Bioinformatics

Schloss Dagstuhl, December 3 – 8, 2000

organized by

Douglas Brutlag, Stanford University
Thomas Lengauer, GMD Sankt Augustin
Martin Vingron, MPI Berlin

Preface

After two Seminars in the years 1992 and 1995 this was the third Dagstuhl seminar with a broad scope ranging over a variety of fields in Computational Biology.
In the five years since the previous Dagstuhl seminar Computational Biology has experienced dramatic growth and a significant shift of focus. In addition to the classical grand challenge problems such as gene identification and protein folding, computational biology has been confronted with a host of application-oriented problems. These problems originate from new experimental data being generated with efforts in genomics and proteomics and ask for unveiling the biological secrets that are hidden in these data. Furthermore, the expected publication of the human genome sequence left its distinct mark on the seminar.
This Dagstuhl seminar intended to focus specifically on these topical issues of the field. Specific topics included

support for large scale sequencing (shotgun sequencing the human genome);
annotating biological sequences (coding and non-coding regions);
comparative genomics;
structural genomics;
functional genomics;
analysis and interpretation of expression data;
modeling of cellular processes and pathways;
medical applications, genetics, genotyping;
proteomics.

Besides the talks and informal discussions there were two evening jam sessions on expression patterns and gene identification in the light of the imminent publication of the human genome sequence. There also was a session on new modes of teaching in computational biology.
One of the highlights of the seminar was the very wet hike in the hills around Dagstuhl on Santa Claus day (December 6).
As probably a premier for Dagstuhl seminars, a crew from German public television visited the group on Tuesday, Dec 5 and aired a short segment on the seminar over national television (Channel: 3sat, Program: Nano) on the following day. Interview partners in the program were Doug Brutlag, Tom Lengauer, and Knut Reinert. An internet version of the segment can be found at http://www.3sat.de/nano/astuecke/13391/index.html.
Many participants expressed the hope that this inspiring seminar series continue.

Program

Monday, December 4, 2000

Morning session **Chair: Thomas Lengauer**

Thomas Lengauer (GMD Sankt Augustin)
Introductory remarks

Jürgen Kleffe (Free University of Berlin) 19
Constrained gene prediction

Teresa Attwood (University of Manchester) 7
The role of gene family databases in functional annotation of sequences

Hanspeter Herzel (Humboldt University, Berlin) 17
DNA-arrays: Reliability, clustering, and detection of transcription factor binding sites

Afternoon session **Chair: Martin Vingron**

Hagit Shatkay (National Institute of Health, Bethasda) 23
Using information retrieval for large scale gene analysis

Matthias Fellenberg (Max-Planck-Institute for Biochemistry, Munich) 14
Integrative analysis of gene expression data

Ralf Zimmer (GMD Sankt Augustin) 29
Structure based target finding exploiting context information

Alexander Zien (GMD Sankt Augustin) 28
Gene expression data analysis

Tuesday, December 5, 2000

Morning session **Chair: Douglas Brutlag**

Knut Reinert (Celera Genomics Corp., Rockville) 21
A hierarchical assembler for the human genome

Daniel Huson (Celera Genomics Corp., Rockville) 17
Algorithmic aspects of genome assembly

William Taylor (National Institute for Medical Research, London) 24
Tiling with basic protein architectures

Stephen Bryant (National Institute of Health, Bethasda) 12
A conserved domain database

Afternoon session **Chair: Stephen Bryant**

Inge Jonassen (University of Bergen) 18
Microarray data analysis

Alvis Brazma (European Bioinformatics Institute, Hinxton) 8
From DNA-chips to reverse engineering of gene regulatory networks

Mark Craven (University of Wisconsin, Madison) 13
Machine learning applied to uncovering gene regulation

Thomas Werner (GSF Neuherberg) 27
Promoter finding on a genomic scale

Evening session

Discussion round on expression patterns

Wednesday, December 6, 2000

Morning session **Chair: William Taylor**

David Gilbert (City University London) 15
Protein Toplogy - techniques for pattern matching, pattern discovery and structure comparison

Juris Viksna (City University London) 26
Pattern matching and pattern learning algorithms for TOPS diagrams

Alfonso Valencia (Centro Nacional de Biotecnologa, Madrid) 25
Detecting networks of protein interaction

Søren Brunak (Technical University of Denmark, Lyngby) 10
Strategies for prediction of orphan protein function

Afternoon

Excursion

Thursday, December 7, 2000

Morning session **Chair: Thomas Lengauer/David Gilbert**

Douglas Brutlag (Stanford University) 11
Automatic discovery of protein motifs

David Gilbert (City University London) 15
Graph-based analysis of biochemical networks and interpretation of expression data

