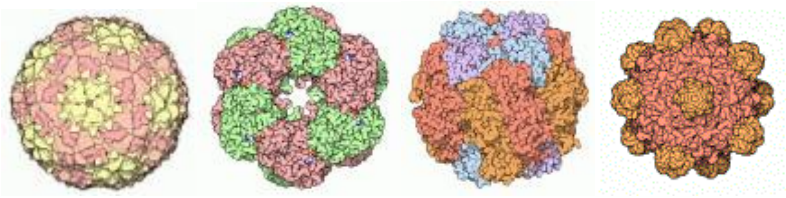
COMPUTATIONAL BIOLOGY



Seminar 02471
SCHLOSS DAGSTUHL
NOVEMBER 17 – 22, 2002

ORGANISED BY:

Russ Altman (Stanford Univ. USA),

David Gilbert (Univ. of Glasgow, GB),

Thomas Lengauer (MPI Saarbrücken, D)

Report compiled by Aik Choon Tan (Univ. of Glasgow, GB)

Preface

This seminar was the fourth seminar on general issues in Computational Biology that was held at Dagstuhl. Three previous seminars on this topic have been held in 1992, 1995 and 2000. In addition, there have been three seminars that concentrated on Metabolic Pathways.

The seminar aimed at exploring traditional as well as some more novel issues in computational biology. The field has expanded greatly in the past years, and the danger has grown of splitting the field into more and more separate sub-disciplines. This seminar attempted to slow down this trend by giving all attendees an overview of the state of the art in widely differing sub-areas of computational biology. These included haplotype analysis, sequence analysis, structure analysis, docking, analysis of expression data and biochemical networks as well as issues in medical applications and software issues in project design.

The days were filled with lectures that had extended discussion periods. Some of the talks had decidedly tutorial character. Early afternoons were set aside for informal discussions. There were evening discussion sessions on Biochemical Pathways (Monday, opened by a statement by David Gilbert), and Bioinformatics and Disease (Tuesday, opened by a statement by Thomas Werner). It was a common sentiment that the broad scope of the seminar is worthwhile and should be maintained in future seminars.

Here is a collection of some of the comments on the seminar by participants:

> "Thanks again for organizing the Dagstuhl workshop. I enjoyed it more than many other conferences I attended."
>
> "I could have a wonderful time at Dagstuhl. Thank you very much for all you have done for this meeting."
>
> "Thanks for organizing such a stimulating meeting."

.. and even from those who didn't attend:

> "Heard Dagstuhl was good"

Program

Monday, November 18, 2002

Session: Networks **Chair: David Gilbert**

David Gilbert and Thomas Lengauer
Welcome and Introductory Remarks

Hanspeter Herzel
Modeling RAS-Singaling Cascades and Gene Regulation

Satoru Miyano
Inferring, Modeling and Simulating Biopathways

Ron Shamir
Analysis of Gene Networks

Benno Schwikowski
Algorithms That Support Manual and Automatic Integration of Biological Networks

Marie-France Sagot
Motif Inference

Gene Myers
Enabling Systems Biology

Session: Regulatory Sequences **Chair: Thomas Lengauer**

Mark Craven
Exploiting Relations in Learning to Identity Regulatory Elements

Patricia Evans
Methods for Finding DNA Motifs

Thomas Werner
Genomic versus Functional Context: Complementary Approaches to Elucidation of Gene Function

Evening Session: Networks **Chair: David Gilbert**

Tuesday, November 19, 2002

Session: Protein Structure **Chair: Sorin Istrail**

Kevin Karplus
Unifying Secondary Structure, Fold-Recognition, and New-Fold Methods for Predicting Protein Structure

Ingolf Sommer
Automation of Protein Structure Prediction

Mario Albrecht
A Case Study on Protein Structure and Function Predictions

Rolf Backofen
Simplified Models of Proteins – How far can we go?

Session: Genetic Variations **Chair: Søren Brunak**

Sorin Istrail
Algorithms for SNPs and Haplotype Inference

Esko Ukkonen
Finding Haplotype Blocks and Founder Sequences

C. David Page
Supervised Learning from SNP Data

Session: Protein Function **Chair: Thomas Lengauer**

Søren Brunak
Protein Function Prediction and Conservation in Feature Space

Joachim Selbig
Elucidation of HIV-1 Drug Resistance

Niko Beerenwinkel
Methods for Optimizing Antiviral Therapies

Evening Session: Bioinformatics and Disease
Chair: Thomas Werner

Wednesday, November 20, 2002

Session: Sequences **Chair: Thomas Werner**

Des Higgins
Boring New Methods for Multiple Sequence Alignment

Ela Pustulka-Hunt
Using Database Technologies to Accelerate Sequence Comparison

Jens Stoye
Algorithms for High-Level Comparative Genomics

Session: Protein Structure **Chair: Thomas Lengauer**

Aik Choon Tan
Integrative Machine Learning of Patterns from Heterogeneous Data Sources

Tal Pupko
Identification of Binding Sites

Thursday, November 21, 2002

Session: Protein Structure and Interactions
Chair: Des Higgins

Juris Viksna
Algorithms for Protein Structure Matching

Francisco Silva Domingues
From Protein Sequence to Function

Irit Gat-Viks
Chain Functions and Scoring Functions in Genetic Networks

Session: Projects **Chair: Thomas Lengauer**

Ross Donald King
The Robot Scientist Project

Mar Albà
Bioinformatics in the Study of Virus-Host Systems

David Gilbert
Modelling Biochemical Pathways – From Network Topology to Dynamics

Session: Expression Arrays **Chair: Thomas Lengauer**

Nir Friedman
Learning About Gene Regulation from Sequences and Expression Data

Dana Pe'er
On Inferring Regulation from Gene Expression Profiles

Alexander Schliep
Dealing with Non-Unique Probes: DNA Chips and Group Testing

Florian Sohler
Analysis of Gene Expression Data on Networks

Lev A. Soinov
Reconstruction of Gene Networks from Expression Data

Friday, November 22, 2002

Session: Docking **Chair: Thomas Lengauer**

Hans-Peter Lenhof
Protein-protein Docking

Alexander Schliep
Time Series Analysis with HMMs

Thomas Lengauer and David Gilbert
Concluding Remarks

A Case Study on Protein Structure and Function Prediction: The Machado-Joseph Disease Gene Product Ataxin-3

Mario Albrecht
Max Planck Institute for Informatics, Saarbrücken

Spinocerebellar ataxia type 3 (SCA3) is a polyglutamine disorder caused by a CAG repeat expansion in the coding region of a gene encoding the protein ataxin-3. We performed a comprehensive computational analysis of carefully selected homologous proteins in order to propose a structural model and structure-based functions for ataxin-3. We describe some of our findings and verifying lab experiments. Since important work of our predictive strategy could be accomplished solely manually, we further outline necessary improvements of automatic structure prediction methods.

Reference:
M. Albrecht, D. Hoffmann, B. O. Evert, I. Schmitt, U. Wüllner, T. Lengauer:
Structural modeling of ataxin-3 reveals distant homology to adaptins.
To appear in Proteins: Structure, Function, and Genetics.

Bioinformatics in the Study of Virus-Host Systems

Mar Albà
Universitat Pompeu Fabra, Barcelona, Spain.

Many virus genomes have already been sequenced and the numbers are increasing at an exponential rate. However much of the information in a genome sequence remains hidden, difficult to interpret or "decode". This can be aided by the use of comparative sequence analysis techniques. We have developed a protocol to cluster sequences on the basis of sequence similarity and created a database of virus sequences, the VIrus DAtabase VIDA (1). The database contains a collection of homologous protein families (HPFs), each characterised by one or more regions of sequence conservation. VIDA includes all sequences available in GenBank from different virus families, in particular from the Herpesviridae, Poxviridae, Coronaviridae, Arterirividae and Papillomaviridae. We have used the HPFs to gain a better understanding on the evolutionary relationships within the Herpesviridae, using the fraction of shared genes to derive distances between pairs of genomes (2). A second application of the HPFs has been to the identification of gene transfer between host and herpesvirus genomes (3). In particular we have used the conserved regions in the HPFs to performed sensitive, profile-based, sequence similarity searches against the products of the human genome. This has lead to the prediction of new herpesvirus and host gene functions.

References:
(1) Albà, M.M., Lee, D., Pearl, F.M.G., Shepherd, A.J., Martin, N., Orengo, C.A., Kellam, P. (2001) VIDA: A virus database system for the organization of virus genome open reading frames. Nucleic Acids Research, Vol. 29: 133-136.
(2)Albà, M.M., Das, R., Orengo, C., Kellam, P. (2001) Genome wide function conservation and phylogeny in the Herpesviridae. Genome Research, Vol. 11: 43-54.
(3)Holzerlandt, R., Orengo, C., Kellam, P., Albà, M.M. (2002). Identification of new herpesvirus gene homologues in the human genome. Genome Research, Vol. 12:1739-1748.

Simplified Models of Proteins: How far can we go?

Rolf Backofen,
Lehrstuhl für Bioinformatik
FSU Jena

In this talk, we introduce simplified models of proteins, which are used in hierarchical approaches to protein structure prediction. The predominant class of simplified models are lattice models, where the positions of amino acids are restricted to positions in a regular lattice. Structure prediction in simplified models is still a hard (to be more precise: NP-hard) problem.

