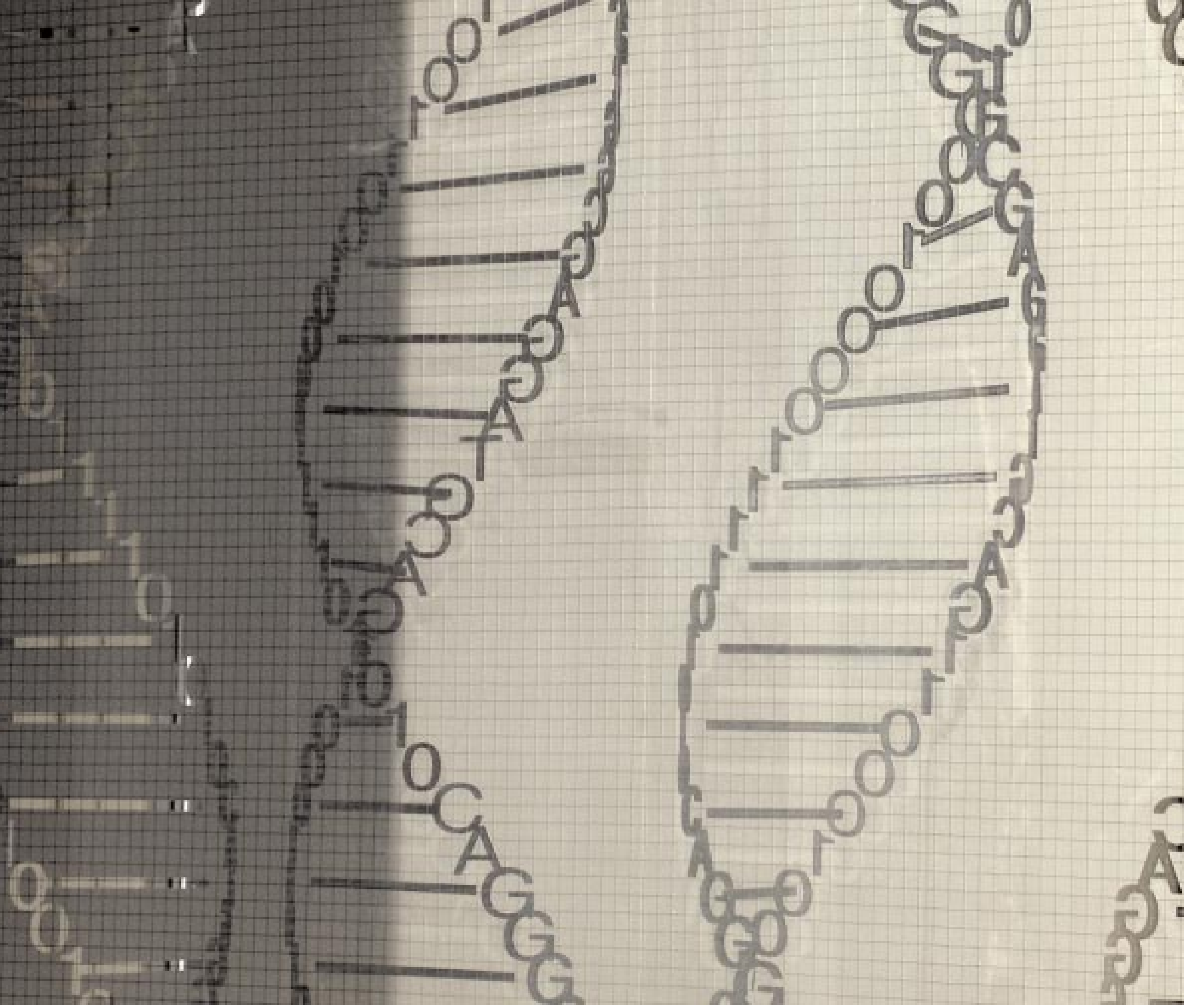


Bioinformatics

in Stockholm and Uppsala

Contents

Companies

Cover Photo: Erik Bongcam-Rudloff

Bioinformatics in the Stockholm–Uppsala Region

Bioinformatics is a new and fast developing field of science, devoted to the interpretation of biological information related to DNA and proteins. This research area requires interaction between computer science, mathematics, statistics, chemistry, biology and medicine. Bioinformatics deals with analysing, interpreting, organising and storing the enormous amounts of data from sequences and therewith associated information, such as mutations, polymorphisms, expression patterns, three-dimensional structures, protein–protein interactions, metabolic pathways, just to mention a few examples.

Huge amounts of data are generated when scientists are unravelling the genomes of different organisms. One example is the Human Genome Project, where researchers all over the world collaborate in investigating the human genome. This information leads to increased understanding of the human organism to be used in finding new routes to cure diseases. It also enhances our knowledge about the function of proteins and gives new insights to evolution. The next challenge for bioinformaticians in collaboration with other scientists is to find out how proteins act and interact in the human organism.

There is a lack of researchers in bioinformatics world-wide. The need of bioinformaticians in both academy and industry is much greater than the available number of researchers. In Sweden, we now see the build-up of both academic centres and companies focused on bioinformatics. With this guide, we will present the current state of bioinformatics in the Stockholm–Uppsala region.

There are three academic bioinformatics centres:

- The *Stockholm Bioinformatics Center*, a collaboration between Karolinska Institutet, the Royal Institute of Technology and Stockholm University
- The *Center for Genomics and Bioinformatics* at Karolinska Institutet, supported by Pharmacia AB, where bioinformaticians work closely with experimentalists
- The *Linnaeus Centre for Bioinformatics*, a collaboration between Uppsala University and the Swedish Agricultural University

Since bioinformatics is a field of strategic importance for both academic research and companies, a Ph.D. programme in medical bioinformatics was initiated in 2001. The research programme is supported by the Knowledge Foundation (KK-stiftelsen) and it is coordinated by the Centre for Medical Innovations (CMI) at Karolinska Institutet. CMI is a non-profit research and development unit at the Karolinska Institutet. CMI has produced this guide to market the bioinformatics research and companies active in the area, and to increase the collaboration between academy and industry. Researchers at the three major bioinformatics centres in the region will be presented, followed by presentation of companies and finally a description of the recently started Ph.D. programme in medical bioinformatics.

Lena Lewin, Ph.D.	Bengt Persson, Associate Professor
Project Manager	Programme Director
Centre for Medical Innovations	Centre for Medical Innovations

Stockholm Bioinformatics Center (SBC)

SBC is a joint undertaking by Stockholm University, the Royal Institute of Technology and Karolinska Institutet. It is funded by a five-year grant from the Foundation for Strategic Research and started in December 1999. SBC is located at the Center for Physics, Astronomy and Biotechnology at Roslagstull.

Presently, the research conducted at SBC is in the areas of:

1. Prediction of subcellular protein localization
2. Annotation of membrane proteins
3. Prediction of membrane protein topology
4. Protein structure prediction and quality assessment thereof
5. Algorithmics for bioinformatics
6. Sequence comparisons
7. Molecular modelling tools
8. Molecular evolution
9. Genetic networks
10. Data integration in the "-omics field" (genomics, proteomics, transcriptomics)

Homepage: www.sbc.su.se

Scientists at the Stockholm Bioinformatics Center

Prediction of Protein Localization and Membrane Protein Topology

Gunnar von Heijne,
Professor and Head of the Center
Dept. of Biochemistry and Biophysics and
Stockholm Bioinformatics Center, SCFAB
Stockholm University
SE-106 91 Stockholm, Sweden

Phone: +46 8 16 25 90
E-mail: gunnar@dbb.su.se
Web: www.dbb.su.se and www.sbc.su.se

Photo: Erika von Heijne

We develop new bioinformatics methods for the prediction of subcellular protein localization and for the identification of integral membrane proteins and prediction of their topology and three-dimensional structure. These projects are carried out in close collaboration with researchers at the Center for Biological Sequence Analysis at the Danish Technical University.

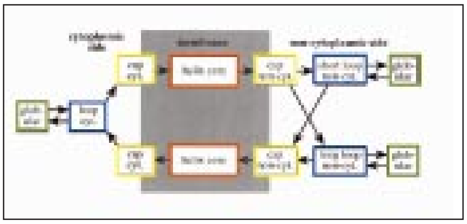
From a functional genomics perspective, the prediction of the subcellular localization of novel proteins can provide vital clues to their function. Our recently developed predictor TargetP is the first well-performing "integrated" prediction method that can discriminate between multiple subcellular locations. We will continue to develop and extend TargetP during the next few years. We will also test whether other machine learning approaches such as support vector machines can be profitably used for these kinds of prediction problems.

The identification of membrane proteins and prediction of their topology is an equally important first step in large-scale functional genomics projects, in particular as membrane proteins are considered among the most important future drug targets. We have recently developed a well-performing hidden Markov model (TMHMM) to identify membrane proteins and to predict their topology. TMHMM will be further improved by, *e.g.*, inclusion of multiple alignment information.

References

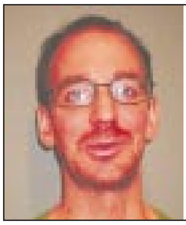
1. Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J.Mol.Biol. 300*, 1005–1016.
2. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to compete genomes. *J.Mol.Biol. 305*, 567–580.

See www.sbc.su.se/gvhpub.html for a full publication list.



The basic HMM architecture of the TMHMM membrane protein topology predictor.

