately, acceptability of Information Society Technologies by anyone, anywhere, at anytime, and through any media and device. The journal publishes research work on the design, development, evaluation, use, and impact of Information Society Technologies, as well as on standardization, policy, and other non-technological issues that facilitate and promote universal access. Paper submissions, in English, should report on theories, methods, tools, empirical results, reviews, case studies, and best practice examples, in any application domain and should have a clear focus on universal access. In addition to the above, the journal will host special issues, book reviews and letters to the editor, news from Information Society Technologies industry, and standardization and regulatory bodies, announcements (eg, conferences, seminars, presentations, exhibitions, education & curricula, awards, new research programs) and commentaries (eg, about new legislation).

The journal will be published by Springer.

Further information:
Journal's website:
http://link.springer.de/journals/uais/
Editor-in-Chief: Constantine Stephanidis – FORTH-ICS
E-mail: cs@ics.forth.gr

CALL FOR PARTICIPATION

8th International Conference on Database Theory

London, 4-6 January 2001

ICDT is a biennial international conference on theoretical aspects of databases and a forum for communicating research advances on the principles of database systems. Initiated in Rome, in 1986, it was merged in 1992 with the MFDBS symposium series initiated in Dresden in 1987. ICDT aims to attract papers of high quality, describing original ideas and new results on theoretical aspects of all forms of database systems and database technology. While special emphasis is put on new ideas and directions, papers on all aspects of database theory and related areas are presented.

ICDT 2001 will be preceeded by a new International Workshop on Web Dynamics which ICDT participants are most welcome to attend.

Further information:
http://www.dcs.bbk.ac.uk/icdt2001/

CALL FOR PARTICIPATION

37th Dutch Mathematical Congress

Amsterdam, 19-20 April 2001

CWI and the Vrije Universiteit in Amsterdam will jointly organise the 37th Dutch Mathematical Congress. This congress, under the auspices of the Dutch Mathematical Society, is the yearly meeting point in The Netherlands of the Dutch mathematicians. Invited talks will be given on cryptography, statistics, analysis and discrete math/combinatorics. Minisymposia will be organised on the same four subjects, supplied with applied math, systems theory, numerical analysis, computer algebra, financial math, and optimization of industrial processes. Location: Vrije Universiteit in Amsterdam.

Further information:
http://www.cwi.nl/conferences/NMC2001

CALL FOR PARTICIPATION

Optimal Control and Partial Differential Equations – Innovations and Applications: Conference in Honor of Professor Alain Bensoussan's 60th Anniversary

Paris, 4-5 December 2000

The conference will be divided into two parts: the first day dedicated to fundamental research animated by international wellknown scientists will present the state of the art in the fields where Alain Bensoussan's contributions have been particularly important: filtering and control of stochastic systems, variational problems, applications to economy and finance, numerical analysis, etc. The second day will devoted to the tight links between research and applications and the impact of research on the 'real world' especially through valorization of research results. This theme will be enlighted through three round tables on the following topics: space, spatial applications, science and technology for information and communication. A fourth panel will be devoted to Europe as a priviledged example of the collaboration between research and industry.

Alain Bensoussan was the first president of ERCIM from 1989 to 1994.

Further information:
http://www.inria.fr/AB60-eng.html

CALL FOR PARTICIPATION

VisSym '01 – Joint Eurographics - IEEE TCVG Symposium on Visualization

Ascona, Switzerland, 28-30 May 2001

Papers and case studies will form the scientific content of the event. Both research papers and case studies are invited that present research results from all areas of visualization. Case studies report on practical applications of visualization to data analysis.

Topics for research papers include, but are not limited to: flow visualization, volume rendering, surface extraction, information visualization, internet-based visualization, data base visualization, human factors in visualization, user interaction techniques, visualization systems, large data sets, multi-variate visualization, multi resolution techniques.

Further information:
http://www.cscs.ch/vissym01/cfp.html

CALL FOR PAPERS

IEA/AIE-2001 – The Fourteenth International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems

Budapest, Hungary, 4-7 June 2001

IEA/AIE continues the tradition of emphasizing applications of artificial intelligence and expert/knowledge-based system to engineering and industrial problems. Authors are invited to submit papers presenting the results of original research or innovative practical applications relevant to the conference. Practical experiences with state-of-the-art AI methodologies are also acceptable when they reflect lessons of unique value to the conference attendees.

Further information:
http://www.sztaki.hu/conferences/ieaaie2001/

CALL FOR PAPERS

International Conference on Shape Modelling and Applications

Genova, Italy, 7-12 May 2001

Reasoning about shape is a common way of describing and representing objects in engineering, architecture, medicine, biology, physics and in daily life. Modelling shapes is part of both cognitive and creative processes and from the outset models of physical shapes have satisfied the desire to see the result of a project in advance.