We give an overview of previous approaches to structure prediction in lattice models, which are currently only able to fold sequences of short size (< 88) and/or in lattices that are not well suited to model proteins (e.g. cubic lattice).

We report on improvements on this problem. Using constraint-based techniques, we are able to improve considerable in sequences size (i.e. fold sequences up to length 180), and in a lattice model that allows a much better modelling of real protein structures (the face-centered-cubic lattice). Thus, the constraint-based method handles a search space that is 4.5^{90} times bigger than the search space modelled by previous methods.

Methods for Optimizing Antiviral Combination Therapies

Niko Beerenwinkel
Max Planck Institute for Informatics, Saarbrücken

Therapeutic success of antiretroviral combination therapies in infected patients is limited by the evolution of drug-resistant viral variants. We present methods for finding drug combinations that are estimated to be maximally active against a given viral strain. Based on viral genomic data we construct a scoring function by predicting phenotypic drug resistance from genotypes for each drug and integrating normalized predictions into a score for drug combinations. The scoring function in shown to correlate with observed clinical response in HIV-1 infected patients.

In order to estimate activity on nearby escape mutants we also score the mutational neighborhood of the considered strain. A tree-like model of evolution of drug resistance is proposed together with an efficient reconstruction algorithm. The model is shown to capture and extend knowledge about viral evolution under drug pressure.

Protein Function Prediction and Conservation in Feature Space

Søren Brunak
Center for Biological Sequence Analysis, Technical University of Denmark

In-dept modeling and understanding of biological systems requires knowledge of the functional role of their subcomponents, in particular the function of proteins. An integrated computational approach is needed to face the challenge of the functional assignment of thousands of new gene products derived from different sequencing projects. Standard functional assignment by homology using proteins of known function is very powerful, but still leaves unassigned proteins belonging to families without known function (orphan families), or isolated sequences (orphan sequences). The number of orphan families and sequences increases over time since experimental functional analysis is highly demanding in time and effort.

Function is a multilevel, complex phenomenon, where different levels are interwoven (chemical, biochemical, cellular, organismal and developmental). We present an indirect approach where predicted structural features, putative posttranslational modifications (glycosylation, phosphorylation), sorting signals (signal peptides, transit peptides), and calculated features (chain length, amino acid composition, isoelectric point, hydrophobicity) are integrated and used to infer the functional class, enzyme categories, and other categories defined by the gene ontology consortium.

The approach presented here predicts functional role categories in the "feature" space of the proteome, rather than using the "sequence" space of the genome. As an example, we present results from a protein feature analysis of the yeast cell cycle.

References:
[1] Prediction of human protein function from post-translational modifications and localization features, L. J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H. H. Stærfeldt, K. Rapacki, C. Workman, C. A. F. Andersen,
S. Knudsen, A. Krogh, A. Valencia, S. Brunak, J. Mol. Biol., 319, 1257-1265, 2002.
[2] Prediction of human protein function according to Gene Ontology categories L.J. Jensen, R. Gupta, H.-H. Stærfeldt, and S. Brunak, to appear in Bioinformatics, 2002.

Exploiting Relations in Learning to Identify Regulatory Elements

Mark Cravens
Dept. of Biostatistics & Medical Informatics, University of Wisconsin

We have been developing models for predictively identifying various regulatory elements in bacterial genomes. This talk starts by describing the models that we have learned from known instances of two types of regulatory elements. We have learned a Bayesian network that uses gene coordinates, codon usage patterns, and gene expression data to identify operons in bacterial genomes. To recognize transcription termination signals, we have a trained a stochastic context free grammar. In the second part of the talk I describe how we can effectively increase the training set sizes for such models by taking into account known relationships among regulatory elements. For example, we know that a terminator will be found shortly downstream from the last gene in an operon. Therefore, we can use known and predicted ends of operons to assemble sets of training sequences that with high probability contain terminators. We refer to such sequences as "weakly" labeled training examples. Typically such examples contain more hidden state than ordinary examples. For example, in the terminator case, we won't know the locations or the types of terminators in these weakly labeled sequences. However, our experiments show that when we have relatively small sets of ordinary labeled training examples, there can be significant value in training with these weakly labeled sequences as well. This result suggests that we can boost the accuracy of our predictions for bacterial genomes for which labeled examples are scarce.

From Protein Sequence to Function

Francisco S. Domingues
Max-Planck-Institut für Informatik

Here we review the different computational approaches available for protein function prediction: homology based methods, genomic context, sequence features and structure based methods. Homology based methods have been widely used for many years, they represent the current standard approach to electronically annotate protein sequences. Identification of homology does not directly imply identification of function. Conservation of function between two proteins has been related to their percentage of sequence identity. A platform for automated annotation of protein sequences using signature databases is described (WILMA). In order to estimate the reliability of the annotations based on signature

database searches, we used WILMA to compare electronically and human curated annotations as obtained from the GOA project. We find that for 70% of the annotated sequences, the functional terms from curated and electronic annotations match. The match is more extensive at the level of molecular function and cellular component than for biological process. More recently, genomic context methods have been developed that allow the identification of proteins that interact or that participate in a common biological process. Another approach to identify the possible biological role of a protein uses sequence features that can be calculated or predicted from the target sequence alone, without relying on external databases. Finally, there are several methods to predict function based on predicted or experimental structural models. Structure comparison is an important tool to identify remote homology, but as mentioned before, homology not always corresponds to similar function. Structural motifs allow the identification of enzyme active sites. Functional sites can be predicted based on the identification of conserved regions that cluster in the protein surface.

Methods for Finding DNA Motifs

Patricia Evans
Faculty of Computer Science, University of New Brunswick

Exhaustive motif enumeration is generally overlooked in favour of faster heuristics that can miss weaker motifs, only report single motifs, or have difficulty determining that no motif is present. Exhaustive methods are very time and space consuming; however, appropriate exploitation of problem restrictions and efficient design can reduce the resource consumption sufficiently and make enumerative methods useable for many reasonable motif sizes. Based on finding common approximate substrings as motifs, we developed two useable exhaustive methods: a tabulation scheme that uses bitstring pairs as indices to produce complete substring presence information for short substrings with any number of errors, and a progressive clique finder that uses lazy graph construction to find longer substrings with fewer errors. Both techniques work for long input sequences, far greater than the shorter lengths needed by many current heuristics. The complete information produced by these exhaustive methods can be used to examine the data used to test heuristics, and reveals the source of heuristics' difficulties with random data.

Chain Functions and Scoring Functions in Genetic Networks

Irit Gat-viks
School of Computer Science, Tel Aviv University

One of the grand challenges of system biology is to reconstruct the network of regulatory control among genes and proteins. High throughput data, particularly from expression experiments, may make this possible in the future. Here we address two key ingredients in any such 'reverse engineering' effort: The choice of biologically relevant, yet restricted, set of potential regulation functions, and the scoring function by which one should evaluate candidate network solutions.

We propose a set of regulation functions which we call chain functions, and argue for their ubiquity in biological networks. We analyze their complexity and show that their number is exponentially smaller than all boolean functions of the same dimension. By limiting network inference to the chain functions, not only do we reduce the size of the search space of possible models, but we also reduce the number of incorrect regulation functions which are scored above the correct solution.

We define a new scoring method which evaluates how well a certain regulation function fits experimental data. We devise two scoring functions, which utilize established statistical methods. The first evaluates the specificity of a group of regulators to the regulated gene. The second function evaluates how well a particular regulation function (for a group of regulators and their regulated gene) fits the data. Both functions are expressed as p-values, and thus

are not very sensitive to over-fitting. Moreover, due to the Gaussian shape of these scoring functions, they always score only few solutions at the high end.

We apply our approach to transcription profiles of the yeast galactose pathway (Ideker et al. 2001). First, we demonstrate the advantages of using the chain functions instead of searching through all boolean functions. Second, we compare several scoring functions previously proposed for network inference, and show that our method outperforms these functions. Third, we show that by using in combination our two scoring functions, we can obtain very high ranking of the correct solution. We expect both chain functions and our function scores to be helpful in future attempts to infer regulatory networks.

Modelling Biochemical Pathways - From Network Topology to Dynamic Behaviour

David Gilbert
Bioinformatics Research Centre, Department of Computing Science, University of Glasgow

Biochemical networks can be of various types, e.g. metabolic, regulatory, signal transduction etc., and several databases exist holding data on these. In general, models of these networks can be based on systems of differential equations, or else based on the structure of networks - e.g. as graphs. Previously we have worked with structural models with the aim of giving an insight as to how connectivities in the graph and biological behaviour correlate [1]. 'Topological' approaches to network structure represent networks as graphs and permit operations such as path searching over the graphs, often with constraints, as well as pathway reconstruction [2,3]. Bipartite graphs are commonly used to avoid ambiguities in representation, and these can be extended to object-oriented models.