Prediction of Protein Structure and Protein Function



Arne Elofsson, Associate Professor
Stockholm Bioinformatics Center, SCFAB
Stockholm University
SE-106 91 Stockholm, Sweden
Phone: +46 8 553 785 68
E-mail: arne@sbc.su.se
Web: www.sbc.su.se

Fold prediction procedures have attracted much attention recently. Indeed, though over 70% of the newly determined protein structures today are found to correspond to a known fold, sequence information alone is often not sufficient to predict this, since proteins with negligible sequence similarity can adopt the same fold. Having powerful methods for predicting the protein fold from the amino acid sequence would therefore be of great help. As function often is conserved within a fold, it also provides help for the prediction of function. Lately we have developed methods that significantly increase the specificity of fold prediction. Two of these methods, described below, have been used by people thousands of times in proteomics and other projects.

Consensus Fold Recognition

During the last years many protein fold recognition methods have been developed, based on different algorithms and utilizing various kind of information. Earlier studies confirm the expectation that for different targets, different methods produce the best predictions and the final prediction accuracy could be improved if available methods would be combined in a perfect manner. A consensus predictor, Pcons, has recently been developed which approaches this task. Pcons attempts to select the best model out of those produced by six prediction servers, each utilizing different methods. Pcons translates the confidence scores reported by each server into uniformly scaled values corresponding to the expected accuracy of each model. The translated scores as well as the similarity between models produced by different servers is being used in the final selection. Pcons outperforms any single server by generating about 8-10% more correct predictions. More importantly the specificity of Pcons is significantly higher than for any individual server. From analysing different input data to Pcons it can be shown that the improvement is mainly due to measurement of the similarity between the different models. Pcons is freely accessible for the academic community through the protein structure prediction metaserver at http://bioinfo.pl/meta/.

Membrane Protein Detection

20% to 25% of the proteins in a typical genome are helical membrane proteins. The transmembrane regions of these proteins have markedly different properties when compared to globular proteins. This presents a problem when homology search algorithms optimised for globular proteins are applied to membrane proteins. We have made modifications of the standard Smith-Waterman and profile search algorithms that significantly improve the detection of related membrane proteins. The improvement is based on the inclusion of information about predicted transmembrane segments in the alignment algorithm. This is done by simply increasing the alignment score if two residues predicted to belong to transmembrane segments are aligned with each other. Benchmarking over a test set of G-protein coupled receptor sequences shows that the number of false positives is significantly reduced in this way, both when closely related and distantly related proteins are searched for.

Description, Modelling and Simulation of Molecular Life Processes

Per Kraulis, Assistant Professor
Stockholm Bioinformatics Center, SCFAB
Stockholm University
SE-106 91 Stockholm, Sweden
Phone: +46 8 553 785 71
E-mail: per@sbc.su.se
Web: www.sbc.su.se

Functional genomics is based on the recent technological developments in biology that make it possible to obtain complete or nearly complete data sets of the constituents of biological organisms or samples. The data sets include the genome (DNA), the transcriptome (RNA), the proteome (proteins) and the metabolome (small molecules). Other emerging experimental techniques are providing information on interactions between *e.g.* proteins, or allow recording the temporal development of the levels of these constituents during defined experimental conditions. The data sets that functional genomics approaches make available pose a number of challenges:

- How should the data sets be adequately described?
- How should the data be stored to allow easy access for different kinds of analysis?
- What patterns occur in the data sets?
- How do these patterns relate to the biology of the organism?
- How should previous knowledge of the processes in biological systems be used to analyse the data sets?
- How should other related data be integrated for proper analysis?
- How should the processes in living organisms be modelled so as to explain the data?

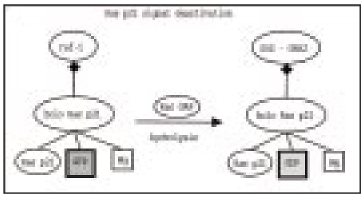
Among my current projects are:

BiPS: Biological Structures and Processes

Although there are numerous databases for describing biological data such as nucleotide and protein sequences and microarray data, there are no databases for a unified description of biological processes, nor of the structures (in a wide sense of the word) involved in these processes. Currently, nearly all of this information is available only in textual, narrative form in the scientific literature. This project investigates approaches towards the formalization of biological knowledge in terms of the material structures and the processes that transform them. The description format will use XML and software will be made available to view and edit the descriptions. Data for the cell cycle in yeast will be used as the test bed.

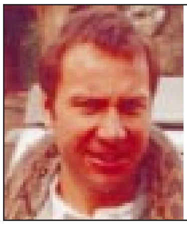
The ExOvo System

This is a software package based on a relational database management system (RDBMS) for handling and analysis of microarray (mRNA) and proteomics data. The system is based on MySQL and Python and will be made available as Open Source software.



A part of a metabolic pathway in man. New approaches towards the systematic description of biological processes may become important tools for data analysis in functional genomics, *e.g.* by showing interactions and pathways for different gene products in a process.

Development of New and Biologically Relevant Algorithms



Jens Lagergren, Associate Professor
Stockholm Bioinformatics Center, SCFAB
Numerical Analysis and Computer Science, KTH
SE-106 91 Stockholm, Sweden

Phone: +46 8 553 785 70
E-mail: jensl@nada.kth.se
Web: www.sbc.su.se

This project focuses on developing new and biologically relevant algorithms in the following areas: genome evolution, phylogeny, and identification of regulatory sequences.

One approach to reveal the function of genes is to correlate phenotype evolution with genome evolution. Genome evolution also provides an opportunity to establish the correspondence between genes in different genomes (orthology analysis), which can be used to translate knowledge of gene function in model organism to the corresponding knowledge for humans.

In a genome, the genes evolve through nucleotide substitutions. The evolution of the genome is also shaped by a multitude of other evolutionary events acting at different organizational levels. Larger genome segments are affected by processes such as duplication, lateral transfer (where a segment of an organism's genome is transferred to the genome of another organism), inversion, transposition, deletion and insertion. Being able to identify genes that have been laterally transferred and count the number of lateral transfer events is crucial for the resolution of the existence of a tree of life. Finally, the whole genome is influenced by speciation and hybridisation of organism lineages (where a new species is created by the fusion of two organisms genomes). The complexity of genome evolution poses a serious challenge in developing mathematical models and algorithms.

A classical problem in computational biology is that of inferring the evolutionary history of a set of species. The evolutionary history is represented by a phylogenetic tree. Due to duplications and lateral transfers gene trees (*i.e.* phylogenetic trees for gene families) and the corresponding species tree may disagree. We have studied the algorithmic problem: for a given set of disagreeing gene trees, find the species tree that explains the disagreement using a minimum number of duplications. We have also given a mathematically rigid and biologically sound model for lateral transfers and a fast algorithm for the problem: given a gene tree and a species tree, find the minimum number of lateral transfers that explains the difference between the given trees.

Recently, we have started developing algorithms for identification of regulatory sequences. There exist basically two algorithmic approaches to this problem. In the first, promoter regions of co-regulated genes are searched for similar substrings. In the second, promoter regions of a gene family are searched for similar substrings. The approaches yields different algorithmic questions, since in the latter case the notion of similarity can be defined relative to a species tree.

Correlating the Evolution of Genes and Proteins with the Evolution of Species

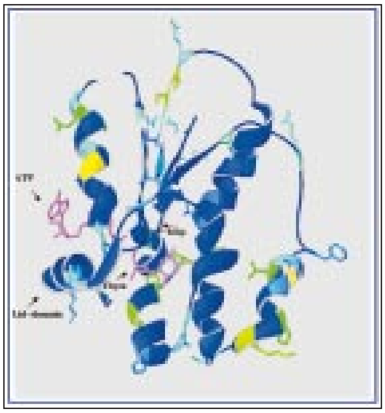
David Liberles, Assistant Professor
Stockholm Bioinformatics Center, SCFAB
Stockholm University
SE-106 91 Stockholm, Sweden
Phone: +46 8 553 785 65
E-mail: liberles@sbc.su.se
Web: www.sbc.su.se

In genomic biology, it is important to develop methods to detect selective pressures for protein encoding genes that change function along specific branches of phylogenetic trees. Along these lines, modifications of the ratio of nonsynonymous to synonymous nucleotide substitution rates (for example, to detect covarion behavior) and the development of new measures, like the Sequence Space Assessment (SSA) statsitic or the Point Accepted Mutation/Netural Evolution Distance (PAM/NED) ratio have been achieved. Further development of methods, including methods that incorporate protein three dimensional structural information are underway.

Such methods are applied to all biological sequences available, in the creation of a database, The Adaptive Evolution Database (TAED). This database includes information on protein coding sequences that appear to be undergoing adaptive evolution or changes of function along specific branches of the tree of life and has been established in a phylogenetic context. In the future, additional modes of evolution (for example changes in gene expression or in alternative splicing patterns) will be added to this database as appropriate datasets become available. This database is a resource for asking the question, what are the genes or molecular events that diverged as different species underwent adaptive divergence from a common ancestor. It is also useful in combination with the Protein Data Bank (PDB) for analysing the effect of protein structure and folding on the fixation of mutations during evolution.

Specific examples of proteins are also of interest. These are studied computationally by examining mutational patterns in a phylogenetic perspective and comparing that with three dimensional protein structures and literature mutagenesis studies to develop models for the roles of specific mutations in functional adaptation (for example in carbamoyl phosphate synthetase and in deoxyribonucleoside kinase). Other proteins of interest are studied further through experimental cloning from additional species and an analysis of mutational patterns through evolutionary history (for example plasminogen activator and myostatin).

This research paradigm fits together at the interface of several fields, enabling us to address basic questions in biology and evolution. An understanding of the role of specific mutations underlying phenotypic effects allows us to better understand the evolution of species.