Joachim Selbig (GMD Sankt Augustin) 22
Machine learning-based analysis of genotypic, phenotypic, and clinical data in order to optimize therapies against HIV

Jens Stoye (German Cancer Research Center, Heidelberg) 23
Protein sequence classification with jumping alignments

Afternoon session **Chair: Douglas Brutlag**

Ivo Grosse (Free University of Berlin) 16
Computational gene recognition – an information theoretic approach

Teresa Attwood, Douglas Brutlag, Mark Craven, David Gilbert, Martin Vingron
Teaching bioinformatics

Evening session

Discussion round on genome sequencing and gene finding

Friday, December 8, 2000

Morning session **Chair: Thomas Lengauer**

Hans-Peter Lenhof (Saarland University, Saarbrücken) 20
An NMR-spectra-based scoring function for protein docking

Inge Jonassen (University of Bergen) 18
Protein structure motif discovery

Thomas Lengauer (GMD Sankt Augustin)
Concluding remarks

The role of gene family databases in functional annotation of sequences

Teresa K. Attwood
University of Manchester, UK

A legacy of the genome projects is sequence data 'overload'. We are now compelled to rationalise the mass of sequence data; to derive more efficient means of data storage; and to design more incisive and reliable analysis tools. The imperative for bioinformatics is to convert sequence information into biochemical and biophysical knowledge, to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences. Identifying evolutionary links between sequences is useful, as this often implies a shared function. We begin to make progress when we gather sequences into families, allowing us to discover conserved structural and functional motifs. Many methods have arisen to encode such patterns of conservation, e.g.: single motif methods (e.g., regexs), multiple motif methods (e.g., fingerprints and blocks) and full domain methods (e.g., profiles and HMMs). Fingerprinting arose because of the diagnostic limitations of exact regex matching. Fingerprints are groups of conserved motifs, used for iterative database searching. Iteration refines the fingerprint, and potency is gained from the mutual context afforded by motif neighbours. Thus, results are biologically more meaningful than from single motifs. Moreover, analyses are hierarchical, allowing more fine-grained diagnoses.

Given the variety and fallibility of databases available, it is important to devise a comprehensive search protocol, and to estimate significance by comparing results and finding a consensus. But why not just use BLAST/PSI-BLAST? The answer is that primary searches won't always allow outright diagnoses: these programs are not infallible; PSI-BLAST can mislead if not adequately supervised; and annotations of retrieved hits may be incorrect. The hope is that since gene family databases contain potent descriptors, distant relationships missed by BLAST may be captured by one or more of the family or functional site distillations.

The family databases now available include: PROSITE; Profiles; PRINTS; Pfam; BLOCKS; and eMOTIF. Which is best? In seeking distant homologues, the hit-or-miss nature of PROSITE regexs can render them worthless; in spite of their complexity, profiles and HMMs are often out-performed by simpler motif methods; the non-weighting system of fingerprints means that Twilight relationships may be missed; the scoring system used to create blocks generates large amounts of noise that may obscure the signal; only PROSITE and PRINTS are fully manually annotated. The bottom line - no method alone is best. This highlights the importance of integrated approaches, such as InterPro. To simplify sequence analysis, the family databases have been integrated within InterPro to create a

central annotation resource, with pointers to its satellite databases. Initial partners were PRINTS, PROSITE, profiles and Pfam; more recently, ProDom and BLOCKS.
Creating and searching family databases lie at different ends of a fallible chain of events. Thus our databases and search routines aren't perfect, but they offer a useful complement to other analysis tools. Until there's more experimental data available, they're among the best tools we have for functional annotation of genome sequences. Gene family databases offer several benefits: by using multiple sequences, trivial sequence errors may be diluted; they can highlight annotation errors if a sequence description differs from that of its family; they allow specific diagnoses, placing sequences in a family context for a more informed assessment of function. There is a long way to go before the databases, individually or together, are complete. Nevertheless, as they grow, their diagnostic potency ensures that gene family databases will play an increasingly important role in genome annotation.