More recently we have become interested in modelling the dynamic properties of biochemical networks as concurrent systems, and are starting a large UK government funded project on this. The aim of this interdisciplinary project is to model diverse biochemical networks and develop an associated computational system to facilitate the analysis of the behaviour of these networks. In contrast to previous modelling approaches we will have a continuous cross check between modelling and real experimental data. As paradigms we will use the (i) regulation of the MAPK network and (ii) apoptosis (programmed cell death). Both networks are at the focus of current drug discovery efforts in important disease areas including cancer, arteriosclerosis, stroke, heart disease, chronic inflammatory and degenerative diseases. Specifically, for signal transduction networks we will investigate the role of threshold effects, enzyme processing, and positive and negative regulatory (feedback) mechanisms of signal propagation in the ERK (MAPK) pathway. The model will be informed by in vitro experiments using core component proteins of the ERK pathway in order to obtain behavioural parameters. Data on apoptosis will be mainly recruited from the literature using a text-mining tool.

We will devise a model of interacting biological processes, within a concurrent calculus setting, incorporating temporal aspects and a type schema permitting quantitative modelling via stochastic processes. This approach will be based on our previous work on process algebra, concurrency theory, temporal reasoning and simulator construction. We will build an analytical simulator, based on the systems we have previously developed with the ability to explore the dynamic characteristics of signal transduction networks, as well as biochemical process equivalences. We will closely couple and refine the design of the theoretical model with to biochemical and biological observations; and we plan to correlate the model with biochemical and biological endpoints, i.e. with patterns of substrate phosphorylation changing during differentiation.

Our approach will permit the following:

- Network topology reconstruction: infer network connectivity (topology) and behaviour of individual components from observations of biochemical behaviour over time; to be verified by assays
- Prediction of the effect of drugs/mutations on a known biological system (topology plus component behaviour)
- Prediction of the overall observable external behaviour of the biological system
- Analysis of the 'internal' behaviour of the system, for example routes and fluxes etc.
- Investigation of the effect of reconfiguring pathways of selecting alternative pathways, which will be manifested in different fluxes although these may result in the same overall observable behaviour of the network

This project is joint with Walter Kőlch, a wet lab scientist, and Muffy Calder, a concurrency specialist, both of the University of Glasgow.

Work on structural aspects of networks is joint with Yves Deville of the Catholic University of Louvain.

References:
[1] Jacques van Helden, Avi Naim, Renato Mancuso, Matthew Eldridge, Lorenz Wernisch, David Gilbert, and Shoshana J. Wodak, Representing and analysing molecular and cellular function in the computer, Journal of Biological Chemistry, Journal of Biological Chemistry, 381 (9-10):921-35, Sep-Oct 2000.
[2] Jacques van Helden, David Gilbert, Lorenz Wernisch, Michael Schroeder, and Shoshana Wodak, Application of Regulatory Sequence Analysis and Metabolic Network Analysis to the Interpretation of Gene Expression Data, in Computational Biology (Olivier Gascuel and Marie-France Sagot, Eds), LNCS 2006, pp147-163, ISBN 3-540-42242-0, 2001.
[3] Jacques van Helden, Lorenz Wernisch, David Gilbert, and Shoshana Wodak. "Graph-based analysis of metabolic networks". in Ernst Schering Research Foundation Workshop Volume 38: Bioinformatics and Genome Analysis. Editors: H.-W. Mewes, B. Weiss, H. Seidel Springer-Verlag, ISBN 3-540-42893-3, pp245-274, Berlin Heidelberg 2002.

Modeling RAS-Signaling Cascades and Gene Regulation

H. Herzel, N. Bluethgen, J. Wolf, Sz. Kielbasa
Institute for Theoretical Biology, Humboldt-University Berlin

C. Sers, R. Schaefer
Universitätsklinikum Charité, Laboratory of Molecular Tumor Pathology

A. O. Schmitt, Epigenomics
J. Walter, Genetics, University Saarbruecken

J. Schuchhardt, D. Beule, Microdiscovery GmbH

The starting point of our project was the genome-wide identification of Ras target genes using PCR-based subtractive suppression hybridization (SSH) [1]. More than 700 genes were found to be differentially regulated in normal fibroblasts compared to cells transformed due to mutated Ras proteins. Genes which are upregulated in transformed cells are potential oncogenes whereas downregulated genes are candidates for tumor suppressors.

The regulation of the genes in cell lines with inducible Ras mutants and in the presence of kinase inhibitors are further analyzed using cDNA arrays on nylon membranes [2] and glass chips (AGILENT and AFFYMETRIX). Consequently, different technologies of expression profiling can be compared quantitatively for a specific set of genes under well-defined conditions.

In order to explore the Ras-dependent response of the cells we model relevant signaling pathways, analyze promoters of coregulated genes, and study the role of DNA methylation

and histone deacetylation. Using inhibitors of MAPK it was shown that a subset of Ras-regulated genes is controlled via the MAPK-cascade. Model simulations of this cascade show that this signaling module can act as an amplifier, switch or feedback controller [3]. We study the robustness of switch-like behaviour with respect to parameter variations [4].

Signaling cascades activate certain transcription factors such as Fos, Elk1 or ATF2. In order to detect transcription factor binding sites we search for promoters of coregulated Ras-targets. We combine data bank information (e.g. DBTSS) with in silico promoter search (PromoterInspector, McPromoter, FirstEF, CONPRO) to get reliable predictions of core promoters. In these regions we identify binding sites of known transcription factors and predict novel motifs, as exemplified for metabolism-related genes in yeast [2,5].

It is known that oncogenes and tumor suppressors can be up- or downregulated via DNA methylation and histone deacetylation. In order to find these genes we measure expression profiles after inhibiting DNA methylation and histone deacetylation by 5-Aza-CdR and TSA. A considerable number of genes are differentially expressed after treatment. Potential differences in the DNA-methylation patterns of the promoter regions are investigated by bisulfit treatment.

In summary, the combination of comprehensive expression profiling, bioinformatic analyses of promoter regions, DNA methylation studies, and mathematical modeling of specific modules will increase our understanding of Ras-signaling and gene regulation in normal and tumorgenic cells.

References:
[1] Zuber, J., Tchernitsa, O.I., Hinzmann, B., Schmitz, A.C., Grips, M., Hellriegel, M., Sers, C., Rosenthal, A,. Schäfer, R.: A genome-wide survey of RAS transformation targets. Nat. Genet. 2000, 24: 144-152.
[2] Herzel, H., Beule, D., Kielbasa, S., Korbel, J., Sers, C., Malik, A., Eickhoff, H., Lehrach, H., Schuchhardt, J.: Extracting Information from cDNA Arrays. CHAOS 2001, 11, 98-107.
[3] Blüthgen, N., Herzel, H.: MAP-Kinase-Cascade: Switch, Amplifier or Feedback Controller? Computation of Biochemical Pathways and Genetic Networks. Logos Verlag, Berlin, 2001, pp. 55-62
[4] Blüthgen, N. Herzel, H.: How robust are switches in intracellular signaling cascades? submitted.
[5] Kielbasa, S.M., Korbel, J.O., Beule, D., Schuchhardt, J., Herzel, H.: Combining frequency and positional information to predict transcription factor binding sites. Bioinformatics 2001, 17: 1019-1026.

Some New Methods for Multiple Alignment

Des Higgins
Department of Biochemistry, University College, Cork, Ireland

Since the late 1980s, multiple alignments have been carried out mainly using the fast heuristic method of “Progressive Alignment”, most commonly using the Clustal package. Clustal is richly parameterised regarding the details of protein alignment but is hard to control or optimise in any general sense and does suffer from a local minimum problem. This is especially apparent with long insertions and deletions or with repeated sequences.

The main systematic way of finding the best multiple alignment is to take an objective function (OF) and optimise it. The usual OF is the weighted sums of pairs. A number of programs can give good solutions in practice using this OF, including SAGA (genetic algorithm: GA), PRRP (iteration), DCA (divide and conquer) but only for fairly small examples. Of these PRRP delivers especially good results and works with realistically large size problems. We used SAGA to explore alternative OFs such as the maximum weight trace (MWT: John Kececioglu). The MWT is especially attractive because of the way it allows us to mix together different types of alignment information. Using the GA in SAGA we could find very high

quality solutions to the MWT that scored very highly when compared to expert alignments of known structures. The disadvantage of the GA was that it was slow for large numbers of sequences.

Finally we developed a heuristic approximation to the MWT alignment problem that delivers very high quality alignments quickly and simply. This is implemented in the T-Coffee program (originated and developed by Cedric Notredame). T-Coffee carries out progressive alignment using weighted pairs of residues as input. Each pair is an amino acid from one sequence, aligned with an amino acid from a second and a weight. These pairs can come from local and global alignments or structure superpositions or existing multiple alignments or a mixture. The resulting alignments are more accurate than those from any other software we have tested.

Using Database Technologies to Accelerate Sequence Comparison

Ela Hunt
Department of Computing Science, University of Glasgow

Comparison of mammalian genomes is a challenging task. Quality comparative gene maps for the rat, the mouse and the human are necessary to interpret the data from experimental models of heart disease, diabetis, and many other inherited diseases where the pattern of gene interaction is not clear. In particular, the interpretation of microarray data sets cannot be performed without comparative gene maps.

Currently available maps are still of low resolution, and they can only be constructed via extensive sequence comparison. To compare mammalian genomes, we need to perform BLAST on three datasets, each comprising around 3GB of DNA sequence. This can only be done with the help of large computing clusters.