The programme will include one day of tutorial and three days of papers programme. Special Sessions will be organized on relevant topics.

The Conference is organized by the Istituto per la Matematica Applicata of the CNR.

Further information:
http://SMI2001.ima.ge.cnr.it/

SPONSORED BY ERCIM

JCAR 2001 International Conference on Shape – The International Joint Conference on Automated Reasoning

Siena, 18-23 June 2001

IJCAR is the fusion of three major conferences in Automated Reasoning: CADE – The International Conference on Automated Deduction, TABLEAUX – The International Conference on Automated Reasoning with Analytic Tableaux and Related Methods and FTP – The International Workshop on First-Order Theorem Proving). These three events will join for the first time at the IJCAR conference in Siena in June 2001. IJCAR 2001 invites submissions related to all aspects of automated reasoning, including foundations, implementations, and applications. Original research papers and descriptions of working automated deduction systems are solicited.

Further information:
http://www.dii.unisi.it/~ijcar/

Order Form

If you wish to subscribe to ERCIM News 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 or fax this form to:

ERCIM NEWS
Domaine de Voluceau
Rocquencourt
BP 105
F-78153 Le Chesnay Cedex
Fax: +33 1 3963 5052
E-mail: office@ercim.org

Data from this form will be held on a computer database. By giving your email address, you allow ERCIM to send you email

Name: .....

Organisation/Company: .....

Address: .....

.....

Post Code: .....

City: .....

Country .....

E-mail: .....

You can also subscribe to ERCIM News and order back copies by filling out the form at the ERCIM web site at ***http://www.ercim.org/publication/Ercim_News/***

ERCIM NEWS

ERCIM News is the in-house magazine of ERCIM. Published quarterly, the newsletter reports on joint actions of the ercim partners, and aims to reflect the contribution made by ercim to the European Community in Information Technology. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. ERCIM News has a circulation of 7000 copies.

ERCIM News online edition is available at http://www.ercim.org/publication/ERCIM_News/

ERCIM News is published by ERCIM EEIG, BP 93, F-06902 Sophia-Antipolis Cedex
Tel: +33 4 9238 5010, E-mail: office@ercim.org
ISSN 0926-4981
Director: Jean-Eric Pin
Central Editor:
Peter Kunz
E-mail: peter.kunz@ercim.org
Tel: +33 1 3963 5040
Local Editors:
CLRC: Martin Prime
E-mail: M.J.Prime@rl.ac.uk
Tel: +44 1235 44 6555
CRCIM: Michal Haindl
E-mail: haindl@utia.cas.cz
Tel: +420 2 6605 2350
CWI: Henk Nieland
E-mail: henkn@cwi.nl
Tel: +31 20 592 4092
CNR: Carol Peters
E-mail: carol@iei.pi.cnr.it
Tel: +39 050 315 2897
FORTH: Constantine Stephanidis
E-mail: cs@csi.forth.gr
Tel: +30 81 39 17 41
GMD: Dietrich Stobik
E-mail: stobik@gmd.de
Tel: +49 2241 14 2509
INRIA: Bernard Hidoine
E-mail: bernard.hidoine@inria.fr
Tel: +33 1 3963 5484
SICS: Kersti Hedman
E-mail: kersti@sics.se
Tel: +46 8633 1508
SINTEF: Truls Gjestland
E-mail:
truls.gjestland@informatics.sintef.no
Tel: +47 73 59 26 45
SRCIM: Gabriela Andrejkova
E-mail: andrejk@kosice.upjs.sk
Tel: +421 95 622 1128
SZTAKI: Erzsébet Csuhaj-Varjú
E-mail: csuhaj@sztaki.hu
Tel: +36 1 209 6990
TCD: Carol O'Sullivan
E-mail: Carol.OSullivan@cs.tcd.ie
Tel: +353 1 60 812 20
VTT: Pia-Maria Linden-Linna
E-mail: pia-maria.linden-linna@vtt.fi
Tel: +358 0 456 4501

Free subscription
You can subscribe to ERCIM News free of charge by:
- sending e-mail to your local editor
- posting paper mail to the ERCIM office (see address above)
- filling out the form at the ERCIM website at http://www.ercim.org/

CALL FOR PAPERS

UAHCI 2001 – 1st International Conference on Universal Access in Human-Computer Interaction

New Orleans, LA, USA, 5-10 August 2001

The 1st International Conference on Universal Access in Human-Computer Interaction (UAHCI 2001), held in cooperation with HCI International 2001, aims to establish an international forum for the exchange and dissemination of scientific information on theoretical, methodological and empirical research that addresses all issues related to the attainment of universal access in the development of interactive software. The conference aims to attract participants from a broad range of disciplines and fields of expertise, including HCI specialists, user interface designers, computer scientists, software engineers, ergonomists and usability engineers, Human Factors experts and practitioners, organizational psychologists, system / product designers, sociologists, policy and decision makers, scientists in government, industry and education, as well as assistive technology and rehabilitation experts. The conference solicits papers reporting results of research work on, or offering insights on open research issues and questions in, the design, development, evaluation, use, and impact of user interfaces, as well as standardization, policy and other non-technological issues that facilitate and promote universal access.