From DNA-chips to reverse engineering of gene regulatory networks

Alvis Brazma and Jaak Vilo
European Bioinformatics Institute, Hinxton, UK

DNA chips are one of the most important recent break-throughs in experimental molecular biology. By enabling researchers to make snapshots of gene expression levels of tens of thousands of genes at a given moment, DNA chip technology is already producing floods of valuable data. The analysis and handling of these data is one of the most important and interesting problems of computational biology.
The raw microarray data are images, which have to be transformed into gene expression matrices – tables where rows represent genes, columns represent various samples such as different tissues, and values at each position characterize the expression level of the particular gene in the particular sample. These matrices have to be analyzed further, if any knowledge about the underlying biological processes is to be extracted. Storing and annotating these data is also a non-trivial problem. We discuss all these mentioned aspects of gene expression data managing and analysis, as well as our efforts to establish international standards for microarray data representation and annotation, and a public repository for such data.
We have been analyzing gene expression data for the systematic discovering of

novel putative regulatory motifs on the full-genome scale.
The analysis pipeline consists of 1) gene expression data clustering, 2) sequence pattern discovery from upstream sequences of genes, 3) a control experiment for pattern significance threshold limit detection, 4) selection of interesting patterns, 5) grouping of these patterns, 6) representing the pattern groups in a concise form and 7) evaluating the discovered putative signals against existing databases of regulatory signals. The pattern discovery is computationally the most expensive and crucial step. Our tool performs a rapid exhaustive search for a priori unknown statistically significant sequence patterns of unrestricted length. The statistical significance is determined for a set of sequences in each cluster with respect to a set of background sequences allowing the detection of subtle regulatory signals specific for each cluster. The potentially large number of significant patterns is reduced to a small number of groups by clustering them by mutual similarity. Automatically derived consensus patterns of these groups represent the results in a comprehensive way for a human investigator.
We are developing a set of Internet tools called collectively Expression Profiler (see http://www.ebi.ac.uk/microarray/) that will allow users to browse and query microarray data stored in a database ArrayExpress at EBI as well as from other databases on the web [1]. The main challenge is the integration of different types of data and presentation of these data in useful form for biologists to perform the analysis and study the complex relationships in the data. Different components of Expression Profiler allows users to cluster and visualize the gene expression data; connect the results of clustering to other tools on the web; perform the extraction of upstream sequences for the gene clusters; perform pattern discovery from the extracted sequences; as well as to visualize the discovered patterns on these sequences.
We will discuss the analysis of the gene expression and promoter data for the yeast Saccharomyces Cerevisiae [2,3], as well as the tools in the Expression Profiler analysis suite.
In the end we propose a new model for describing gene regulatory networks that can capture discrete (boolean) and continuous (differential) aspects of gene regulation. After giving some illustrations of the model, we study the problem of the reverse engineering of such networks, i.e., how to construct a network from gene expression data. We prove that for our model there exists an algorithm finding a network compatible with the given data. We also describe some generalizations of the model, discuss their relevance to the real-world gene networks and formulate a number of open problems.

References

[1] A. Brazma, A. Robinson, G. Cameron, and M. Ashburner, One stop shop for microarray data, Nature 403:699-700 (2000).

[2] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen, Predicting gene regulatory elements in silico on a genomic scale Genome Research 8:1202-1215 (1998).
[3] J. Vilo, A. Brazma, I. Jonassen, A. Robinson, and E. Ukkonen, Mining for putative regulatory elements in the yeast genome using gene expression data, ISMB-2000, August 2000, AAAI Press, (pp. 384–394).

Strategies for prediction of orphan protein function

Søren Brunak
Technical University of Denmark, Lyngby, Denmark

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 will increase 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 protein modifications, sorting signals, and gene expression data from DNA array experiments are integrated and used to infer the functional class.

References

[1] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, Protein Eng. 10:1-6 (1997).
[2] H. Nielsen, S. Brunak, and G. von Heijne, Machine learning approaches to the prediction of signal peptides and other protein sorting signals, Protein Eng. 11:3-9 (1999).
[3] J.E. Hansen, O. Lund, N. Tolstrup, K. Rapacki, and S. Brunak, Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility, Glycoconjugate J. 15:115-130 (1998).

[4] N. Blom, S. Gammeltoft, and S. Brunak, Sequence and structure-based prediction of eukaryotic protein phosphorylation sites, J. Mol. Biol. 294:1351-1362 (1999).
[5] P. Baldi and S. Brunak, Bioinformatics: The machine learning approach, MIT Press, Cambridge, Mass., 1998.