We aim to produce faster, and qualitatively better sequence analysis tools. We want to perform exhaustive searching, and speed it up by using an index to the DNA sequence. We experiment with the suffix tree data structure. We demonstrated that suffix trees can now be built for data of any size, using our new partitioned tree building algorithm. We showed that this algorithm performs in practice as well as the optimal linear time algorithm (Hunt, Atkinson and Irving, VLDB Conference 01).

Our next step was to run the Smith-Waterman algorithm (with a unit cost similarity matrix) on the persistent suffix tree. We demonstrated that by using a suffix tree index we reduce the size of the DP matrix to a small percentage of the whole (Hunt, Atkinson and Irving, VLDB Journal 2002).

Our further work will examine alternative index data structures, arbitrary cost matrices, and hardware optimisations exploiting indexing.

Algorithms for SNPs and Haplotype Inference: Don't Block Out Information

Sorin Istrail
Celera/Applied Biosystems

Joint work with Bjarni Halldorsson, Vineet Bafna, Russell Schwartz & AndyClark

It is widely hoped that variation in the human genome will provide a means of predicting risk of a variety of complex, chronic diseases. A major stumbling block to the successful identification of association between human DNA polymorphisms (SNPs) and variability in risk of complex diseases is the enormous number of SNPs in the human genome. The large

number of SNPs results in unacceptably high costs for exhaustive genotyping, and so there is a broad effort to determine ways to select SNPs so as to maximize the informativeness of a subset.

We will present an overview of the algorithmic issues in haplotype reconstruction and mapping. We will also present and compare two methods for reducing the complexity of SNP variation: haplotype tagging, i.e. typing a subset of SNPs to identify segments of the genome that appear to be nearly unrecombined (haplotype blocks), and a new block-free model.

We will present a statistic for comparing haplotype blocks and show that while the concept of haplotype blocks is reasonably robust there is substantial variability among block partitions.

We will also discuss a measure for selecting an informative subset of SNPs in a block free model. We show that the general version of this problem is NP-hard and give efficient algorithms for two important special cases of this problem.

Unifying Secondary-Structure, Fold-Recognition, and New-Fold Methods for Protein Structure Prediction

Kevin Karplus, Rachel Karchin, Richard Hughey
Dept. of Computer Engineering, University of California at Santa Clara

We have recently implemented a fragment-packing program that allows us to combine information from several different techniques.

We use iterated search with hidden Markov models (HMMs) to make a multiple alignment of probable homologs of the target sequence.

We use the multiple alignment as input to a neural net to make predictions of secondary structure (or other local structural properties). The predictions are in the form of probability vectors over the local structure alphabet for each position of the target.

We create multi-track HMMs that have emission tables for amino acids (from the multiple alignments) and secondary structure (from the neural net). The multi-track HMMs are used to score every template in the template library (currently over 7000 structures). The local structure track substantially improves performance over amino-acid-only HMMs.

We combine scores for multi-track target HMMs using different local structure alphabets with amino-acid-only template HMMs scoring the target sequence. The combined scores are used to select templates for alignment.

We generate target-template alignments for the top hits using several different alignment parameter settings.

Using a new program in the SAM tool suite, fragfinder, we search the template library for the top 10 or so short gapless alignments (fragments) at each position in the target sequence.

Undertaker, which tries to optimize burial, takes in all the fold-recognition alignments, the fragfinder fragments, and a large generic library of very short fragments, then uses a genetic algorithm to generate conformations of the target sequence. The method is similar to Baker's Rosetta program in concept, but uses an all-heavy-atom representation of the conformation, and allows inserting full alignments, and not just contiguous fragments.

Our score function for this generate-and-test method is not yet tuned or tested, but we used it for CASP5 this summer anyway. The results were interesting, but we do not yet know whether they were right.

The Robot Scientist Project

Ross King
Computer Science, University of Wales

Biological science is being revolutionised by a host of remarkable new technologies: microarrays, proteomics, metabolomics, etc. The combination of increased automation, with new technologies, is producing a flood of data replete with undiscovered biological and medical knowledge. The extraction of this knowledge is becoming the largest bottleneck in the scientific process. We describe a novel route towards removing this bottleneck based on integrating data generation with intelligent data analysis - the Robot Scientist. We have physically implemented the Robot Scientist and applied it to functional genomics.

The problem domain is the use of auxotrophoic growth experiments to determine the function of knockout mutants in S. cerevisiae. The Robot Scientist: uses abductive logic programming to form an initial hypotheses set, devises near optimal experiments to select between competing hypotheses, directs a robot to perform these experiments, automatically analyse the experimental results, revises its hypothesis set in the light of the experimental results, and then repeats the cycle until the user's criteria for selection of the best hypothesis are met or resource limits are met. We have selected as a model system the aromatic amino-acid synthesis pathway. For this pathway we have access to 9 metabolites and 15 knockout mutants. We have by-hand carried out multiple times all possible single and double metabolite experiments for this domain. From the literature and the results of the single metabolite experiments we have developed a logical model which explains all but ~2% of the double metabolite experiments. To test our "intelligent" experiment selection procedure we are comparing it with the naive approach of choosing the cheapest experiment, and a random selection procedure. In simulation the intelligent procedure can achieves the same accuracy as random selection for lower cost, and is faster than the naive approach. The Robot Scientist is currently physically repeating these simulation experiments. Two rounds of repeats have been completed and the results are consistent with the simulation results.

Protein-Protein Docking

Hans-Peter Lenhof
Zentrum für Bioinformatik, Universität des Saarlandes

In this talk I will give an overview on our protein-protein docking project. New approaches for realizing protein fexibility, for filtering using a NMR scoring function, and for predicting the solvation free energy of protein complexes will be presented and experimental results obtained with the new techniques will be discussed. The use of non-local electrostatics for solvation free energy computations is the main focus of the talk. The accurate prediction of solvation free energies of molecules in water, especially of large molecules like proteins, is still a largely unsolved problem, which is mainly due to the complex nature of the water solute interactions. We have developed a scheme for the determination of the electrostatic contribution to the solvation free energy that is based on non-local elctrostatics. The new approach has been tested on simple ions, small molecules, and on proteins and quantitative comparisons with the former standard approach have been carried out.

Inferring, Modeling and Simulating Biopathways

Satoru Miyano
Human Genome Center, Institute of Medical Science, University of Tokyo

In the post-genome era, biopathway information processing will be one of the most important issues in Bioinformatics.

One is to infer the relations between genes from cDNA microarray data obtained by various perturbations such as gene disruptions, shocks, etc. We have developed a new method for inferring a network of causal relations between genes from cDNA microarray gene expression data by using Bayesian networks [1, 2]. We employed nonparametric regression for capturing nonlinear relationships between genes and derive a new criterion called BNRC (Bayesian Network and Nonlinear Regression) for choosing the network in general situations. Theoretically, our proposed theory and methodology include previous methods based on Bayes approach [3]. We also extended our method to (1) Bayesian network and nonpaametric heteroscedastic regression and (2) dynamic Bayesian network and nonparametric regression for time series gene expression data. We applied the proposed methods to the *S. cerevisiae* cell cycle data and cDNA microarray data of 120 disruptants (mostly transcription factors). The results showed us that we can infer relations between genes as networks very effectively.

The other approach to this issue is our development of Genomic Object Net [4]. This software aims at describing and simulating structurally complex dynamic causal interactions and processes such as metabolic pathways, signal transduction cascades, gene regulations. We will released Genomic Object Net (ver. 1.0) in 2002. With this system, we have shown that we can reorganize and represent various biopathway information so that biopathways can be modeled and simulated for new hypothesis generation and testing (see [5, 6, 7]). As its basic architecture, Genomic Object Net employs the notion of hybrid functional Petri net (HFPN) that is a newly defined notion obtained by extending hybrid Petri net [7] and hybrid object net [8] so that various aspects in biopathways can be smoothly modeled while inheriting good traditions from the research on Petri net. In HFPN, hybrid system of continuous and discrete events with functional enhancement for transitions is very suited for modeling various interactions and reactions, and the hierarchization of objects provides diversity in intuitive creation of complex objects. Although Petri net has been studied independently of biology, its affinity to biopathways is surprisingly good. Such enhancement to Petri net changed it to an excellent architecture for biopathway modeling and simulation.

References

[1] Imoto, S., Goto, T. & Miyano, S. (2002). Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. Proc. Pacific Symposium on Biocomputing 7, 175-186.
[2] Imoto, S., S.-Y. Kim, Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., & Miyano, S. (2002). Proc. IEEE Computer Society Bioinformatics Conference, IEEE Computer Society Press, 219-227.
[3] Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000). Using Bayesian Network to Analyze Expression Data. J. Comp. Biol., 7 601-620.
[4] http://www.GenomicObject.Net/
[5] Matsuno, H., Doi, A., Nagasaki, M., Miyano, S. (2000). Hybrid Petri net representation of gene regulatory network. Pacific Symposium on Biocomputing 5, 338-349.
[6] Matsuno, H., Doi, A., Hirata, H., Miyano, S. (2001). XML documentation of biopathways and their simulations in Genomic Object Net. Genome Informatics 12, 54-62.
[7] Alla, H., David, R. (1998). Continuous and hybrid Petri nets. J. Circuits, Systems, and Computers 8 (1), 159-188.
[8] Drath, R. (1998). Hybrid object nets: An object oriented concept for modeling complex hybrid systems. Proc. Hybrid Dynamical Systems, 3rd International Conference on Automation of Mixed Processes, ADPM'98, 437-442.