A homology model of human thymidine kinase 2 is shown. The residues that have mutated since its last common ancestor with fruit fly deoxyribonucleoside kinase are indicated, color coded by mutational severity. The ATP phosphate donor and the thymidine that is specifically bound by the human, but not the fruit fly protein are indicated in magenta.

Evolving Solutions to the Protein Folding Problem

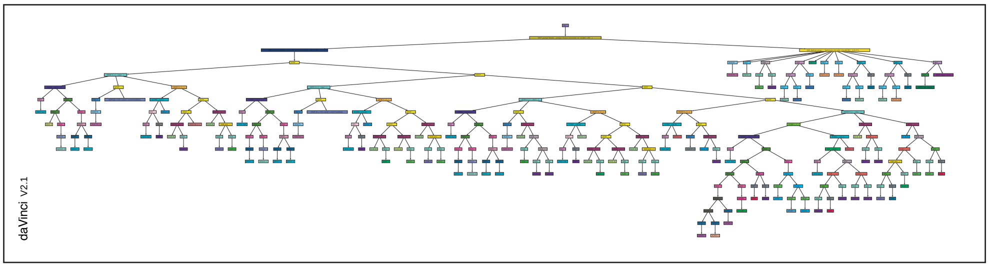
Bob MacCallum, Assistant Professor
Stockholm Bioinformatics Center, SCFAB
Stockholm University
SE-106 91 Stockholm, Sweden
Phone: +46 8 553 785 67
E-mail: maccallr@sbc.su.se
Web: www.sbc.su.se

Our major interest is to develop machine learning and evolutionary computation techniques to understand protein folding and structure. A protein starts life inside the cell as a simple linear chain of amino acids, but quickly and efficiently folds into a specific three-dimensional shape which dictates how it interacts with other molecules in the cell or organism. This interplay of proteins and other biomolecules is the very essence of life.

An organism's DNA specifies the exact sequence of amino acids for every protein. In turn, the sequence of amino acids specifies the three-dimensional shape of the folded protein. The folding process is difficult to observe experimentally and currently impossible to simulate accurately in a computer. Therefore our limited understanding of protein folding has not yet resulted in a computational prediction method which can accurately produce a 3D structure given a sequence as input.

Living organisms have overcome many difficult challenges, such as surviving in extreme climates or flying through the air, by the process of evolution, driven by chance mutation events, selection and reproduction. Evolutionary computation is an area of computer science which takes inspiration from Nature as it tries to evolve solutions to difficult problems. We hold the view that it may be possible to evolve simplified but accurate models of protein folding that no human has so far been able to come up with. We are particularly interested in genetic programming, which is the automatic evolution of computer programs, as the means to achieve this. We pay particular attention to developing evolutionary algorithms which follow the biological lead. One example is our exploration of meta-evolution – where not only the "organisms" evolve but also the evolution mechanisms.

At the same time as working on algorithms, we are also looking at aspects of sequence–structure relationships with other techniques which we hope will give valuable insight into protein structure and function. The ultimate goal is to understand complex biological systems with exciting new biologically inspired machine learning algorithms.



A tree-representation of an automatically generated computer program to predict protein structure.

Medical Bioinformatics for Comparisons and Predictions

Bengt Persson, Associate Professor
Dept of Medical Biochemistry and Biophysics
Stockholm Bioinformatics Center
Karolinska Institutet
SE-171 77 Stockholm, Sweden

Phone: +46 8 728 77 30,
+46 8 553 785 69
E-mail: bpn@mbb.ki.se
Web: www.cbb.ki.se

Our research aims at detection of protein sequence patterns and relationships, which will be used to develop prediction algorithms and to do functional assignments. The research programme also includes developments of bioinformatics tools, *e.g.* programs to automate these tasks and databases to simplify the handling of the protein sequences. The generally applicable methods are tested on several protein families in order to understand protein function, evolution and architecture. Our research interests are in three main areas – prediction algorithms, sequence comparisons/pattern searches, and molecular modelling.

Membrane prediction algorithms are developed in collaboration with prof. Gunnar von Heijne. The aim is to increase the success rate and also to estimate the prediction reliability. A first pilot study, using five different prediction methods based upon different principles, has been published [1]. We will now develop a prediction method based upon consensus predictions that will have high success rate without the need of multiply aligned sequences.

Thanks to the ongoing genome projects, the available amount of sequence information now increases at a fast pace. This sequence information can be used to extract knowledge about gene families, including functional and structural properties, sequence patterns, evolutionary relationships. As an example, a comparison of the completed genomes from man, insect, worm and plant shows that eight SDR (short-chain dehydrogenase/reductase) orthologue clusters are present in all the genomes now compared [2]. Four of these represent extended SDR forms, a subgroup found in all life forms. Four represent classical SDRs with activities involved in differentiation and signalling. We also find 18 SDR genes that are present only in the human genome, reflecting enzyme forms specific of mammals. Close to half of these gene products represent steroid dehydrogenases, emphasising the regulatory importance of these enzymes [2].

In the sequence pattern searches, one needs access to a non-redundant database, *i.e.* a database containing only unique sequences and not doublets or sub-sequences that are identical to another sequence in the database. We have therefore created a non-redundant protein sequence database, denoted KIND (Karolinska Institutet Non-redundant Database) which is regularly updated and has been made public via anonymous ftp (ftp.mbb.ki.se) [3].

Molecular modelling techniques are used to study molecular interactions and sequence variations in relation to structural changes. We have applied these methods on alcohol dehydrogenase, where the binding of bile acids was investigated showing that *iso*-ursodeoxycholic acid binds with optimal distances while the ursodeoxycholic acid cannot bind properly [4]. Molecular modelling and docking has also been performed for the SDR enzyme 17ß-hydroxysteroid dehydrogenase, where different substrates were tested [5].

References

1. Nilsson, J., Persson, B. & von Heijne, G. (2000) *FEBS Lett.* *486*, 267–269.
2. Kallberg, Y., Oppermann, U., Jörnvall, H., & Persson, B. (2002). *Prot. Sci.* *11*, 636–641.
3. Kallberg, Y. & Persson, B. (1999) *Bioinformatics* *15*, 260–261.
4. Marschall, H.-U., Oppermann, U. C. T., Svensson, S., Nordling, E., Persson, B., Höög, J.-O. & Jörnvall, H. (2000) *Hepatology* *31*, 990–996.
5. Nordling, E., Oppermann, U. C. T., Jörnvall, H. & Persson, B. (2001) *J. Mol. Graph. Model.* *19*, 514–520, 591–593.

Evolutionary Bioinformatics

Bengt Sennblad, Assistant Professor
Stockholm Bioinformatics Center (SBC),
SCFAB
SE-106 91 Stockholm, Sweden
and Center for Genomics and
Bioinformatics (CGB), Karolinska Institutet
SE-171 77 Stockholm, Sweden

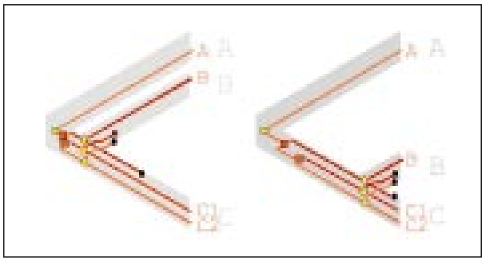
Phone: +46 8 553 785 72 (SBC) or
+46 8 728 66 88 (CGB)
E-mail: bengt.sennblad@sbc.su.se
Web: www.cgb.ki.se/cgb/groups/
sennblad/index.html

Studies of genome evolution have become increasingly important in the molecular era. The current wealth of genome projects provides an abundance of molecular data but also creates new applications for such studies. This includes orthology analysis, studies of selection and adaptation, and evolutionary studies of genome organization. These applications have strong potential value for, *e.g.*, prediction of protein function, genetic engineering and drug design. For studies of genome evolution, there has emerged an urgent need for evaluation in a probability perspective and, more important, a need to put a reliability measure or a confidence on results.

The long-term aim of this project is to develop a unified probabilistic framework and software system for genome-wide analyses of molecular evolution, with possibilities to incorporate data and models relating to various aspects of genome biology, *e.g.*, orthology, sequence evolution, selection, adaptation, and genome organization. Hierarchical Bayesian analysis combined with Markov Chain Monte Carlo (MCMC) methods provides such a probabilistic framework for genome evolution studies. It also provides a powerful reliability measure, "posterior probability", allowing us to determine the probability for individual hypotheses or to identify confidence sets comprising the most probable hypotheses.

Current Projects Include

- Probabilistic orthology analysis (see Figure)
- Detection of adaptation using context-dependent models



Probabilistic orthology analysis. Orthology analysis is important for prediction of gene function. Orthologous genes are closest related through speciations and are more likely to have the same function. Orthology analysis is performed by mapping a gene tree on to a species tree. A problem is that such a mapping will depend on the time between speciations. In the figure, the same gene tree (red) mapped on two species trees (gray) with the same topology, but with different times between speciations (yellow circles). Because of the shorter time for gene duplications (red squares) to occur, the left mapping is less probable than the right. We are developing probabilistic models for orthology analysis that takes this into account.

Computational Biology

Jesper Tegnér, Assistant Professor
Stockholm Bioinformatics Center, SCFAB,
Numerical Analysis and Computer
Science, KTH
SE-106 91 Stockholm, Sweden

Phone: +46 8 553 785 64
E-mail: jespert@nada.kth.se
Web: www.nada.kth.se/~jespert,
www.sbc.su.se/invest.html

Our mission is to develop tools and concepts in order to identify, understand, and control fundamental principles of living matter. We use theory and computational techniques to address issues ranging from basic research to biotechnological applications. Current themes are:

Algorithms for Identification of Cellular Networks

Remarkable progress in genomic research produces an ever increasing catalogue of genes and proteins participating in the machinery of life. To identify those cellular communication networks, we design algorithms coupled with computational models to support inference of the their architecture from micro-array like measurements. One class of techniques we have developed is based on systematic perturbations of the ongoing gene dynamics using synthetic gene circuits. Together with Cellicon Biotechnologies and researchers at Boston University we now validate our tools experimentally and investigate their further use for drug design.