Automatic discovery of protein motifs

Douglas L. Brutlag
Stanford University, USA

We have developed two methods for discovering conserved sequence motifs from families of aligned protein sequences. The first method is called EMOTIF (http://motif.stanford.edu/emotif). Given an aligned set of protein sequences, EMOTIF generates a set of regular expressions, termed eMOTIFs, with a wide range of specificities and sensitivities. EMOTIF can also generate motifs that describe possible subfamilies of a protein superfamily. A disjunction of such motifs can often represent the entire superfamily with high specificity and sensitivity. eMOTIF permits one to develop highly specific motifs that can be used to search entire proteomes rapidly and with no false predictions.
We have also used the PSI-BLAST search program to generate protein families and superfamilies and to identify highly conserved regions within these families (eBLOCKs: http://eblocks.stanford.edu/). Starting with the Swiss-Prot database of 79,000 proteins, we generated 19,000 protein families with over 81,000 conserved eBLOCKs. We have applied the eMATRIX method to convert these conserved regions into scoring matrices capable of discovering the function of unknown proteins or genes.
We have used eMOTIFs and eMATRICES built from eBLOCKs to identify newly sequenced proteins in humans, non-humans and Drosophila proteins. eBLOCKs were able to assign functions to 70% of the human proteins, 60% of the non-human proteins and 50% of the proteins in the Drosophila proteome. These identifications agreed with homology methods when the homology method found a significant homolog. However, the eMATRIX method was able to assign functions to 40% for which there were no known homologs.

A conserved domain database

Stephen Bryant
National Institute of Health, Bethasda, USA

The molecular biology information services currently operated by the NCBI are based on a model of links and neighbors. Objects from different databases, such as protein structures and MedLine citations point to one another. Objects from the same database are neighbored by similarity, via the VAST algorithm for structure comparison, for example. This model creates a powerful retrieval system, but does not provide summary classifications, such as identification of previously described protein domain families. To add this capability my colleagues and I have begun a project to build a database of conserved domain alignments based on sequence and structure similarity.

Our goal for CDD (a Conserved Domain Database) is to produce core-structure alignments that identify the homologous residues conserved across a domain family. This may be used for inheritance of annotation and inference of 3D structure, when a protein matches a profile made from the CDD alignment. The different lines of evidence to consider in making alignments are sequence conservation, location of compact domains as identified from 3D structure, location of functional sites, and 3D structure similarity. To do so we must devise algorithms and procedures to combine these lines of evidence in a largely automatic way, with a minimum input from human curators.

Seed alignments for CDD are taken from curated domain alignment collections, currently Pfam, SMART and LOAD. These sources supply information on sequence conservation and domain boundary location. Using BLAST comparison, the sequence fragments in these sources are matched to sequences in the public database, and when possible substituted with sequences of known 3D structure. The alignments are automatically modified to exlcude deletions in reference sequence, as required for construction of PSSMs (position-specific score matrices). Alignments at this level of processing are used in the public CD-search service, available at http://www.ncbi.nlm.nih.gov/Structure. This provides reverse PSI-BLAST search of the CDD profile library, with display of multiple alignments, including Cn3D graphics display of included structures.

CDD alignments are further processed for quantitative reconciliation with 3D structure. Only residues consistently aligned with the reference structure are retained, to create an alignment consisting of several ungapped blocks. Outlier sequences that introduce insertions within aligned blocks consistently identified in other sequences are excluded. Residues aligned to the reference structure in a chemically implausible way are also excluded, such as those adjacent to another in sequence, but aligned to sites in the reference structure that are too far apart.

When more than one structure in the aligned family is known, structurally aligned residues from the reference structure are added to the CD alignment, provided they extend existing blocks only by addition of previously unaligned residues.
Automatic processing cannot reasonably identify a conserved core structure in all cases, however. It is common, for example, for N- or C-terminal strands in a beta sheet to be excluded, even though they correspond to a central strand that is necessary for maintenance of the structure. To correct these deficiencies the blocks defining the core elements in a CDD alignment may be edited by human experts, with the goal of identify a geometrically compact core structure, consistent with aligned blocks and any functional sites apparent from the structure or sequence annotations. In cases where the seed alignment identifies only a short motif or repeat, the edited CDD alignment may significantly extend the boundaries of the conserved domain. Curation of core structure alignments is currently in progress for a representative sample of domains in the collection.
CDD core structure alignments are critical for automated update by addition of new sequences. New domain family members are detected by PSI-BLAST search, but then merged into the alignment by the core-structure threading procedures previously developed in our laboratory, and tested in the CASP competitions. Manual intervention by curators is required only when threading detects a potential conflict of a new family member with the previously defined core structure. It remains to be seen how much curator time will be required for maintenance of up-to-date domain alignments, representing all known subfamilies, with separate subfamily PSSMs when required for sensitive searching. Anecdotal evidence so far, such as from the GAF family, indicates that it is typically rather straightforward to define and update a core structure alignment, whenever a 3D structure is known, and that this leads to improved search performance of PSSMs constructed from this alignment.