Systematic Genomics

Gene Myers
University of California at Berkeley

This talk is primarily a position piece designed to provoke deeper thinking on where genomics should be headed. A survey of current techniques reveal that while we can accurately sequence, we cannot accurately annotate, we can measure RNA expression only crudely,

and we can observe only abundant proteins. Biological discovery has been accelerated by genomics, but it has not fundametally changed from the single hypothesis paradigm into a discovery science. We propose a program of whole genome sequencing of 10 or more species around a target genome followed by comparative informatics to obtain gene annotations of an accuracy sufficient to enable the systematic empirical verification and refinement of every prediction by full-length cDNA sampling, probed cDNA selection, and RT-PCR as necessary. One should capture all transcripts in gateway vectors as a global resource, and then and only then proceed to the development of optimized global expression arrays for the given genome.

Supervised Learning from Unphased SNP Data: A Case Study in Multiple Myeloma

C.David Page
Dept. of Biostatistics & Medical Informatics, University of Wisconsin

Joint work with Michael Waddell (University of Wisconsin), Bart Barlogie (University of Arkansas Medical Sciences) and John Shaughnessy (University of Arkansas Medical Sciences)

Single-nucleotide polymorphisms (SNPs) hold the promise of making it possible quickly gain insight into genetic factors of disease. We study 3000 SNPs from patients with the cancer multiple myeloma. 40 of these patients developed the disease before age 40, while 40 of the patients developed the disease after age 70. We test the hypothesis that patients who contracted the disease before age 40 have genetic differences from patients who contract the disease after age 70.

We employ support vector machines (SVMs), a type of supervised learning algorithm, to construct a classifier to distinguish between the two age groups based on SNP profile alone. The classifier achieves over 70 percent cross-validation accuracy, which is significantly better than chance, lending support to the hypothesis of genetic difference.

It is important to note that we use "unphased" SNP data (haplotypes are not known or estimated), because we do not have sufficient data to accurately estimate haplotypes. Performance might be improved by obtaining more data and attempting to construct haplotypes. It appears likely that performance also can be improved by using a larger number of SNPs.

On Inferring Regulation from Gene Expression Profiles

Dana Pe'er
The Hebrew University of Jerusalem

The talk consists of two papers that deal with two different algorithms to reconstruct regulation from gene expression data, included is the abstract for each.

Regulatory relations between genes are an important component of molecular pathways. Here, we devise a novel global method that finds a small set of relevant active regulators from a set of gene expression profiles, identifies the genes that they regulate, and derives their functional annotations. We show that our algorithm is capable of handling a large number of genes in a short time and robust to a wide range of parameters. We apply our method to a combined dataset of S. cerevisiase expression profiles, and validate the resulting model of regulation by cross-validation and extensive biological analysis of the selected regulators and their derived annotations. A living cell is a complex system that performs multiple functions and responds to a variety of signals. To achieve this, the cell is organized as a network of interacting functional modules, where a module consists of a set of genes that are co-regulated to optimize a response to different conditions. In this paper, we present module

networks, a novel, fully-automated, probabilistic method for discovering regulatory modules based on gene expression data. The method identifies both the genes composing each module and the regulators controlling their activation. We applied this method to 173 gene expression arrays measuring the response of Saccharomyces cerevisiae to various stress conditions. We validated the inferred modules using gene annotations, and known and novel binding site motifs, demonstrating the method's ability to identify highly coherent modules and their regulators. The method also suggests testable novel biological hypotheses about gene regulation in the form regulator 'X' regulates process 'Y' under conditions 'W'. We offer experimental results supporting four of our method's hypotheses, suggesting regulatory roles for previously uncharacterized proteins, including two putative transcription factors and two putative signal transduction molecules.

Automated Identification of Functionally Important Regions on the Surface of Proteins

Tal Pupko
School Of Computational Science & Information Technology, Florida State University

(Joint work with Fabian Glaser and Nir Ben-Tal from the Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel.)

In this work we develop a new algorithm to identify functionally important patches on the surface of proteins. We first apply the maximum likelihood (ML)-based rate4site algorithm to assign a rate for each amino-acid site. The goal here is to identify those positions with low evolutionary rates that correspond to conserved regions, and thus might be functionally important. However, conservation alone is not enough. A functional site is likely to include several conserved regions, that are close to each other in the 3D space, and that are all exposed rather than buried in the protein core. To this end, our goal was to develop an algorithm that searches for a "patch" of several amino acids that are both conserved and physically close to each other. We thus use both the evolutionary information (for assigning a rate for each site) and the 3D information (for searching the functional regions). The algorithm is based on a probabilistic approach: for each possible "patch" we compute a probability; The probability of a patch is computed by contrasting the size of the observed patch, with patches of the same evolutionary rate that are expected by chance alone. We then chose the patch that maximizes this probability.

We validate our novel algorithm on two real examples (the SH2 Src domain and the Bcl-xl protein). The preliminary results are promising. We were able to find the known functional regions *in silico*. We believe that this algorithm is a step forward in the development of automated tools to aid in protein function prediction.

Motif Inference

Marie-France Sagot
Inria Rhône-Alpes and Laboratoire de Biométrie et Biologie Évolutive, University Lyon I, France

This talk introduces various mathematical models and combinatorial algorithms for inferring network expressions that appear exactly or approximately repeated in a word or are common to a set of words, where by "network expression" (Mehldau and Myers, 1993) is meant a regular expression without Kleene closure. A network expression is therefore any expression built up of concatenation and union operators. This has many applications in molecular biology. The network expressions considered may contain spacers where by "spacer" is meant any number of don't care symbols. "Constrained spacers" are consecutive don't care symbols whose number ranges over a fixed interval of values. Network expressions with possibly don't care symbols but no spacers are called "simple" while network expressions with spacers are called "flexible" if the spacers are unconstrained and "structured" otherwise.

A few applications to the identification of promoter and regulatory sequences in bacteria are then mentioned. One of them is used to illustrate some of the remaining difficulties with all currently existing methods.

Dealing with Non-Unique Probes: DNA Chips and Group Testing

Alexander Schliep
MPI für Molekulare Genetik

For a number of applications DNA-chips which can robustly identify the presence of target organisms in a sample are needed; efficient computational methods have been proposed recently. Their practical applicability is limited by the high redundancy of closely related biological sequences, as finding unique probes for each individual target is often impossible due to oligo length and melting temperature constraints.

We propose to select sets of probes, which are non-unique with respect to the targets but still allow to infer information about the presence of one or several targets in a sample. This can be accomplished by using a non-adaptive statistical group testing approach, similar to the one successfully employed in the context of screening clone libraries. For DNA chips, a probe defines a group of targets as those it hybridizes to.

The two main theoretical problems are choosing an optimal group testing design as a subset of columns of the target-probe hybridization incidence matrix and how to infer the presence of one or several targets in a sample from the result of the hybridization experiments for each probe. The former can be addressed with a greedy heuristic and the latter with a Bayesian approach using Markov Chain Monte Carlo for computations.

Time Series Analysis with HMMs: Gene Expression Data

Alexander Schliep
MPI für Molekulare Genetik

In microarray experiments the expression levels of thousands of genes are being measured simultaneously. When performing microarray experiments consecutively in time we call this experimental setting a time course of gene expression profiles. One goal of such a setting is the detection of the underlying cellular processes, to set up regulatory networks and to assign function to time courses. As a prerequisite we want to identify time courses of equivalent qualitative behaviour.

We propose a model-based approach. Instead of defining a distance measure and grouping data points in a way that minimizes a distance-based scoring function (such as Euclidean distancea and k-means), statistical models are used to represent clusters. Cluster membership is decided based on maximization of a data point's likelihood given a model/cluster.

We apply Hidden Markov Models (HMMs) as our model class. Besides their prevalent use for biological sequence analysis, HMMs have been successfully applied for analysing time series data in a wide range of different problem domains. They are particularly suitable, if essential types of qualitative behavior can be proposed, as ''grammatical'' or ''structural'' constraints in the data can be effectively and explicitly modeled. We present a method to partition a set of expression time course data into clusters by use of HMMs. Given a number of clusters, each of which is represented by one Hidden Markov Model from a finite collection encompassing typical qualitative behavior, an iterative procedure finds cluster models and an assignment of data points to these models, which maximizes the joint likelihood of the clustering.

Algorithms that support Manual and Automatic Integration of Biological Networks

Benno Schwikowski
The Institute for Systems Biology

Genomic sequencing projects have provided us with an approximate parts list of biological systems. We are now moving towards an understanding of the dynamic interplay between the parts in these systmes. This development is fueled by high-throughput technologies, such as DNA microarrays, and mass spectrometry. These technologies allow to measure the dynamic expression of the parts, as well as technologies such as yeast two-hybrid, that define the interactions between the parts. Computational support is essential in the assembly of these data into models. Computationally integrating the data that we obtain from various technologies is critical for the following three reasons:

- The raw data that "high-throughput" technologies that allow to collect require primary processing and assembly on a scale that cannot be performed by hand.