Learning Algorithms in Natural and Synthetic Systems

Using tools from statistical physics we have analysed spike time-dependent learning rules as found in the brain. We have designed a self modifying learning rule which also accounts for cortical deprivation experiments. Currently we investigate whether new machine learning algorithms can be derived from our rule. Plastic self adapting systems operating at different timescales could be fundamental to biological and engineered networks.

Transient and Sustained Forms of Memory

We study working memory, the ability to hold things in the mind without sensory input, crucial for thinking, and future planning in man and machines. In one study of visuospatial working memory, we have introduced a novel principle of disinhibition which accounts for available data and could form a basis for building cognitive architectures for decision, attention, and working memory capacity. Our memory networks can be used in robotics and for sub-cellular computations.

Computing the Natural Way and Network Dynamics

Together with Boston University we participate in the NSF program for biocomputing, here defined as the ability of cells to make decisions and computations based on external stimuli. Synthetic gene networks may provide a means to control complex biochemical systems in much the same way that digital and analog circuits provide a means to control electronic and mechanical systems. We have identified design criteria, for the experimental construction of a library of plasmids, such as Boolean circuits and genetic memory networks. Our work set the stage for using cellular control and computation at the DNA level, thus permitting sophisticated schemes for drug delivery and providing new tools to unravel mechanisms of gene regulation. In parallel, we examine computational capabilities of networks of genes, proteins and neurons.

Center for Genomics and Bioinformatics (CGB)

The CGB is a department at Karolinska Institutet, which was started as a collaboration with Pharmacia Corporation in 1997. The CGB hosts a Bioinformatics Unit with about 20 members.

Presently, the bioinformatics research conducted at CGB is in the areas of:

1. Data integration of genomic, proteomic, transcriptomic, phenotypic, and other information
2. Analysis of protein families, protein domains, and their evolutionary origin
3. Automated methods for orthology assignment
4. Integration of predicted functional and structural features of proteins
5. Prediction of antisense oligonucleotides for inhibition of gene expression
6. Analysis of sequences regulating transcription
7. Regulatory region detection by phylogenetic footprinting
8. Analysis and prediction of functional genome variations
9. SNP genotyping data mining for research into complex disease
10. Automated design of SNP genotyping assays
11. Population genetics

Homepage: www.cgb.ki.se

Bioinformaticians at CGB

Genome Analysis

Björn Andersson, Associate Professor
Center for Genomics and Bioinformatics
Karolinska Institutet, Berzelius väg 35
SE-171 77 Stockholm, Sweden
Phone : +46 8 728 39 87
E-mail: bjorn.andersson@cgb.ki.se
Web: www.cgb.ki.se

The main focus of my group is genome research based on a genome sequencing activity. Our main sequencing project is genome sequencing in the protozoan parasite *Trypanosoma cruzi*. Other areas of interest is functional genomics and bioinformatics. The latter provides both an interesting field of research, but also tools that are necessary for the production and analysis of genomic data. Our bioinformatics research thus contains both the development of tools for routine analysis of data as well as more advanced research into new algorithms for sequence analysis.

The largest project to date is the development of a new program for the assembly of shotgun sequences, Tandem Repeat Assembly Program (TRAP). This set of new methods was developed in order to solve the problem of erroneous assemblies of nearly identical repeats, which has been a major drawback of the shotgun approach for large scale sequencing. TRAP identifies the positions that differ between repeat copies using statistical methods and uses these to produce correct assemblies. TRAP and the underlying algorithms have recently been published in two papers. Currently, the program is being improved further using several different approaches, and a version that can be exported in a user-friendly manner is being produced.

As a spin-off of this project, a program for error correction of shotgun data is under development. This program is designed to be used in conjunction with assembly programs to produce error-free data sets that can improve the performance of these programs.

A major effort has been put into the handling and analysis of *T. cruzi* data. The tools developed include a web-based system for uploading and handling large amounts of data and a software package for analysis and semi-automated annotation of genomic data. The latter system is designed to carry out all steps required for rapid analysis, including gene finding, database searches, feature finding etc, and display them in a format where they can be viewed by the user. We are currently adding to these systems with several local databases.

A major project in the group is the development of new methods for functional classification of genes and the identification of regulatory sequences in the *T. cruzi* genome. Several possible methods are available for this and we are currently testing different approaches.

In summary, a major driving force of the field of bioinformatics has been genome sequencing. It has turned out that the combination of genomics and bioinformatics under the same roof and the close communication between the activities has in many cases resulted in very strong research. We have therefore built our bioinformatics activity close to where data is being produced and working with issues generated within the genome projects. This has resulted in the projects listed above amongst others and to the presence of several trained bioinformaticians.

Cataloging and Experimental Utilization of Genome Variation

Anthony J Brookes, Professor
Vice-Prefect, Clinical Genomics Unit
Coordinator
Center for Genomics and Bioinformatics
Karolinska Institutet, Berzelius väg 35
SE-171 77 Stockholm, Sweden

Phone : +46 8 728 66 30
E-mail: anthony.brookes@cgb.ki.se
Web: www.cgb.ki.se

Major Public SNP Database HGBASE

HGBASE was created in mid-1998 and is run today by a pan-European consortium including our group at Karolinska Institutet, Stockholm, teams at the EBI, UK, and teams at EMBL, Germany. Using a MySQL platform, plus Visual basic, Perl, and Java procedures, we collect and represent all publicly available human genome variations (from databases, literature publications, and via receipt of submissions). As of October 2001, a wide range of high-quality records had been combined into a non-redundant list of >1,000,000 sequence variations – indicating for each the original data source and submitter. Data exchange with dbSNP (SNP repository at the NCBI) was established at the end of year 2000. Population allele frequencies are included whenever available, and considerable efforts are made to relate polymorphisms to human genes, detailing consequences for coding regions, promoters and splice sites. Core HGBASE data is also represented within the ENSEMBL project at EBI. Automated and manual data checking ensures internal consistency, and addresses errors sometimes present in the original source information. HGBASE is freely available at http://hgbase.cgr.ki.se and complete downloads are also possible (XML, Fasta, SRS and tagged text files). Sophisticated online search tools facilitate data interrogation by sequence similarity and by keyword queries. Major ongoing developments include the provision of genotyping assays for every SNP, functional predictions based upon various aspects of protein structure and amino acid conservation, genotype and haplotype representations, phenotype descriptions, and genotype-phenotype correlations. The latter is being enabled by a broader collaboration with GDB (Canada) and a USD 1M application to the NIH now under consideration.

Software to Assist Laboratory Analysis of Genome Variation

Bioinformatics support is essential for effective SNP utilization, and we are developing a wide range of purpose-built software and genotyping database resources. These include; i) enhanced local query facilities for HGBASE, ii) a local database system for storing validated genotyping assays, extendable to a laboratory management system, iii) an internet-based DASH Users-Network service (DASH-Net) [DASH is a locally developed SNP genotyping procedure], iv) software for automated DASH assay design (also to be applied to all HGBASE entries), v) an in-house database for storage, integration and retrieval of genotype data, and vi) extensive statistical resources for data interpretation including both main-stream packages and novel tools based upon data randomization. These components need to be developed further to i) increase their relative integration (*e.g.*, linking the genotype database to the statistics tools), ii) enable the import of data from other genotyping platforms and laboratories, and iii) convert each to stand-alone modules that can be exported to other laboratories.

Large Scale Data Analysis for Molecular Biology

Mark Reimers, Assistant Professor
Center for Genomics and Bioinformatics
Karolinska Institutet, Berzelius väg 35
SE-171 77 Stockholm, Sweden
Phone : +46 8 728 78 26
E-mail: mark@kisac.cgb.ki.se
Web: www.cgb.ki.se

The massive data sets that are now coming out of micro-array experiments, and will be coming out of proteomics experiments, offer a new kind of challenge for data analysis. Mark Reimers comes to KI from a biotech company in Boston, USA; his background is mathematics, statistics, and computer science. He works in large scale data analysis using approaches from classical statistics, Bayesian inference, and machine learning.

At KI his role includes working closely with experimentalists. The key to doing useful work with experimentalists is rapid and robust inference. It must be rapid, in order to generate null distributions through repeated randomisation of the large data set. The inference must be robust, because inevitably some results in a high-throughput experiment fail, or generate questionable results. A skilled experimentalist will be able to recognize many of these failures from experience; the challenge for the data analyst is to incorporate some of the "common sense" of the experimentalist explicitly into the analysis protocol.

Specifically, Mark works with design and analysis of micro-array experiments. He has written code for efficient analysis of multiple-treatment experiments, and also for time-series studies. He has written programs for rapid estimation of haplotype frequencies using high-throughput genotype data. Mark also takes over management of the KI Sequence Analysis Center, which provides bioinformatics tools and consulting locally to researchers at KI.

Statistics has undergone a tremendous change over the last twenty years, as computer-intensive techniques have become widely used. We are only beginning to see the changes in approach that accompany these new techniques, such as routinely generating empirical null distributions, and generating p-values for path-dependent processes. Molecular biology has also experienced a rapid acceleration of certain experimental techniques. The conjunction of these two changes should be fruitful for both disciplines.