Machine learning applied to uncovering gene regulation

Mark Craven
University of Wisconsin, Madison, USA

We have been applying machine learning methods to the task of uncovering regulatory elements in E. coli and other prokaryotes. A first step in this line of research has been to develop an approach to predicting operons. Our approach induces predictive models for this task from a rich variety of data types including sequence data, gene expression data, and functional annotations associated

with genes. We use multiple learned models that individually predict promoters, terminators and operons themselves. A key part of our approach is a dynamic programming method that uses our predictions to map every known and putative gene in a given genome into its most probable operon. We have also been investigating the use of stochastic context free grammars (SCFGs) for recognizing terminators in E. coli. Our experiments indicate that SCFGs provide superior predictive accuracy to a variety of competing approaches. In our most recent work, we are developing a novel algorithm for refining the structure of a stochastic context free grammar during the learning process. Initial results indicate that this approach can provide more accurate learned models.

Integrative analysis of gene expression data

Matthias Fellenberg
Max-Planck-Institute for Biochemistry and Biomax Informatics AG, Munich, Germany

DNA microarrays allow the measurement of gene expression profiles on a genomic scale. I presented three instances of an integrated analysis of gene expression data combining the expression data with systematic functional and metabolic data and with other large-scale data sets. These analyses have been performed for yeast using gene expression data that is publicly available (Diauxic Shift, DeRisi et al., Science, 1997).
After clustering the ≈6100 genes using the topology conserving self-organizing map neural network algorithm we do a functional projection: a category from the hierarchically organized MIPS functional catalogue is selected and the distribution of the genes of this category over the clusters is computed. For many categories we found that the genes of the categories are largely contained in a group of neighboring clusters. For the diauxic shift experiment the functional projection revelaed that the cytoplasmatic and the mitochondrial ribosomal proteins are located in two clearly seperated regions of the cluster map.
The metabolic mapping approach allows to evaluate the clustering in terms of metabolic pathways. For a group of selected clusters we map the genes to the corresponding bioreactions via the EC numbers annotated for the genes. These reactions are linked via common metabolites resulting in a number of metabolic networks of reactions whose genes are coregulated.
Using the protein-protein interaction data collected at MIPS, we can compute clusters of interacting proteins that are coregulated in the gene expression experiment analyzed (cf. Fellenberg et al., ISMB 2000).

Protein Toplogy - techniques for pattern matching, pattern discovery and structure comparison

David Gilbert
City University London, UK

We present a formal description of protein topology based on TOPS representations of protein structures. These comprise Secondary Structure Elements (SSEs), beta-sheet connectivities and certain chiralities. We have also defined topological patterns which are like structure descriptions with inserts of SSEs permitted between pattern SSEs. We have developed a pattern matching algorithm which exploits the constraints imposed by the ordering on SSEs to prune the search space. Our pattern discovery technique works by repeated pattern matching and pattern extension and is of complexity linear in the number of examples. We have developed a method to divide a set of protein domains into subsets each with its own characteristic pattern, using a rating function based on the "goodness" of the pattern (using a compression measure) and their coverage of the example set. Such unions of patterns can be used to characterise the CATH or SCOP protein hierarchy with greater discrimination than simple (non-union) patterns.
In addition a distance between two proteins can be computed by using a common discovered pattern to produce a structural alignment of their SSE sequences, and then computing a sum of the edit distances over non-matching subsequences plus a penalty for non-matched arcs (hbonds and chiralities).
This work is collaborative with David Westhead (Leeds, UK) and Juris Viksna (Latvia).

Graph-based analysis of biochemical networks and interpretation of expression data

David Gilbert
City University London, UK

The PFBP group at EBI led by Shoshana Wodak and managed by Jacques van Helden is developing a database on protein function and cellular processes - aMAZE (http://www.ebi.ac.uk/research/pfbp). This covers metabolic pathways and gene regulation.
Together with Jacques van Helden and Lorenz Wernisch, I have been developing prototype tools to permit constrained path finding, pathway building and graph

extraction. We have developed a graphical interface that permits the display of automatically laid out pathway graphs.
We have used our system to demonstrate the possibility of analysing results from gene expression data by using "seed" nodes derived from cluster analysis to extract subgraphs from metabolic pathway databases.
We now plan to develop algorithms to compare pathways and to perform pattern discovery over pathways.
This work has been performed whilst a collaborator with the PFBP group.

Computational gene recognition – an information theoretic approach

Ivo Grosse
Free University of Berlin, Germany

One basic task in the analysis of DNA sequences is the identification of protein coding regions. Since biochemical techniques alone are not sufficient for identifying all coding regions in every genome, researchers from many fields have been attempting to find statistical patterns that are different in coding and noncoding DNA. Many such patterns have been found, but none seems to be species-independent. Hence, traditional coding measures based on these patterns need to be trained on organism-specific data sets before they can be applied to identify coding DNA. This training-set dependence limits the applicability of traditional coding measures, as many new genomes are currently beeing sequenced for which training sets do not exist.
We explore if there exist universal statistical patterns that are different in coding and noncoding DNA and can be found in all living organisms, regardless of their phylogenetic origin. We find that (i) the *mutual information function* has a significantly different functional form in coding and noncoding DNA. We further find that (ii) the probability distributions of the *average mutual information* (AMI) are significantly different in coding a noncoding DNA, while (iii) they are almost the same for organisms of all taxonomic classes. Surprisingly, we find that the AMI is capable of predicting coding regions as accurately as organism-specific coding measures.