- Each experimental technology highlights only one aspect of living systems. On the way to comprehensive and predictive models, the views from multiple experimental technologies need to be combined.

- Some of the corresponding experimental technologies are still in their infancy; their error properties are frequently unknown, which means that there needs to be a high level of error tolerance that can only be achieved through integration with other data.

We highlight several examples of our work in this area; algorithms that integrate protein interaction data with protein sequence and gene expression data, and present implementations in our visualization and analysis software platform Cytoscape.

Elucidation of HIV-1 Drug Resistance

Joachim Selbig
MPI of Molecular Plant Physiology, Golm/Potsdam

In treatment of HIV-1 infections physicians use two types of drug-resistance tests. Genotypic testing tries to determine drug resistance based on the virus's genetic makeup. Phenotypic testing, in contrast, evaluates how well the virus growes in the presence of different drugs. Whereas phenotypic tests may miss clinically important indicators the prediction of phenotypic resistance factors only from genotypes is a bioinformatics challenge.

We applied two types of machine learning methods, decision tree generation and support vector machines, to analyze correlations between HIV-1 genotype and resistance phenotype based on more than 650 samples. In addition, we derived mutual information profiles which quantify the statistical significance of the sequence positions of two viral enzymes for the discriminination between susceptibility and resistance.

The good predictive power of our classification models allowed us to provide a freely available prediction service.

Analysis of Gene Networks

Ron Shamir
Tel Aviv University

We are involved in an ongoing effort to model, understand and eventually reverse engineer genetic networks based on high-throughput data. Initial work on expansion of a known genetic

network based on gene expression data proved promising and validated our modeling assumption. We developed a biclustering algorithm to detect tight regulatory modules, which allowed annotations of numerous unknown yeast genes at high specificity, and experimentally validated the suggested functionality of some genes.

We have carried out a genome-wide in-silico determination of transcriptional regulation modules controlling cell cycle in human cells.

We are also studying and validating experimentally fitness function to evaluate putative regulatory functions.

Joint work in parts with Amos Tanay, Roded Sharan, Rani Elkon, Chaim Linhart, Irit Gat-Viks (School of Computer Science, TAU), Martin Kupiec (Faculty of Life Sciences, TAU) and Yossi Shiloh (Faculty of Medicine, TAU).

Analysis of Gene Expression Data on Biological Networks with ToPNet

Florian Sohler
Institut für Prak. Informatik & Bioinformatik, Ludwig-Maximilians-Universität München

Due to improvement in experimental techniques gene expression data have become available in large quatities. Data on biological networks is coming in from different sources such as yeast-2-hybrid scans, text mining and manually curated databases. Although networks are still very sparse, and network and gene expression data are often very noisy, we believe that the combination of these data types can aid the interpretation and validation of either one.

Automatic methods alone are not capable yet to accomplish that goal, so visualization tools and user interaction are required. Several such tools are being developed in commercial and academic institutions. We present ToPNet, a software developed in cooperation with biologists, designed as a visualization aid and a platform for method development. ToPNet is capable of integrating and visualizing different data types, linking network objects (genes, proteins, reactions) to reference databases like SwissProt or KEGG. A special feature is the ability to display Medline abstracts that have been found with the text mining tool ProMiner.

Some algorithmic methods like a significant area search, i.e. the search for a connected subnet with significant gene expression data, are also included. For more information see the ToPNet website: http://cartan.gmd.de/ToPNet/.

Reconstruction of Gene Networks from Expression Data

Lev A. Soinov
EBI-EMBL, Cambridge, UK

Reconstructing and modelling gene expression networks is one of the most challenging problems of functional genomics. Most network models can be described as graphs in which nodes represent genes and the edge between two nodes indicates the existence of an interaction between the connected genes. Control or influence functions associated with each node reflect how input signals affect particular genes. The reconstruction of a gene network means the reconstruction of its architecture as well as of the influence functions. Network architecture does not pose conceptual problems - knowing the list of connections for each node is enough to answer the question. At the same time the appropriate choice of influence functions is one of the main theoretical and practical issues.
Here we describe a machine-learning approach for reconstructing elements of gene networks. This approach is based on building classifiers – functions that determine the state of a particular gene's transcription machinery via the expression levels of other genes in the system. It allows us to identify genes affecting the target gene directly from the classifier and

to solve the problem of signal discretization without any subjective assumptions. All expression profiles are treated as examples for classification algorithms designed to learn from these examples. Classifiers given in the form of decision trees/tables/rules are easy to interpret and compare to the existing knowledge.
To build a mathematical model for a real-world problem, we usually need to make some assumptions. The gene whose expression is to be predicted is called here "the predicted gene", while the genes that are used to make the prediction - "the explaining genes". We assume that the transcription machinery of a predicted gene can be in a finite number of different states, which depend on the expression levels of other genes, and that the expression of the gene is determined by its state. For simplicity we allow here the predicted genes to be only in two states, “upregulated” and “downregulated”, though the model can be generalised to any number of states in a straightforward manner. However, this does not imply an *a priori* discretization of expression measurements for the explaining genes. On the contrary, there is no single threshold for the expression levels – one explaining gene may affect the state of different predicted genes at different thresholds. The way to determine the levels of expression sufficient for switching regulated genes between different states is to have the discretization be a part of the classification procedure, optimising the accuracy of predictions. Here, our approach is rather different from Boolean networks, and, in fact, from any approach that depends on *a priori* discretization of expression data.
It is also important that our classifiers are not black boxes – they consist of sets of simple rules that can be used for building and explaining gene networks. For each predicted gene we know precisely which genes affect it and how.
Of course expression networks alone are not sufficient for the reconstruction of the whole picture of biochemical processes. A combination of expression data with information about protein-protein interactions, metabolic pathways, etc., should be used to address this question. However, the presented approach can be applied to both relatively small gene sets of particular interest (e.g., involved in the same pathway, say, in order to find missing participants or to verify the interrelationships between known ones) and to complete genomes for large-scale investigations.

Reference:
1. Soinov L., Krestyaninova M., Brazma A. Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biology*, 2002 (in press).

Automation of Protein Structure Prediction

Ingolf Sommer[1], Niklas von Öhsen[2]
[1] – Max-Planck-Institute for Informatics, [2] – FraunhoferInstitute forScientific Computing and Algorithms

In the talk we gave an overview of our fully automated protein structure prediction server Arby, which combines the results of several fold recognition methods to find suitable templates in a database of structural representatives of protein domains.

The method starts by constructing a set of subsequences from the query sequence, each subsequence representing a hypothesis for a possible protein domain. This is done by scanning against the InterPro database and using hits as domain hypotheses. Additional hypotheses are constructed using a secondary structure prediction from PSIPRED. Segments of predicted loops are used as potential domain boundaries. Finally, the set of subsequences is reduced to a reasonable size by removing subsequences that are highly similar or short.

For each subsequence a multiple alignment is constructed by searching the NR database using PSI-BLAST. A frequency profile is calculated from this multiple alignment using a slightly modified version of the Henikoff-Henikoff sequence-weighting algorithm.

Each of the potential domains is then subjected to four different fold recognition methods. Each method searches for an optimal structure in our template database. The template database is a representative subset of the SCOP domains with pairwise sequence identity lower than 40%. For each of these template domains, a frequency profile was constructed as

described above for the targets. The first fold recognition method is PSI-BLAST, which is used to search through our set of template domains (augmented by the NR sequence database). The second one is the 123D threading program. It uses frequency profiles on the target side and 3D structural information on the template side. The third one is the JProp profile-profile alignment method recently developed in our group. It compares frequency profiles on the target side with profiles on the template side using the log average scoring approach. The fourth method is again the JProp profile-profile alignment program, but in this version it makes use of additional secondary structure information on the target and template side (publication in preparation).

The quality of each of these search results is assessed using confidence measures. For PSI-BLAST, these are readily available, for the other methods, these were developed in a recent study.

The target sequence is then annotated with all the produced quadruplets (subsequence, fold recognition method, search result, confidence value). Finally, we select a set of non-overlapping annotations along the sequence, by performing combinatorial optimization of a heuristic score based on the confidence values. For each of these selected annotations, a separate protein domain is predicted. The structure of this domain prediction is computed by aligning the subsequence against the template structure using JProp.

The underlying machinery is a Java based data flow engine, designed for stability. Since it is general and independent of the specific pipeline (as the one described above), it can be used as infrastructure for other projects as well: we developed a component framework in which all algorithms and programs are encapsulated in small Java classes. Each of these components specifies an algorithm to be executed along with its input parameters, the output that it produces, and possible error conditions. The accompanying engine provides a number of features for the components: First of all, the input/output dependencies of components are resolved. If all inputs for a specific algorithm have been determined, the algorithm itself is being scheduled for execution. The components are executed in parallel on any number of CPUs, in our case 10 CPUs of a SunFire 4800 server. A frequent problem in fully automated systems is reliable error handling. We solve this problem by catching potential error conditions and adaptively pruning the data-flow tree. Additionally, persistence of the computed results is accomplished by using a relational database, thus offering convenient and fast access to previously computed results for identical input parameters.