Computational Tools to Predict Novel Protein Sequences

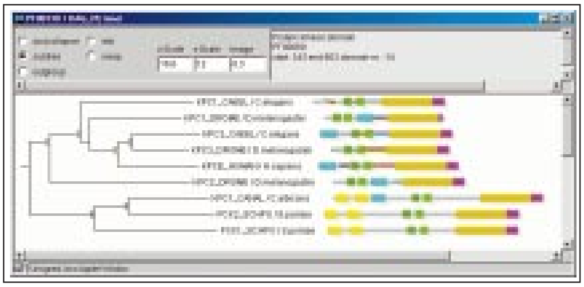
Erik Sonnhammer, Professor
Center for Genomics and Bioinformatics
Karolinska Institutet, Berzelius väg 35
SE-171 77 Stockholm, Sweden
Phone : +46 8 728 63 95
E-mail: erik.sonnhammer@cgb.ki.se
Web: www.cgb.ki.se

Professor Erik Sonnhammer's research group is working on development and application of bioinformatics methods for functional inferences in genes and genomes. A major goal is to use computational tools to predict the function of novel sequences, which can then be selected for experimental study by experimental groups. Sequence similarity-based methods such as hidden Markov models and phylogenetic trees are used, together with graphical visualization tools.

Many proteins contain multiple domains, and methods are being developed for detection and analysis of protein domains. The Pfam database plays a central part in this line of work. Protein function is usually derived from sequence homology; in addition we are using various prediction models, e.g. for transmembrane topology to better understand protein function. To assist experimental analysis, we also developed a model for predicting efficient antisense oligos.

Current Projects

- Development of novel methods for analysis of protein families, protein domains, and their evolutionary origin.
- Automated methods for orthology assignment.
- Functional classification of receptors in the genome of the worm C. *elegans* and extraction of functional counterparts (orthologs) in human and fly.
- Discovery of novel protein domains and families.
- Prediction of optimally effective antisense oligonucleotides for inhibition of gene expression.
- Functional and structural analysis of protein sequences using compositional sequence properties.



A phylogenetic tree of a protein kinase domain.

Gene Regulation Bioinformatics

Wyeth Wasserman, Assistant Professor
Center for Genomics and Bioinformatics
Karolinska Institutet, Berzelius väg 35
SE-171 77 Stockholm, Sweden
Phone : +46 8 728 63 91
E-mail: wyeth.wasserman@cgb.ki.se
Web: www.cgb.ki.se

Our research is motivated by a desire to understand how sets of genes can be coordinately activated at the transcriptional level within cells in response to external stimuli. By applying the tools of bioinformatics to experimental data, our research aims to: (1) develop systems to identify transcriptional regulatory regions within genomic sequences; (2) identify sequence patterns characteristic of transcription factor binding sites; and (3) integrate biological information to facilitate access to genome-scale data for life scientists.

Identification of Modules Regulating Transcription

In the past, computational approaches to the analysis of regulatory sequences within the human genome suffered from low predictive specificity. Laboratory evidence has accumulated indicating transcriptional regulation is often directed by a combination of transcription factors binding to locally dense clusters of regulatory elements termed modules. This combinatorial structure provides the regulatory specificity and diversity required by cells, and suggests alternative computational approaches for the delineation of regulatory sequences within a genome. We have developed computational models to identify regulatory regions that direct transcription to specific tissues, and have recently shown that these models can be applied on a genome-scale to accurately predict patterns of gene expression.

De Novo Identification of Binding Sites in Co-regulated Genes

In most cases, data is too sparse to generate module models. To address sets of co-regulated genes in these cases, one must identify the initial set of regulatory regions computationally. In order to produce meaningful results in the analysis of human genomic sequences, we have been developing approaches based upon "phylogenetic footprinting." Over the course of evolution, sequences with a specific biological function (such as gene regulation) tend to be conserved. Thus comparisons between orthologous genomic sequences allows one to differentiate functional regions within a background of more varied sequences. By applying pattern searching algorithms to the reduced "conserved sequence space", potential transcription factor binding patterns can be identified. We are working to extend this approach based on known architectures of regulatory modules (for instance neighbor-neighbor interactions and spacing between transcription factor binding sites).

Integration and Summarization of Biological Data

The greatest limitation in the development of algorithms to address biologically relevant questions is the availability of well-defined and carefully constructed data collections. While high-throughput technologies are producing enormous pools of data, it is preferable to initiate the computational studies with a set of well-understood controls. To this end, we have invested considerable effort to produce well-curated datasets for our investigations. As a tool for our efforts, we have developed a "gene-centric portal" to the human genome. The system is available online at http://www.genelynx.org

The Future

Computational biology evolves rapidly. Our core interest in understanding how cells perceive signals to alter the expression of batteries of genes will remain constant, but the specifics will change over time. In the future the group will apply computational approaches to increase understanding of the impacts of chromatin structure on gene regulation.

Linnaeus Centre for Bioinformatics in Uppsala (LCB)

LCB is a joint initiative of Uppsala University and the Swedish University of Agricultural Sciences. It was established in 1999 and is funded by a five-year grant from the Knut and Alice Wallenberg Foundation. LCB is located at the Biomedical Center in Uppsala.

Presently, the bioinformatic research conducted at LCB is in the areas of:

1. Development of object-oriented comparative sequence databases
2. Genome assembly, annotation and comparative genome sequence analysis
3. Statistical methods for genetic analyses of multifactorial traits and human disorders
4. Clustering methods for comparative analyses of RNA-expression data
5. Visualization tools for RNA- and protein expression information
6. Methods for protein structure prediction
7. Algorithms and network models for phylogenetic inference
8. Mathematical modelling of molecular evolutionary processes
9. Mathematical modelling of genetic control circuits in bacteria
10. Proteometrics and pharmacometrics for protein classification and rational drug design

Homepage: www.linnaeus.bmc.uu.se

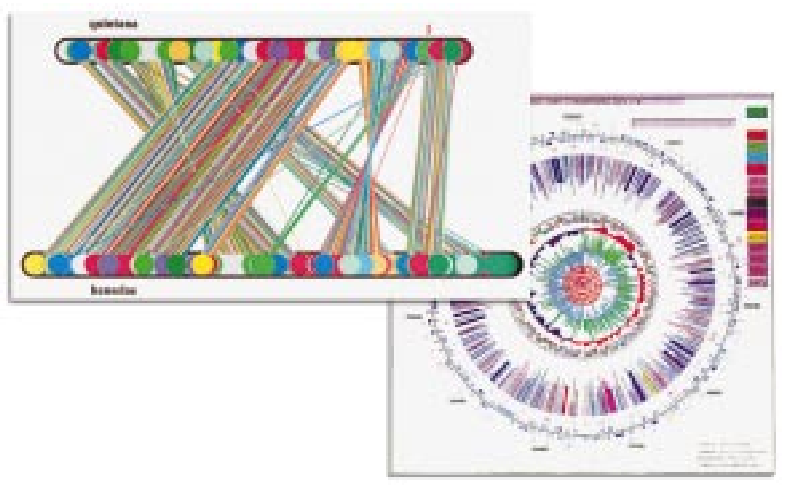
The Linnaeus Centre for Bioinformatics

Microbial Genomics and Bioinformatics

Siv Andersson, Professor and Chair,
Department of Molecular Evolution
Linnaeus Centre for Bioinformatics (LCB)
and Dept. of Molecular Evolution,
Uppsala University
Evolutionary Biology Center,
Norbyvägen 18 C,
SE-752 36 Uppsala, Sweden
Phone: +46 18 471 43 79
E-mail: siv.andersson@ebc.uu.se
Web: web1.ebc.uu.se/molev/staff/siv.html

We exploit new technologies and whole-genome based approaches to attack fundamental questions concerning the emergence and evolution of infectious diseases. So far, we have sequenced four microbial genomes including the genomes of the typhus pathogen, the causative agents of trench fever and cat-scratch disease as well as the genome of an aphid endosymbiont. By comparative genome analyses of closely related strains and species of human pathogens, we try to find the correlation between patterns of genome evolution and reproductive strategies such as population structures, virulence features and modes of transmission. Based on the genome sequence information we have initiated comparative analyses of gene expression profiles in response to environmental stimulus.

As a complement to our experimental studies, we are heavily engaged in bioinformatics research. For example, we have developed an annotation system that provides a extensive links between different sources of data as well as a tool for comparative analyses of genomic architectures and gene order structures. To study the origin and evolution of biochemical pathways in microbes, we have developed automatic methods for large-scale phylogenetic reconstructions. In order to categorize the gene products identified in the genome projects based on their expression patterns and interactions, we are currently developing visualization tools for comparative analyses of gene expression profiles that are based on charts for general descriptions of metabolic pathways.



At the department for molecular evolution, we develop methods and software for visualization of genomic architectures and whole-genome based phylogenetic analyses.

Evolutionary Analysis of the Mammalian Genome

Tomas Bergström, Assistant Professor
Linnaeus Centre for Bioinformatics &
Dept. of Genetics and Pathology
Section of Medical Genetics
Rudbeck Laboratory
Uppsala University
SE-751 85 Uppsala, Sweden
Phone: +46 18 471 48 17
E-mail: tomas.bergstrom@genpat.uu.se
Web: http://beluga.genpat.uu.se/

Research

The vast amount of DNA sequences that are currently available in various databases offers new and exciting possibilities for evolutionary analysis as well as challenges for data handling. By analysing substitution pattern in orthologous nucleotide sequences in primates, it is possible to address questions regarding both nucleotide sequence evolution (*e.g.* the effect of base pair composition on mutation processes, selection) and evolution at the level of populations and species. Furthermore, to understand the effect of genes associated with human diseases, it is often fruitful to make sequence comparisons with orthologous genes in order to reveal regions that are conserved by functional constrains.