DNA-arrays: Reliability, clustering, and detection of transcription factor binding sites

Hanspeter Herzel, Dieter Beule, Szymon Kielbasa, Jan Korbel, Christine Sers, Arif Malik, Holger Eickhoff, Hans Lehrach, Johannes Schuchhardt
Humboldt University and Max-Planck-Institute for Molecular Genetics, Berlin, Germany

High-density DNA-arrays allow measurements of gene expression levels (messenger RNA abundance) for thousands of genes simultaneously. We analyze arrays with spotted cDNA used in monitoring of expression profiles. A dilution series of a mouse liver probe is deployed to quantify the reproducibility of expression measurements. Saturation effects limit the accessible signal range at high intensities. Additive noise and outshining from neighboring spots dominate at low intensities. For repeated measurements on the same filter and filter-to-filter comparisons correlation coefficients of 0.98 are found.
Next we consider the clustering of gene expression time series from stimulated human fibroblasts which aims at finding co-regulated genes. We analyze how preprocessing, the distance measure, and the clustering algorithm affect the resulting clusters. Finally we discuss algorithms for the identification of transcription factor binding sites from clusters of co-regulated genes.

Algorithmic aspects of genome assembly

Daniel Huson
Celera Genomics Corp., Rockville, USA

In modern biology, knowledge of the complete genome of a species is seen as a fundamental step towards it's full understanding. The unraveling of the human genome, in particular, is of great scientific importance, and publication of a first draft of the human genome is expected by the end of the year. We will first review a number of different sequencing and assembly strategies and then focus on some of the associated algorithmic problems.

Microarray data analysis

Inge Jonassen
University of Bergen, Norway

Microarray technology makes it possible to measure simultaneously the transcript abundance of large numbers of genes. In such an experiment one obtains for each gene an expression profile which can be represented as a vector in multidemensional space. The genes' expression levels can be analysed through an analysis of the corresponding vectors. In this talk we describe a number of different methods that have been proposed to analyse such data sets including hierarchical and particional clustering, projection methods, and self-organising maps. A tool called J-Express implementing all these methods in an integrated manner is used to illustrate the principles of the different methods. Furthermore we discuss how gene expression data can be used to learn about gene regulation. An approach is described for finding putative regulatory signals whose occurrences in gene promoter regions is correlated with the genes' expression profiles. First the genes are clustered based on similarity of expression profiles. Second the (putative) promoter regions of the genes in each cluster are anlysed to find patterns that are over-represented in these upstream regions as compared to their occurrences in all the promoter regions in the genome. The approach has been implemented in a system called Expression Profiler by Jaak Vilo at the European Bioinformatics Institute. Experimental results obtained in an anlysis of Yeast (S. Cerevisiae) data is presented. The results show that for strong clusters (with respect to similarity of gene expression profiles) one tends to find more significant patterns (with respect to how over-represented they are in the genes' upstream regions). Finally we briefly discuss the possibilities of learning regulatory networks from microarray data.

Protein structure motif discovery

Inge Jonassen
University of Bergen, Norway

We briefly give an introduction to the problem of protein structure comparison and discuss some methods for pairwise structure comparison/alignment. The SAP algorithm is described in some more detail. It is based on double dynamic programming and was developed by Taylor and Orengo. Most methods for structure comparison, including SAP, can only compare pairs of structures and can

only be extended to do multiple comparison through a strategy where a multiple alignment is built up through a series of pairwise alignments. Next we describe a more direct approach to the discovery of structure motifs in sets of protein structures. A method called SPratt is described which is able to analyse a large set (up to thousands) of protein structures simultaneously and find patterns common to at least some (user specified) minimum number of the structures. The method avoids the indirect and computationally expensive steps of pairwise structure comparisons. Instead it explores a set of possible patterns which are matched against the structures and non-trivial patterns matching the minimum number of structures are reported. The approach is based on using a string representation (neighbour string) of each residue's spatial neighbourhood. A neighbour string includes the residue and secondary structure type of all the residues within a certain spatial distance and they are ordered as along the protein's backbone. A search method is devised for finding patterns matching the minimum number of neighbour strings so that the geometric configuration of the residues constuting each of the matches are similar to each other. As a result the method is able to find patterns consisting of individual residues coming close together in space with constraints on the residues' secondary structure and amino acid types as well as their three-dimensional coordinates. We show how the patterns found can be further analysed by using SAP and that the patterns can in some cases improve the alignments found by SAP.