The power of the structure prediction server is based on the use of modern profile-profile algorithms for fold recognition, the quality assessment using confidence measures, and the stable and powerful Java data flow engine. In future work, we will use the latter technology as a basis for our bioinformatics computing environment.

Algorithms for High-Level Comparative Genomics

Jens Stoye
Bielefeld University

We have presented methods for comparing gene orders in completely sequenced genomes. This is a standard approach to locate clusters of functionally associated genes.

Often, gene orders are modeled as permutations. Given k permutations of n elements, a k-tuple of intervals of these permutations consisting of the same set of elements is called a common interval. We considered several problems related to common intervals in multiple genomes. We presented an algorithm that finds all common intervals in a family of genomes, each of which might consist of several chromosomes. We presented another algorithm that finds all common intervals in a family of circular permutations. A third algorithm finds all common intervals in
signed permutations. We also investigated how to combine these approaches. All algorithms have optimal worst-case time complexity and use linear space.

Finally we have discussed how the gene order and the number of common gene clusters of different species might be used to infer evolutionary trees.

This is joint work with Steffen Heber, UC San Diego.

Integrated Machine Learning of Patterns from Heterogeneous Data Sources – An application in Characterising α-amylases Superfamily.

Aik Choon Tan and David Gilbert
Bioinformatics Research Centre, Dept. of Computing Science, University of Glasgow

Research in bioinformatics and molecular biology is mostly driven by the experimental data. Current biological databases are populated by vast amounts of experimental data. Human experts are unable to cope with this fast growing trend, and require some automatic yet intelligent approaches to help them to extract the useful biological information underlying these databases. Machine learning is one such approach which has been widely applied to bioinformatics and has gained a lot of success in this research area.

One of the current research trends in machine learning applied to bioinformatics is to combine several sophisticated learning algorithms in order to increase a classifier's predictive accuracy (credibility) and its explanatory power (comprehensibility). When trying to learn from large and diverse data sets (e.g. biological databases) it is important to produce hypotheses that encapsulate all the information from different sources. The classifiers that are used to characterise and/or classify the data must be accurate and easily understandable by the human expert. Most methods in bioinformatics only concentrate on the classifier's credibility and less often emphasise its comprehensibility

The aim of this research is to construct a novel approach to induce invariant relationships from distributed heterogeneous biological data sources using knowledge discovery and ensemble learning techniques. Specifically, the objective is to produce credible yet comprehensible hypotheses to assist biologists in the data mining and knowledge discovery process.

We applied this technique in characterising proteins that belongs to the α-amylases superfamily. In this system, we applied decision trees as the base learning systems and inductive logic programming as the combined learner. The combined classifiers integrate sequence patterns, TOPS patterns, enzymatic functional class and also CATH structural classification learning from the individual dataset. We searched the PDB using the combined classifiers and successfully retrieved other members of the protein family. Since we are using rule-induction learners in our systems, the combined classifiers can be translated into a set of IF-THEN rules. We have shown that our approach is capable to discriminate positive and negative examples as well as improving its expressive power.

Finding Haplotype Blocks and Founder Sequences

Esko Ukkonen
Department of Computer Science, University of Helsinki

Haplotyped DNA sequences and dense SNP sequences of a sample of individuals from the same species, such as humans, open up new possibilities to uncover the variational structure in the genomes. We discuss two computational problems arising in this context; our results appear in the WABI2002 and PSB2003 conferences.

First, we describe a new method for finding so-called haplotype blocks based on the use of the minimum description principle. We give a rigorous definition of the the quality of a

segmentation of a genomic region into blocks, and describe a polynomial time dynamic pragramming algorithm for finding the optimal segmentation with respect to this measure. We also describe a method for evaluating the significance of each block boundary. The method has been applied to the published data of Daly et al. The results are in relatively good agreement with the published results, but also show clear differences in the predicted blocks and their strengths. We also give results on the block structure in population isolates in Finland.

Second, we consider the problem of reversing the recombination structure inherent in genomic sequences. Here we assume that our sequences have been sampled from a population isolate that was founded some generations ago by a relatively small number of settlers. Then the sequences in our given sample should be a result of recombinations of the corresponding sequences of the founders, possibly corrupted by (rare) point mutations. We are interested in finding plausible reconstructions of the sequences of the founders. We give a precise combinatorial formulation of the founder reconstruction problem, with a convenient equivalent formulation as a string coloring problem. Polynomial-time algorithms for some special cases as well as a general solution by dynamic programming are given.

Algorithms for Protein Structure Comparison

Juris Vīksna and David Gilbert
Glasgow University, UK

The objective is to develop efficient methods for protein structure comparison and analysis based on topological structure representations (TOPS diagrams). At this level proteins are described as ordered graphs and the basic problems are: which algorithmic questions about graph properties are the most adequate from the biological point of view; are there any particular constraints in existing data that could be exploited to obtain more efficient algorithms; which of these algorithmic problems still can be solved in a reasonable time for the existing data.

The current work has been centred on the study of subgraph isomorphism and maximal common subgraph problems in these graphs. For the existing data these seems to be already quite adequate notions for search and discovery of biologically meaningful motifs. We have developed a new fast SI algorithm for ordered graphs; this allows solving MCS problem by exhaustive search of all possible subgraphs still in reasonable time for the existing data. An advantage of such approach is that the time is just proportional to the number of samples (and in practice multiple comparisons tend to be even faster than pairwise); as a downside the complexity grows very fast with the size of samples and the method cannot be applied to much larger structures than we have in topology database.

We have evaluated our method in comparison with CATH and SCOP classifications. The conclusion is that topological information makes a difference (TOPS works much better that just amino acid or SSE sequences and is faster than coordinate based comparison methods) and can be surprisingly good, if the proteins have sufficiently rich secondary structure. However, this is the case for about 35% of α-β and 25% of β proteins. To improve this there is an ongoing work on inclusion of additional topological information in the database.

Genomic versus Functional Context: Complementary Approaches to Elucidation of Gene Function

Thomas Werner
Genomatix Software GmbH, Landsberger Strasse 6, D-80339 München, Germany

Elucidation of gene function is no longer restricted to the functional properties of proteins. In order to elucidate the functional context of genes it is necessary to include functional partners whether they interact physically with the gene product (e.g. in protein complexes) or not (e.g.

in signaling or metabolic cascades). Part of the functional context can be derived from the regulatory context as functionally interacting genes are often coregulated on a transcriptional level. A significant part of this coregulation is also determined within gene promoters via so-called promoter modules that act as endpoints of singaling cascades. The example of the chemokine RANTES shows that promoter analysis and computational modeling can indeed reveal partner genes from the functional context with high selectivity in a database search and can reveal functional connection that are very difficult to elucidate by proteomics approaches. It is most useful to combine data and algorithms that are independent, i.e. do not use common experimental basis or common features for analysis. Such combinations (e.g. expression data with genomic sequence analysis of promoters) will reduce the amount of results dramatically. Although some good results will be inevitably lost during the process as well, overall quality of the remaining results justifies that sacrifice.

List of Participants

Mario Albrecht
MPI für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken, (D)
Tel: +49 681 9325 327
mario.albrecht@mpi-sb.mpg.de

Mar Albà
Universitat - Pompeu Fabra
Biomedical Informatics
Passeig Maritim de la Barceloneta
E-08003 Barcelona, (E)
malba@imim.es
http://www1.imim.es/~malba/

Rolf Backofen
Friedrich-Schiller-Universität
Fakultät f. Mathematik & Informatik
LST für Bioinformatik
Ernst-Abbe-Platz 1-4
D-07743 Jena, (D)
Tel: +49-3641-9-46451
backofen@informatik.uni-jena.de

Niko Beerenwinkel
MPI für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken, (D)
Tel: +49 681 9325 304
beerenwinkel@mpi-sb.mpg.de
http://www.mpi-sb.mpg.de/~niko/

Soren Brunak
Technical University of Denmark
Center for Biological Sequence Analysis
Bio-Centrum DTU / Bldg. 208
DK-2800 Lyngby, (DK)
Tel: +45-45 25 24 77
brunak@cbs.dtu.dk
http://www.cbs.dtu.dk

Mark Craven
University of Wisconsin
Dept. of Biostatistics & Medical Informatics
5730 Medical Sciences Center
1300 University Avenue WI 53706
Madison, (USA)
Tel: +1-608-265-6181
craven@biostat.wisc.edu
http://www.biostat.wisc.edu/~craven/

Ingvar Eidhammer
University of Bergen
Dept. of Informatics, HIB
Høyteknologisenteret, Thormohlensgt. 55
P.O.B. 7800 N-5020 Bergen, (N)
Tel: +47-5558-4164
ingvar@ii.uib.no

Patricia Evans
University of New Brunswick
Faculty of Computer Science
Office E25 - Head Hall
P.O. Box 4400
E3B 5A3 Fredericton (New Brunswick), (CDN)
Tel: +1-506-458-7276
pevans@unb.ca
http://www.cs.unb.ca/profs/pevans/

Anders Fausboll
Technical University of Denmark
Center for Biological Sequence Analysis
Building 208
DK-2800 Lyngby, (DK)
Tel: +45-45 25 2477
fausboll@cbs.dtu.dk