Few software are currently available for evolutionary analysis and data handling of large genomic sequences. We are therefore developing a coherent framework of databases and software for evolutionary analysis of genomic regions. To successfully integrate data and applications it is critical that data and applications can be accessed in a standardized manner that is transparent to the user and offer interoperability between different platforms. We are therefore developing the software in the JAVA programming language and to integrate software and databases the Common Object Request Broker Architecture (CORBA) is used. This is thus a distributed computing approach where we can integrate software that are already available into our interface. For example, there is no need to re-implemented the Jukes Cantor algorithm that corrects for multiple hits. Third party applications, that we integrate in our interface, can thus be written in any programming language, be running on any platform or operating system and be physically located on any networked computer. In the same way, data from many different databases can be integrated to present a more complete picture of the available information of a particular genomic region.

Current Projects

- Development of an annotated object-oriented comparative sequence database with a CORBA layer
- Development of a Java interface for Evolutionary analysis
- Analysis of substitution pattern in non coding regions

Bioinformatics and Comparative Genome Analysis

Erik Bongcam-Rudloff
Assistant Professor
Linnaeus Centre for Bioinformatics (LCB)
Uppsala and Dept. of Animal Breeding
and Genetics, SLU
Linnaeus Centre for Bioinformatics
BMC
Box 598
SE-751 23 Uppsala, Sweden
Phone: +46 18 471 45 25
E-mail: erik.bongcam@bmc.uu.se
Web:
http://bioinformatics.bmc.uu.se/bongcam
http://www.embnet.org/

Scientific Manager of EMBnet's (European Molecular Biology network) Swedish node. The Swedish node was established in 1989 to serve as a redistribution hub for bioinformatics databases to some 7–9 universities where the actual end-user services were set up. A national bioinformatics teaching resource was added. Courses were given yearly at locations ranging from Lund University in the south to Umeå University in the north. Bioinformatics is now becoming an ordinary academic discipline at most Swedish universities, with the Stockholm Bioinformatics Centre and the Linnaeus Centre for Bioinformatics in Uppsala as examples of strong commitments. Consequently, the role of the national EMBnet node has changed. The node has been affiliated to the Linnaeus Centre. Its focus has changed from redistribution and teaching to becoming a knowledge resource, setting up and testing systems that can serve as a model. Keeping up international contacts in the field of Bioinformatics is an important part of the node responsibilities. A limited end-user sequence retrieval and database service will be maintained.

Current Research

A "working draft" of the human genome sequence is now available and several other genomes are under way. Even though a finished, high-quality sequence of the human genome will not be available until after one or two years, the working draft sequence has the information, sufficiently close to complete, that it marks a major milestone in modern biology.

Comparisons with the sequences of mouse and other species is a powerful approach to identifying functional segments of the non-coding regions, such as gene regulatory elements. We are at the moment creating novel analytical methods to discover cis-regulatory elements based on cross-species comparison.

The aim of this work is to create an eukaryotic promoter-database and a public web-portal with freely available software. The use of this portal, in conjunction with the expanding array of publicly available resources, should make analysis of non-coding regions accessible to all interested investigators.

Computational Functional Genomics Approximate Reasoning Methods in Biology

Jan Komorowski, Professor in Bioinformatics
Linnaeus Centre for Bioinformatics, BMC
Box 598
SE-751 24 Uppsala, Sweden
Phone: +46 18 471 66 97
E-mail: jan.komorowski@lcb.uu.se
(from Summer 2002, full time at Linnaeus Centre for Bioinformatics)
Norwegian University of Science and Technology, Trondheim, Norway
Dept. of Computer and Info. Science
E-mail: jan.komorowski@idi.ntnu.no
(spring 2002)

Dr. Komorowski's current research focus is Computational Functional Genomics and approximate reasoning methods from uncertain and incomplete data applied to a variety of biomedical problems. He develops theories, methods and tools for the analysis of complex systems in biomedicine from measured data and background knowledge. Dr. Komorowski is a graduate of Warsaw University in Computer Science and did his PhD work at Linköping University. He was Assistant Professor with Harvard University and Research Associate with Harvard Medical School from 1982 to 1988. From 1990 to 2002 he was Professor of Computer Science and since 2000 Director of Computational Biology Laboratory with the Norwegian University of Science and Technology in Trondheim. Dr. Komorowski has been nominated Professor of Bioinformatics at Linnaeus Centre for Bioinformatics in November 2001 and will head bioinformatics developments at the centre from January 2002, part-time, and from Summer 2002 full-time.

Together with his group and colleagues, Dr. Komorowski established some of the first supervised learning approaches to the synthesis of computational models of complex biological systems. Using Gene Ontology and time-series gene expression data, this methodology assigns function to unknown genes and provides hypotheses of new function for known genes. In an earlier research, he and his group have made major contributions to datamining methods based on rough sets, a mathematically well-defined formalism to reason with approximate data. In co-operation with Warsaw University he has lead the development of the ROSETTA system for datamining. The system has now almost 3,000 user's worldwide (December 2001). Rough set methods and ROSETTA appeared to be well suited to the analysis of biomedical data and, in particular, to the analysis of gene expressions using background knowledge.

In another line of research, the PubGene system that mines publicly available gene and text databases (the bibliome) for high-throughput gene expression analysis has been developed in his group and is publicly available following the publication in Nature Genetics in May 2001.

Current Projects

- Gene expression studies in gastric carcinoma
- Changes of gene expressions in carcinoma cells exposed to unsaturated fats
- Time series analysis of yeast shock treatments
- Responses of *Arabidopsis* to nutrient changes and the corresponding gene expression changes
- Insect–plant interactions
- The Norwegian Microarray Consortium Datawarehouse

Phylogenetics – Evolution in Retrospect

Mikael Thollesson, Assistant Professor
Linnaeus Centre for Bioinformatics
and Evolutionary Biology Centre
Molecular Evolution
SE-752 36 Uppsala, Sweden

Phone : +46 18 471 64 01
E-mail: mikael.thollesson@ebc.uu.se
Web: http://web1.ebc.uu.se/molev/staff/
mikael.html

Research

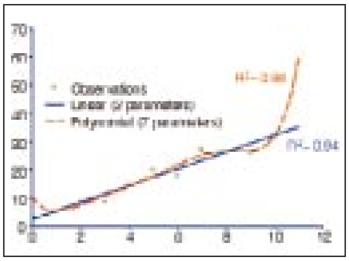
Biological replicators, such as genes or species, share a more or less inclusive evolutionary history – a phylogeny. The implications of this are recognized in most areas of biology. It is evident that the phylogenies are important when the aim is to untangle the evolution of things like metabolic pathways, but they are also important in applied contexts. For example, phylogenies have applications in disease tracking (like the origin of HIV, HIV transmission between people, and prediction of features of next years influenza virus) and the evaluation of biodiversity for conservation. They are also important in comparative studies; species (or strains, or genes in a gene family,...) are not independent samples because of their shared history. Comparative studies must include the phylogenetic background to deal with this fact. For example, evaluating the effects of a medical treatment on various pathogens without a phylogenetic context can lead to serious errors. There is thus a widespread need for tools that produce good phylogenetic hypotheses along with credibility measures for these hypotheses.

From a bioinformatic viewpoint, phylogenetic inference faces three different kinds of problems, requiring components from mathematics and computer science as well as biology: (1) To model evolutionary change adequately and efficiently. (2) To find ways to evaluate the huge number of phylogenetic hypotheses that are possible also for a small number of genes or species. (3). To adapt the inference process to be able to cope with the ever-increasing amount of data that are produced.

Bayesian inference implemented using Markov Chain Monte Carlo (MCMC) algorithms makes the use of complicated models in phylogenetic inference feasible. One line of investigation of the project is to develop, adapt and implement models for phylogenetic inference using this technique. To increase computational speed calculations may be done concurrently; by splitting the problem into smaller parts, these can be processed by several computers, *e.g.*, on a cluster of standard Linux computers (a.k.a. Beowulf cluster). Implementing phylogenetic inference methods on a Bewowulf cluster is a second line of investigation of the project. A third is to tie component software together to automate the process of phylogenetic inference (usually using Perl and the BioPerl framework), with a special focus on creating tools for the evaluation of the results.

Current Projects Include

- Model based phylogenetic inference using gene order data
- Non-homogenous models in MCMC framework
- Model selection and overadaptation
- Implementing algorithms for concurrent processing using message passing (MPI)
- Evaluation and presentation of large numbers of partly incongruent phylogenetic trees



Overadaptation – using too many parameters – in a model. Two different models are adapted to a data set (linear with random variation). A polynomial model (seven parameters, a to g; $y=a+bx^1+cx^2+dx^3+ex^4+fx^5+gx^6$) fit the data better than a linear model (two parameters, a and b; y=a+bx). However, the estimate of a specific parameter the models have in common, *e.g.*, the y-intercept a in this case or a tree in a phylogenetic context, will be worse for the complex model. The problem of selecting an adequate model for phylogenetic inference is one of the research topics of the project.

Amersham Biosciences

Lennart Björkesten
Amersham Biosciences
Uppsala, Sweden
Phone: +46 18 612 05 57
E-mail: lennart.bjorkesten@eu.amershambiosciences.com
Web: www.amershambiosciences.com/sweden/

The completion of the human genome has become the catalytic agent for many high throughput efforts addressing systematic mapping of gene function. While the human genome was a huge effort, the real challenge is still ahead. While the three billion nucleic acid bases show some variation among individuals, gene function adds several dimensions to the task. Proteins are expressed at different rate in different cells at different time governed by internal and external conditions. Proteins are modified and interact in extremely complex reaction patterns. The implication on data analysis is significant.