Deadline for paper abstract submission: 5 November 2000

Contact and submission details:
http://uahci.ics.forth.gr/

CALL FOR PAPERS

ICANNGA 2001 – 5th International Conference on Artificial Neural Networks and Genetic Algorithms

Prague, Czech Republic, 22-25 April 2001

The focus of ICANNGA is on theoretical aspects and practical applications of computational paradigms inspired by natural processes, especially artificial neural networks and evolutionary algorithms.

ICANNGA 2001 will include invited plenary talks, contributed papers, poster session, tutorials and a social program. The conference is organized by the Institute of Computer Science, Academy of Sciences of the Czech Republic.

Further information:
http://www.cs.cas.cz/icannga

Forthcoming Events sponsored by ERCIM

ERCIM sponsors up to eleven conferences, workshops and summer schools per year. The funding is in the order of 2000 Euro. Conditions and an online application form are available at:
http://www.ercim.org/activity/sponsored.html

FORTE/PSTV 2000 – Formal Description Techniques for Distributed Systems and Communication Protocols,
Pisa, Italy, 10-13 October 2000
http://forte-pstv-2000.cpr.it/

Sofsem 2000 – Current Trends in Theory and Practice of Informatics,
Prague, 25 November-
2 December 2000
http://www.sofsem.cz/

IJCAR – International Joint Conference on Automated Reasoning,
Siena, Italy, 18-23 June 2001
http://www.sofsem.cz/

MFCS'2001 – 26th Symposium on Mathematical Foundations in Computer Science,
Marianske Lazne, Czech Republic,
27-31 August 2001
http://math.cas.cz/~mfcs2001/

CWI Incubator BV helps researchers starting a company: In July 2000, CWI has founded CWI Incubator BV (CWI Inc). CWI Inc will generate high-tech spin-off companies, based on results of CWI research. Earnings from CWI Inc. will be re-invested in fundamental research at CWI. During the last 10 years, CWI created 10 spin-off companies with approximately 500 employees. Generation of spin-off companies is an important method for institutes like CWI to convert fundamental knowledge to applications in society and to create high-level employment at the same time. CWI expects CWI Inc to attract enterprising researchers and to have a positive influence on the image of mathematics and computers science for students. For more information, see: http://www.cwi.nl/cwi/about/spin-offs.

CWI – Data Distilleries Raises $24million. One of CWI's succesful spin off companies is Dutch CRM developer Data Distilleries BV. In July 2000, Data Distilleries has closed a fundinground to raise $24m. The Amsterdam-based company will use the cash to expand into the UK and six other European countries and says it plans to be in the US market by the end of next year. Data Distilleries has developed technology that enables users with customer information from multiple channels to analyze it and draw up corresponding marketing and promotions customized for individual customers.

INRIA's 2000-2003 four-year contract signed: Roger-Gérard Schwartzenberg, Minister of Research, and Christian Pierret, Minister of Industry, signed INRIA's 2000-2003 four-year contract with Bernard Larrouturou, Chairman of INRIA. The four-year contract provides for a significant increase in INRIA personnel, which will go from 755 to 1180 as of 2003, as well as in funding to keep pace with the augmentation of personnel. As early as 2001, the Institute's funding will be raised by 9.15 million Euros and the staff will be increased by 180 employees. France will make a significant effort for research in information and communication technology in the coming years in terms of both means and staff to further and consolidate the role of France and in the larger context, Europe, in terms of innovation and economic competitiveness. Along with other institutions and university departments, INRIA is at the heart of the French state research system upon which this policy is based. Through this contract, the State, aiming for ambitious goals, has conferred on INRIA the labour and material resources for a well-planned growth with precise strategic objectives.

Photo: Alexandre Eidelman/INRIA

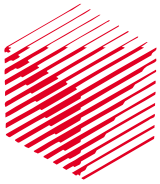
From left: Christian Pierret, French Minister of Industry, Roger-Gérard Schwartzenberg, French Minister of Research, Bernard Larrouturou, Chairman of INRIA.