Constrained gene prediction

Jürgen Kleffe
Free University of Berlin, Germany

The human genome is as good as ready and understanding its message is the next important step to achieve those medical advances newspapers speculate about each day. The possibly simplest question to begin with is; where are the genes? There are three major sources of gene annotation. The experimental approach, which is expensive, the model based gene prediction, which is inaccurate, and sequence matching with ESTs, mRNAs, and protein sequences of other organisms, which is also error prone.
Hence, the current state of the art is not satisfactory. Every second published gene annotation is not entirely correct. On the other hand about 90% of the multi exon gene annotations are not entirely wrong. There are many problems with short single exon genes.
This situation may not improve soon since we know far too little about the precise working of gene regulation, mRNA transcription, splicing, and translation inside

the cell. Even though the 3-D structure of the ribosome is available, we know almost nothing about how it determines the start of translation. It seems obvious that correct gene prediction and improved mathematical modeling of these processes are interacting fields of continuos research. Published gene annotations must be considered temporary and prone to change as knowledge increases in the course of time.
This makes it necessary to carefully review each gene annotation before it can be used in other research projects. Do the latest sequence data bases provide new clues for a better location of the gene? Are there alternative gene predictions, and how can we combine such information in order to find the currently most likely gene annotation? These are some of the questions we ask and need software tools to find the answers.
Constrained gene prediction is one such tool. We describe the GeneGenerator program and its application to combine many different outputs of gene finder programs in to a single and hopefully improved consensus solution that also accommodates results from sequence matching with new ESTs, mRNAs or cDNAs as well as biological facts known about the gene of investigation.

References

[1] Kleffe et al. Bioinformatics 14:232-243 (1998
.

An NMR-spectra-based scoring function for protein docking

Hans-Peter Lenhof
Saarland University, Saarbrücken, Germany

A well studied problem in the area of Computational Molecular Biology is the so-called Protein-Protein Docking problem (PPD) that can be formulated as follows: Given two proteins A and B that form a protein complex, compute the 3D-structure of the protein complex AB. Protein docking algorithms can be used to study the driving forces and reaction mechanisms of docking processes. They are also able to speed up the lenghty process of experimental structure elucidation of protein complexes by proposing potential structures. In this paper, we are discussing a variant of the PPD-problem where the input consists of the tertiary structures of A and B plus an unassigned ^{1}H-NMR spectrum of the complex AB. We present a new scoring function for evaluating and ranking potential complex structures produced by a docking algorithm. The scoring

function computes a "theoretical" ^{1}H-NMR spectrum for each tentative complex structure and subtracts the calculated spectrum from the experimental spectrum. The absolute areas of the difference spectra are then used to rank the potential complex structures.
In contrast to formerly published approaches (e.g. Morelli et al., 2000) we do not use distance constraints (intermolecular NOE constraints). We have tested the approach with four protein complexes whose three-dimensional structures are stored in the PDB data bank and whose ^{1}H-NMR shift assignments are available from the BMRB database (BioMagResBank).
In all examples, the new scoring function produced very good rankings of the structures. The best result was obtained for an example, where all standard scoring functions failed completely. Here, our new scoring function achieved an almost perfect separation between good approximations of the true complex structure and false positives.
Unfortunately, the number of complexes with known structure and available spectra is very small. Nevertheless, these first experiments indicate that scoring functions based on comparisons of one- or multi-dimensional NMR spectra might be a good instrument to improve the reliability and accuracy of docking predictions and perhaps also of protein structure predictions (threading).

A hierarchical assembler for the human genome

Knut Reinert
Celera Genomics Corp., Rockville, USA

Two different strategies for determining the human genome are currently being pursued: one is the "clone-by-clone" approach, employed by the public Human Genome Project, and the other is the "whole genome shotgun" approach, favored by researchers at Celera Genomics. We describe the design, implementation and operation of a "hierarchical assembler" that makes use of preliminary data from both assembly projects and produces, at an early stage, a draft of the genome that is much more complete than is obtainable from either source separately. The pipeline is divided into three stages, *fragment recruitment*, *assembly*, and *tiling*. In the talk I discuss how these three stages are implemented and which tools are developed to ensure the correctess of the resulting assemblies.