Centre of Excellence in Uppsala

The Amersham Biosciences software and informatics department in Uppsala is engaged in several internal and external research projects. Two examples are being outlined below. Other activities are devoted to various aspects of mass spectrometry data analysis.

Massive Parallel Differential Protein Expression Analysis

Differential 2D-gel electrophoresis (DIGE) is a novel technology for the analysis of protein abundance. The Uppsala group is developing algorithms and data analysis tools allowing for accurate differential analysis of several thousands of proteins simultaneously. Time spent on data analysis is reduced from days to hours. Figure 1.

Sub-cellular Mapping of Protein Relocation

Automated analysis of spatial, spectral and temporal information from living (or dead) cells is crucial in high-content cell screening applications. Algorithms for cell segmentation and feature analysis are being developed together with Centre for Image Analysis at Uppsala University. Figure 2.

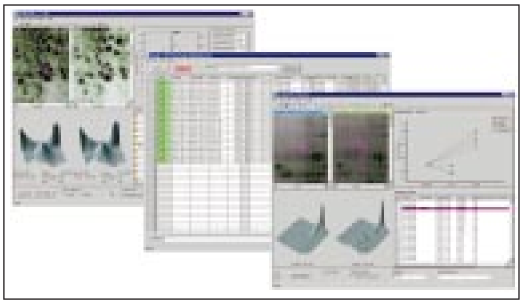


Figure 1. The DeCyder™ software package with novel algorithms for expression analysis.

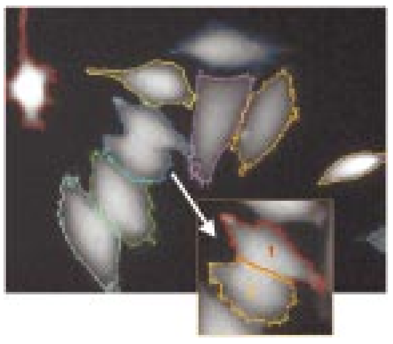


Figure 2. Cytoplasm signature in fluorescent cells.

DeCyder is a trademark of Amersham Biosciences Limited.

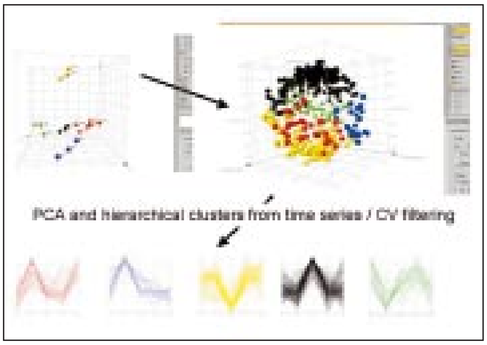
AstraZeneca

Hugh Salter
AstraZeneca, Södertälje, Sweden
Phone: +46 8 553 234 06
E-mail: hugh.salter@astrazeneca.com
Web: http://www.astrazeneca.com

AstraZeneca is one of the world's leading pharmaceutical companies. Backed by a strong research base and extensive manufacturing and commercial skills, the company provides a powerful range of innovative products that meet patients' needs in important areas of healthcare. Sales in 2000 totalled $15.8 billion (£10.6 billion, SEK150.8 billion) with an operating profit of $4 billion (£2.7 billion, SEK38 billion).

Within the Stockholm region, basic research within AstraZeneca is focused on CNS and pain indications, which includes research into diseases such as Alzheimers and multiple sclerosis. AZ has an applied bioinformatics research group working in this area. The group is based at Novum Research Centre in Huddinge and headed by Hugh Salter. As a whole, AstraZeneca has around 80 bioinformatics staff around the world, working in a range of problems from infrastructural projects to algorithmic development.

Research within the Stockholm group is focussed on the development and application of bioinformatics methods to drug discovery problems. The group is currently focused on genome/SNP analysis, comparative genomics, sequence & domain analyses, and on multivariate analysis of microarray and proteomics data. Group members are actively involved in teaching and education programs at several local universities, including KI, SU, UU and Södertörn Högskola, and at any given time there are several university students working within the group on MSc or PhD theses.



Data from a gene expression experiment deconvoluted using multivariate analysis techniques (analysis by Kerstin Nilsson).

Biovitrum

Bioinformatics Core Group
Biology Department Biovitrum AB
SE-112 76 Stockholm, Sweden
Phone: +46 8 697 32 66 (Staffan Lake)
Phone: +46 8 697 24 49 (Sarah Hunter)
E-mail: staffan.lake@biovitrum.com
Web: http://www.biovitrum.com

Bioinformatics is a driving and supporting technology for Biovitrum Research. The priority is to provide the necessary Bioinformatics resources to our current programs where Bioinformatics plays an important role. Secondly we implement the necessary infrastructure to maximize our investments in other enabling technologies, such as DNA microarrays. We recognize and identify Bioinformatics needs that are specific to our therapeutic area, from those that are general. Significant efficiency gains can be realized by having all common needs addressed by our core group.

Project Support

The participation in target and drug discovery project is our highest priority. This support comes from a Bioinformatics scientist participation in the research projects. Some activities that would be specific to support Metabolic Diseases projects are:

- Metabolism and Signalling pathways
- Access to information about model organisms: the implementation of the NIDDM project based on pathway analysis in model organisms would require databases for organisms with completed genomes

There are certain aspects of project support that benefit from a generalized approach:

- Project databases, organized so all the data that is relevant to a given project can be easily accessed and shared by all project participants
- Molecular platforms: in many cases our interest is focused on specific protein families
- Technology platforms, for example sequencing and expression analysis technologies will have common needs
- Extended collaborations: we maximize the potential benefits from such interactions

Bioinformatics Research

In order to establish a Bioinformatics approach that would provide a strong leading role in our new discovery paradigm, we perform activities that go beyond the support functions. These activities explore new methodologies and develop new approaches in informatics-based research. These are some activities in this area:

- Pathway analysis
- Comparative Genome Analysis
- Database mining
- Structural Bioinformatics

End User Support

This function provides the Biovitrum with access to information and analysis tools that are necessary for their research activities. Many activities are performed:

- Access to external and external sources of information
- Analysis tools
- Commercial vendor evaluations
- Training

Global Genomics

Sten Linnarsson
Global Genomics
Tomtebodavägen 21 B
SE-171 77 Stockholm, Sweden
Phone: +46 8 728 47 18
E-mail: sten@globalgenomics.com
Web: www.globalgenomics.com

Global Genomics AB develops and commercialises innovative technologies in biomedicine, specializing in post-genomics applications combining state-of-the-art molecular biology, bioinformatics and nanotechnology. The company was founded in 2000 by Professor Patrik Ernfors and Dr Sten Linnarsson. It is getting ready to launch its PCR-based, truly global, gene expression analysis technology as a web-based service in 2002. The method uses advanced combinatorial assignment algorithms to find an optimal matching between a gene database and a pattern of expressed RNAs, thus simultaneously identifying and quantifying virtually all genes in a sample without sequencing or hybridisation.

Gene expression profiling is currently the most powerful tool for gene discovery, drug target discovery, disease classification and for the prediction of drug response. In response to the growing demand for higher-quality, lower-cost and truly global gene expression analysis, Global Genomics delivers a PCR-based gene expression profiling method which simultaneously quantifies and identifies more than 95% of all genes in just a few hours. The method is based on an innovative application of proven molecular biology techniques, combining PCR with proprietary in-house developed computational algorithms, together achieving the Holy Grail of gene expression analysis – solid data on virtually all genes in a single experiment.

The method, Tangerine, is delivered as a research service over the web. Customers submit RNA samples and receive high-quality data with a turnaround of less than two weeks. Built-in quality control ensures consistent reliability. In addition, the web-based interface provides data management and analysis tools and the opportunity to share and purchase gene expression data.

Founded on the sensitivity and reliability of PCR gene detection, Tangerine combines a proprietary fragment display protocol with advanced in house-developed bioinformatics. By first displaying all genes in multiple pattern and then using combinatorial algorithms to simultaneously quantify and identify them, FragmentPrinting is capable of simultaneously quantifying and identifying more than 95% of all genes in a sample without sequencing or hybridisation.

Prevas Bioinformatics

Hans Fondelius, Vice President
Prevas Bioinformatics
Vallongatan 1
SE-752 28 Uppsala, Sweden
Phone: +46 18 56 27 00
E-mail: hans.fondelius@prevas.se
Web: www.prevas.se

Mission/Background

Prevas bioinformatics is a business area in Prevas AB founded January year 2000. The business idea is to increase the chance of success for companies in the biotechnology, pharmaceutical and medical technology sectors by giving new and established companies in Scandinavia an opportunity to focus on their core competence. Prevas bioinformatics offers a reliable and competent resource for business development, software development and after-sales service in the life science area. Prevas' application of its ISO certified project model has resulted in high quality and high delivery reliability, and consequently represents security for our customers. Our aim is that our customers shall regard us as an IT partner with broad experience of various bioinformatics application areas. We therefore offer employees with experience of the area, a number of whom have an education in molecular biology combined with information technology.

Technology/Methodology

Prevas bioinformatics use a large set of tools and are continuously learning to handle new tools related to specific customer needs. The basic tools are described below.

Project Model

- Prevas was the first IT consultant in Sweden to be ISO 9000 certified. The project model was part of the quality system to be certified and is still used with success.
- The Rational toolset with Rational Unified Process (RUP) and Rational Rose is an emerging *de-facto* standard for software development and has become an important complement to the existing Prevas project model.

Application Development

- The general programming languages C++, Java and Visual Basic are standard tools.
- COM and CORBA together with J2EE from SUN, provides platform independence and general functionality for technical architecture.