The department of Computer Science in Trinity College Dublin hosted the first Irish Workshop on Eye-Tracking in May this year, in co-operation with the psychology department of University College Dublin. The analysis of eye movements can provide valuable insights which can be exploited in a range of research areas, and an interchange of ideas between the different fields is highly desirable. This workshop brought together an interdisciplinary group of researchers to share experiences in using eye-tracking hardware, and analysing eye movements. Topics included: Computer interfaces for the disabled, eye-tracking for video encoding, map-processing and other cognitive tasks, the treatment of schizophrenia, and applications in computer graphics. Link: http://isg.cs.tcd.ie/iwet/

VTT Information Technology sells its VIP Intelligent Pagination system to Unisys Corp.: VTT Information Technology has sold the VIP Intelligent Pagination system developed by it to Unisys Corporation. The VIP software has been developed to automate the layout of classified advertisements in newspapers. VTT Information Technology has already supplied the software to a number of European newspapers, including the Helsingin Sanomat in Finland and Dagens Nyheter in Sweden. Unisys will continue to develop the software in Finland. The product complements the US company's own, globally-marketed advertising systems which, until now, lacked a pagination function.

CWI upgraded its connection to SURFnet to gigabit level on 8 August. SURFnet is the national Dutch computer network for research and education organisations. CWI is the first client to get this very fast access to the SURFnet backbone, which will have a capacity of 80 Gbit/s next year.

Patrick Valduriez, Research Director at INRIA from the Caravel research team and his co-authors C. Bobineau, L. Bouganim and P. Pucheral from the PRISM Laboratory of Versailles University have been distinguished by the "Best Paper Award' at the VLDB (Very Large Database) Conference for their paper 'PicoDBMS: Scaling down Database Techniques for the SmartCard'. VLDB, which is one of the most prominent conferences in the database field took place in Cairo in September 2000.



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.

Central Laboratory of the Research Councils
Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire OX11 0QX, United Kingdom
Tel: +44 1235 82 1900, Fax: +44 1235 44 5385
http://www.cclrc.ac.uk/

Consiglio Nazionale delle Ricerche, IEI-CNR
Via Alfieri, 1, I-56010 Pisa, Italy
Tel: +39 050 315 2878, Fax: +39 050 315 2810
http://www.iei.pi.cnr.it/

Czech Research Consortium for Informatics and Mathematics
FI MU, Botanicka 68a, CZ-602 00 Brno, Czech Republic
Tel: +420 2 688 4669, Fax: +420 2 688 4903
http://www.utia.cas.cz/CRCIM/home.html

Centrum voor Wiskunde en Informatica
Kruislaan 413, NL-1098 SJ Amsterdam, The Netherlands
Tel: +31 20 592 9333, Fax: +31 20 592 4199
http://www.cwi.nl/

Danish Consortium for Information Technology
DANIT co/CIT, Aabogade 34, DK - 8200 Aarhus N, Denmark
Tel: +45 8942 2440, Fax: +45 8942 2443
http://www.cit.dk/ERCIM/

Foundation for Research and Technology – Hellas
Institute of Computer Science, P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
Tel: +30 81 39 16 00, Fax: +30 81 39 16 01
http://www.ics.forth.gr/

GMD – Forschungszentrum Informationstechnik GmbH
Schloß Birlinghoven, D-53754 Sankt Augustin, Germany
Tel: +49 2241 14 0, Fax: +49 2241 14 2889
http://www.gmd.de/

Institut National de Recherche en Informatique et en Automatique
B.P. 105, F-78153 Le Chesnay, France
Tel: +33 1 3963 5511, Fax: +33 1 3963 5330
http://www.inria.fr/

Swedish Institute of Computer Science
Box 1263, S-164 29 Kista, Sweden
Tel: +46 8 633 1500, Fax: +46 8 751 7230
http://www.sics.se/

Swiss Association for Research in Information Technology
Dept. Informatik, ETH-Zentrum, CH-8092 Zürich, Switzerland
Tel: +41 1 632 72 41, Fax: +41 1 632 11 72
http://www.sarit.ch/

Stiftelsen for Industriell og Teknisk Forskning ved Norges Tekniske Høgskole
SINTEF Telecom & Informatics, N-7034 Trondheim, Norway
Tel :+47 73 59 30 00, Fax :+47 73 59 43 02
http://www.informatics.sintef.no/

Slovak Research Consortium for Informatics and Mathematics
Dept.of Computer Science, Comenius University, Mlynska Dolina M SK-84248 Bratislava, Slowakia
Tel: +421 7 726635, Fax: +421 7 727041
http://www.srcim.sk

Magyar Tudományos Akadémia – Számítástechnikai és Automatizálási Kutató Intézete
P.O. Box 63, H-1518 Budapest, Hungary
Tel: +36 1 4665644, Fax: + 36 1 466 7503
http://www.sztaki.hu/

Trinity College, Department of Computer Science,
Dublin 2, Ireland
Tel: +353 1 608 1765, Fax: 353 1 677 2204
http://www.cs.tcd.ie/ERCIM

Technical Research Centre of Finland
VTT Information Technology, P.O. Box 1200, FIN-02044 VTT, Finland
Tel:+358 9 456 6041, Fax :+358 9 456 6027
http://www.vtt.fi/