Nir Friedman
The Hebrew University of Jerusalem
School of Engineering & Computer Science
Ross Bldg, Room 203
Givat Ram 91904 Jerusalem, (IL)
Tel: +972-2-658-4720
nir@cs.huji.ac.il
http://www.cs.huji.ac.il/~nir/

Irit Gat-viks
Tel Aviv University
School of Computer Science
Faculty of Exact Sciences, Schreiber Bldg. Room 11
Ramat Aviv 69978 Tel-Aviv, (IL)
Tel: +972-3-640-5397
iritg@post.tau.ac.il
http://www.cs.tau.ac.il/~iritg/

David Roger Gilbert
University of Glasgow
Dept. of Computing Science
Bioinformatics Research Centre
8-17 Lilybank Gardens

G12 8QQ Glasgow, (GB)
Tel: +44 141 330 2563
drg@dcs.gla.ac.uk
http://www.brc.dcs.gla.ac.uk/~drg

Hanspeter Herzel
Humboldt Universität
Institut für Biologie
Institute for Theoretical Biology (ITB)
Invalidenstr. 43 D-10115 Berlin, (D)
Tel: +49-30-2093-9101
h.herzel@biologie.hu-berlin.de
http://itb.biologie.hu-berlin.de/

Des Higgins
University College Cork
Dept. of Biochemistry
Lee Maltings, Prospect Row
Mardyke Cork, (IRL)
d.Higgins@ucc.ie

Alfred Hofmann
Springer-Verlag
Tiergartenstr. 17
D-69121 Heidelberg, (D)

Sorin Istrail
Celera Genomics Corp.
45 West Gude Drive
MD 20850 Rockville, (USA)
sorin.istrail@celera.com
http://www.celera.com/

Peter Jeavons
Oxford University
Computing Laboratory
Parks Road, Wolfson Bldg.
OX1 3QD Oxford, (GB)
Peter.Jeavons@comlab.ox.ac.uk
http://web.comlab.ox.ac.uk/oucl/people/peter.jeavons.html

Thomas S. Jensen
Technical University of Denmark
Center for Biological Sequence Analysis
Building 208 DK-2800 Lyngby, (DK)
Tel: +45-45 25 24 85
skot@cbs.dtu.dk

Kevin Karplus
University of California at Santa Clara
Dept. of Computer Engineering
CA 95053 Santa Clara, (USA)
Tel: +1-831-459-4250
karplus@soe.ucsc.edu
http://www.soe.ucsc.edu/~karplus/

Ross Donald King
University of Wales, Computer Science
Ceredigion, SY23 3DD Aberysthwyth, (GB)
Tel: +44-1970-622-432
rdk@aber.ac.uk
http://users.aber.ac.uk/~rdk/

Lars Kunert
MPI für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken, (D)

Thomas Lengauer
MPI für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken, (D)
Tel: +49 681 9325 300
lengauer@mpi-sb.mpg.de
http://www.mpi-sb.mpg.de/~lengauer/

Hans-Peter Lenhof
Universität des Saarlandes
Zentrum für Bioinformatik
Geb. 36.1 PF 15 11 50
D-66041 Saarbrücken, (D)
Tel: +49-681-302-64701
lenhof@bioinf.uni-sb.de
http://www.zbi-saar.de/chair/

Jochen Maydt
MPI für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken, (D)
jochen.maydt@gmx.de

Satoru Miyano
University of Tokyo
Human Genome Center - Institute of Medical Science
4-6-1 Shinokanedai
Minato-ku 108-8639 Tokyo, (J)
Tel: +81-3-5449-5615
miyano@ims.u-tokyo.ac.jp
http://bonsai.ims.u-tokyo.ac.jp/

Gene Myers
University of California at Berkeley
Dept. of EECS
775 Soda Hall
CA 94720-1776 Berkeley, (USA)
gene@EECS.Berkeley.EDU

C.David Page
University of Wisconsin
Dept. of Biostatistics & Medical Informatics
Medical Science Center, Rm 6743
1300 University Avenue
WI 53706 Madison, (USA)
Page@biostat.wisc.edu
http://www.cs.wisc.edu/~dpage/

Dana Pe'er
The Hebrew University of Jerusalem
School of Engineering & Computer Science
Givat Ram 91904 Jerusalem, (IL)
danab@cs.huji.ac.il
http://www.cs.huji.ac.il/~danab/

Tal Pupko
Florida State University
School of Computational Science & Information Technology
150-D Dirac Science P.O. Box 4530
FL 32306-4120 Tallahassee, (USA)
Tel: +1-850-645-0314
pupko@csit.fsu.edu
http://www.csit.fsu.edu/~pupko

Ela Pustulka-Hunt
University of Glasgow
Dept. of Computing Science
Room F163 8-17 Lilybank Gardens
G12 8QQ Glasgow, (GB)
ela@dcs.gla.ac.uk
http://www.dcs.gla.ac.uk/~ela/

Jörg Rahnenfuehrer
MPI für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken, (D)

Somak Ray
MPI für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken, (D)

Marie-France Sagot
Université Claude Bernard Lyon 1
Lab. de Biométrie et Biologie Évolutive
INRIA Rhones-Alpes, Projet Helix
43, Boulevard du 11 Nov. 1918
F-69622 Villeurbanne, (F)
marie-france.sagot@inria.fr
http://www.inrialpes.fr/helix/people/sagot/

Alexander Schliep
MPI für Molekulare Genetik
Abt. Comp. Molecular Biology
Ihnestr. 73 D-14195 Berlin, (D)
Tel: +49-30-8413-1166
schliep@molgen.mpg.de
http://www.molgen.mpg.de/~schliep/

Benno Schwikowski
The Institute for Systems Biology
1441 North 34th Street
WA 98103-8904 Seattle, (USA)
Tel: +1-206-732-1296
benno@schwikowski.de
http://www.gmd.de/People/Benno.Schwikowski/

Joachim Selbig
MPI für Molekulare Pflanzenphysiologie
Am Mühlenberg 1 D-14476 Golm, (D)
Tel: +49-331 567 8208
selbig@mpimp-golm.mpg.de
http://www.mpimp-golm.mpg.de

Ron Shamir
Tel Aviv University
School of Computer Science
Ramat Aviv 69978 Tel-Aviv, (IL)
Tel: +972-3-640-9356
shamir@math.tau.ac.il
http://www.math.tau.ac.il/~rshamir/

Francisco Silva Domingues
MPI für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken, (D)
Tel: +49 681 9325 326
Fax: +49 681 9325 399
doming@mpi-sb.mpg.de
http://www.mpi-sb.mpg.de/~doming/

Florian Sohler
Ludwig-Maximilians-Universität München
Institut für Prak. Informatik & Bioinformatik
Theresienstr. 39 D-80333 München, (D)
sohler@bio.informatik.uni-muenchen.de

Lev A. Soinov
EMBL Outstation
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton CB10 1SD Cambridge, (GB)
lev@ebi.ac.uk

http://www.ebi.ac.uk/microarray/

Ingolf Sommer
MPI für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken, (D)
Tel: +49 681 9325 306
sommer@mpi-sb.mpg.de
http://www.mpi-sb.mpg.de/~sommer

Jens Stoye
Universität Bielefeld
Technische Fakultät
AG Genominformatik
Universitätsstr. 25
10 01 31 D-33501 Bielefeld, (D)
Tel: +49-521-106-3852
stoye@techfak.uni-bielefeld.de
http://www.techfak.uni-bielefeld.de/~stoye/

Priti Talwar
MPI für Informatik
Stuhlsatzenhausweg 85
D-66123 Saarbrücken, (D)
Tel: +49 681 9325 328
ptalwar@mpi-sb.mpg.de

Aik Choon Tan
University of Glasgow
Dept. of Computing Science
Bioinformatics Research Centre
8-17 Lilybank Gardens
G12 8QQ Glasgow, (GB)
Tel: +44-141-330-3371
actan@brc.dcs.gla.ac.uk
http://www.brc.dcs.gla.ac.uk/~actan/

Esko Ukkonen
University of Helsinki
Dept. of Computer Science
Teollisuuskatu 23
P.O. Box 26 FIN-00014 Helsinki, (FIN)
Tel: +358-9-19144172
ukkonen@cs.Helsinki.FI
http://www.cs.helsinki.fi/u/ukkonen/

Juris Viksna
University of Glasgow
Dept. of Computing Science
Bioinformatics Research Centre
8-17 Lilybank Gardens
G12 8QQ Glasgow, (GB)
jviksna@cclu.lv, juris@brc.dcs.gla.ac.uk
http://www.brc.dcs.gla.ac.uk/~juris/

Thomas Werner
Genomatix Software
Landsberger Str. 6
D-80339 München, (D)
Tel: +49 89 599766 0
werner@genomatix.de
http://www.genomatix.de

Roland Yap
National Univ. of Singapore
School of Computing
Dept. of Computer Science / S16 Level 5
3 Science Drive 2, 117543 Singapore, (SGP)
Tel: +65-6844-2972
ryap@comp.nus.edu.sg
http://www.comp.nus.edu.sg/~ryap/

Ulrik de Lichtenberg
Technical University of Denmark
Center for Biological Sequence Analysis
Building 208 DK-2800 Lyngby, (DK)
Tel: +45-45 25 24 85
ulrik@cbs.dtu.dk