Data Management

- XML from the World Wide Web consortium (W3C) provides a standardised data format.
- ORACLE and MS SQL server serves many solutions developed by Prevas bioinformatics.

Product

Prevas bioinformatics offers services for business development, software development and support in the life science area. Our undertaking spans from the early discussions about what to be achieved, and it proceeds via the development project to the final maintenance and support state of the delivered solution.

Pyrosequencing AB

Lennart Beckman, Director of instrument
and SW R&D
Pyrosequencing AB
Phone: +46 18 56 59 00
E-mail:
lennart.beckman@pyrosequencing.com
Web: http://www.pyrosequencing.com

Pyrosequencing AB develops, manufactures and sells complete solutions for rapid applied genetic analysis based on its proprietary Pyrosequencing™ technology, a simple-to-use DNA sequencing technique. Pyrosequencing leads the global market in Applied Genomics with over 120 systems sold to major pharmaceutical and biotech companies and prestigious research institutions worldwide.

Pyrosequencing™ is broadly applicable for the analysis of single nucleotide polymorphisms (SNPs) and for the identification and quantification of short DNA sequences used in bacterial and viral typing.

Pyrosequencing AB formed a Molecular Diagnostics Business Unit earlier this year to establish the Company's proprietary technology as a standard platform for clinical genetic analysis. Capitalizing on Pyrosequencing's worldwide market leadership in applied genetic analysis, the Molecular Diagnostics Business Unit is pursuing a global strategy to identify new diagnostic product opportunities, develop clinically useful molecular diagnostic assays, and collaborate with academic and commercial partners in the fields of disease diagnosis, clinical prognosis and pharmacogenomics.

The Company's products include the benchtop PSQ™96 System and a high-throughput PTP™ system which utilize proprietary software and reagents. Among Pyrosequencing's customers are AstraZeneca, GlaxoSmithKline, Merck, the NIH, the Harvard Center for Cancer Prevention, the Karolinska Institute, Biogen, Oxagen, Ltd., and DuPont Agriculture.

Pyrosequencing and Bioinformatics

The Pyrosequencing product offering contain substantial elements of software from real time instrument control to basecalling and and pattern recognition. We are particularly active in the following areas of "bioinformatics":

- Algorithms for basecalling
- Pattern recognition
- Integration with database platforms (Oracle™, Access™)
- Primer/assay design software
- Integration with external systems such as robotics and LIMS
- Real time instrument control
- Image analysis

We have a strong internal team of more than 10 software specialists and we collaborate with leading external commercial and academic groups.

Virtual Genetics Laboratory

Jack Robinson
Managing Director
Phone: + 46 8 508 844 07
E-mail: jack.robinson@vglab.com
Web: http://www.vglab.com

Jack Robinson, Joakim Cöster (Director Research)
Photo: Eva Ankarvall Tolgraven

Virtual Genetics is a Swedish bioinformatics company with expert resources in the areas of Life Science, Advanced Mathematics and Information Technology. The products and services offered are focused on advanced text and data mining solutions, which can be integrated into the existing research and knowledge management processes.

Virtual Adapt

Virtual Adapt is a text mining and information retrieval system designed to handle large document-based databases such as Medline. The system addresses the main challenges that face researchers in the life science area with the following features:

Queries

- Search expressions on multiple databases
- Word truncation and wildcards
- Numerical searching within ranges
- Proximity and synonym searching (thesaurus)
- Similarity searching based on keywords, text or documents
- Relevance feedback to "learn" from previous searches
- Query history and saved queries across sessions

Results

- Ranking of results based on occurrence, frequency, date of publication.
- Highlighting of relevant text (*e.g.* protein names)
- Export of documents in XML format
- Viewing of summaries, abstracts or full text
- Extraction of keyword and key phrases

Virtual Predict

Virtual Predict is a data mining platform for the discovery of rules that describe significant patterns in data of varying complexity. In addition to providing an insight into a particular domain, the rules that are discovered can be used for predictive modeling, decision support and automatic classification of new data. The following projects have benefited from the use of Virtual Predict: Prediction of the aqueous Solubility of Molecules.

- Mutagenicity of Molecules
- Secondary protein structure prediction
- Toxicity of candidate drugs for treatment of Alzheimer's disease
- Identification of brain-specific proteins
- Analysis of micro array data

Future Development

The company's vision for the future development of software tools goes beyond bioinformatics and includes several scientific collaborations. Projects exist for the extraction of protein names and their contextual relationships from text and sequence databases. Developed tools allow the establishment of databases of documented protein interactions. A planned continuation of this collaboration will focus on advanced techniques for the description of discrete and dynamic interactions in protein networks. This type of analysis is one of the building blocks in the emerging branch of computational biology called *systems biology*.

Visual Bioinformatics

Tim Wood
Visual Bioinformatics AB
Box 700 06
Teknikringen 30, plan 7
SE-100 44 Stockholm, Sweden
Phone: +46 8 790 66 62

E-mail:
tim.wood@visual-bioinformatics.com
Web: http://www.visual-bioinformatics.com

Visual Bioinformatics is a Swedish bioinformatics company founded in 1999 by scientists from the Royal Institute of Technology and Karolinska Institutet in Stockholm. We develop integrated software solutions for managing data in the rapidly expanding field of functional genomics. In addition to competence in state-of-the-art information technologies such as Java, Oracle and XML we also provide value-added biological databases, which capitalise on our expertise within genomics, genetics and drug target discovery. Visual Bioinformatics became a subsidiary of Affibody at the end of June 2001. Affibody AB is a proteomics and biotherapy company that develops small, robust protein ligands (Affibodies) that mimic antibodies in a variety of biological and functional applications.

Visual Bioinformatics' foundation product, GeneWeaver, is a software and database solution for analysis of gene expression data in an integrated project management environment. GeneWeaver can manage data derived from the major types of experimental methods, divided into sequencing-based (EST/SAGE) and hybridisation-based (micro/macro arrays) methods. Key functionality includes:

- Management of multi-method gene expression research projects
- Comparison of publicly available data sets with locally generated data
- Mapping of expression data against different gene databases
- "Virtual Chips" for viewing data projected onto different classification schemes
- Algorithms for normalization, pair-wise comparisons, multi-sample trend finding and clustering
- Immediate access to more information through WWW-links
- Seamless and efficient creation of reports for visualization and publishing

GeneWeaver has been designed as an extendible platform capable of importing and analysing experiment data and annotation both from the user's own laboratory and from public domain databases. The Virtual Chip concept facilitates the rapid integration of experimental data with genomic annotation and reveals the relationships between such experimental data and biological knowledge as easy-to-follow visual patterns. The extendibility and intuitively flexible graphical interface of GeneWeaver means that the platform is applicable to a variety of types of genomic and proteomic data. Designing novel strategies to integrate this data into GeneWeaver is a major component of our research at Visual Bioinformatics.

Ph.D. Programme in Medical Bioinformatics

Karolinska Institutet coordinates a Ph.D. programme in medical bioinformatics. The purpose is to build up competence in bioinformatics with special emphasis on biomedical and clinical applications. The organisation of the Ph.D. programme is located at Karolinska Institutet but the studies may be pursued at any university, provided that the premises are right. The programme is open to Ph.D. students from all Sweden.

Bioinformatics is a new science at the border between computer science and biology, aiming at handling and interpreting large amounts of biological data. The need of bioinformaticians in the pharmaceutical and biotech industry is much greater than the available number of researchers. A fast expansion of the Ph.D. education is therefore necessary.

The Ph.D. programme in medical bioinformatics is unique in Sweden, combining a research education in bioinformatics with industrial and clinical collaboration. The needs of the end-users are in focus. The Ph.D. programme integrates research, companies and hospitals. Regional companies, hospitals and universities will be very important in this process.

The overall goal with the Ph.D. programme is to provide an education of high quality. It shall:

- be adapted to suit the specific project, company and Ph.D. student
- provide good supervising
- lead to good research results
- gather knowledge and competence at universities, industries, organisations etc.
- lead to a career in the industry, health care or academia

The Ph.D. student will be financed to fifty percent by the Knowledge Foundation (KK-stiftelsen) and to fifty percent by the participating company or hospital. A Ph.D. student will cost the company or hospital maximally 400 000 SEK per year. The Knowledge Foundation will contribute totally with 48 million SEK.

The Ph.D. student will continuously provide the company or hospital with competence. Courses within the programme will be open also to co-workers from the participating company. A part of the research project can be performed at the company/hospital, which will be beneficial both in a short-term and a long-term perspective. The participating companies will also increase their contacts with the academic world.

Bioinformatics Centres in the Stockholm–Uppsala Region

Center for Genomics and Bioinformatics
http://www.cgb.ki.se

Stockholm Bioinformatics Center
http://www.sbc.su.se

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

For information about the Ph.D. Programme in Medical Bioinformatics, please contact:

Lena Lewin
Centre for Medical Innovations
Nobels väg 15a
Karolinska Institutet
SE-171 77 Stockholm, Sweden
Phone: +46 8 728 60 64
E-mail: lena.lewin@cmi.ki.se
http://www.cbb.ki.se/fmb

Bengt Persson
Centre for Medical Innovations
Nobels väg 15a
Karolinska Institutet
SE-171 77 Stockholm, Sweden
Phone: +46 8 728 77 30
E-mail: bengt.persson@cmi.ki.se

KONGL. CAROLINSKA MEDICO CHIRURGISKA INSTITUTET

Karolinska Institutet
A medical university dedicated to improving people's health through higher education, research and information.

This guide is produced by Centre for Medical Innovations at Karolinska Institutet, March 2002.