Machine learning-based analysis of genotypic, phenotypic, and clinical data in order to optimize therapies against HIV

Joachim Selbig
GMD National Research Center for Information Technology,
Sankt Augustin, Germany

Human immunodeficiency virus type 1 (HIV-1) is the primary etiologic agent for Aquired Immune Deficiency Syndrome (AIDS). HIV-1 is a lentivirus, a separate genus of the Retroviridae, which are complex RNA viruses that integrate into the genome of host cells and replicate intracellularly. Currently, for treating HIV infected patients there are two possibilities to interfere with the replication cycle of the virus: Inhibitors of the two viral enzymes protease (PR) and reverse transcriptase (RT) are available. Since HIV shows a very high genomic variability, even under the usual combination therapy (HAART - highly active antiretroviral therapy) consisting of several drugs, mutations occur, that confer resistance to the prescribed drugs and even to drugs not yet prescribed (cross resistance). Therefore the treating physician is faced with the problem of finding a new therapy rather frequently.
Genotypic and phenotypic resistance tests have the potential to help identify which drugs in a regimen are failing and to guide the selection of drugs for new regimens. However, the relations between observed mutations, phenotypic resistance and therapy success are poorly understood so far.
The goal of our investigations is to develop bioinformatics methods that help to understand these connections and that contribute directly to therapy optimization. In a database, set up in collaboration with university hospitals and virological institutes, clinical data, sequence data and phenotypic resistance data are collected. Machine learning methods are then used to learn and predict properties like therapy success or drug resistance. Additionally, the mechanisms of drug resistance will be studied at the molecular level. To this end we will carry out force field based calculations on enzyme-inhibitor complexes.

Using information retrieval for large scale gene analysis

Hagit Shatkay
National Institute of Health, Bethasda, USA

The abundance of information resulting from DNA microarray experiments, accompanied by a significant increase in the amount of literature discussing gene-related discoveries, presents a major data analysis challenge.
Current methods for interpreting gene expression data typically rely on cluster analysis of expression levels. While clustering indeed reveals potentially meaningful relationships among genes, it can not explain the underlying biological mechanisms. Such explanations rely a great deal on human expertise and on scanning through the literature for information about each gene involved in the experiment. Automating the extraction of relevant information from the literature is a necessary step towards complete analysis of genome-wide data, complementing existing expression clustering techniques.
We present a new approach for utilizing the literature to establish functional relationships among genes on a genome-wide scale, automatically searching for the literature relevant to each studied gene, summarizing it, and relating the genes to each other. Our method is based on revealing coherent themes within the literature, using a new Expectation-Maximization algorithm. This algorithm produces sets of PubMed documents with a unifying theme, along with a list of terms characterizing each theme. The algorithm is used to find content-based relationships among abstracts, that are translated into functional connections among genes. A comparison of preliminary results, produced by applying our algorithms to a database of yeast-related abstracts, with well-established yeast genes functions, demonstrates the effectiveness of our approach.
Joint work with Stephen Edwards, John Wilbur and Mark Boguski.

Protein sequence classification with jumping alignments

Jens Stoye
German Cancer Research Center, Heidelberg, Germany

In our presentation we address the problem of classifying uncharacterized protein sequences. In particular, we present a method for comparing a protein sequence

to a given protein family. The rationale is to exploit both vertical and horizontal information of a multiple alignment in a well balanced manner. This is in contrast to established methods like profiles and hidden Markov models which focus on vertical information as they model the columns of the alignment independently.
In our setting we want to select from a given database of "candidate sequences" those proteins that belong to a given superfamily. In order to do so, each candidate sequence is separately tested against a multiple alignment of the known members of the superfamily by means of the so-called jumping alignment algorithm. This algorithm is an extension of the Smith-Waterman algorithm and computes a local alignment of a single sequence and a multiple alignment. In contrast to traditional methods, however, this alignment is not based on a summary of the individual columns of the multiple alignment. Rather, the candidate sequence is at each position aligned to one sequence of the multiple alignment, called the "reference sequence". In addition, the reference sequence may change within the alignment, while each such jump is penalized in order to prevent arbitrary jumping.
To evaluate the discriminative quality of the jumping alignment algorithm, we compared it to hidden Markov models on a subset of the SCOP database of protein domains. The discriminative quality was assessed by counting the number of false positives that ranked higher than the first true positive (FP-count). For moderate FP-counts above five, the number of successful searches with our method was considerably higher than with hidden Markov models.
Joint work with Rainer Spang, ISDS, Duke University and Marc Rehmsmeier, DKFZ Heidelberg.

Tiling with basic protein architectures

William R. Taylor
National Institute for Medical Research, London, UK

With the large number of protein structures not known, it is difficult to gain an overview of their variety of forms and even more difficult to comprehend how each structure relates to its neighbours. Despite systematic attempts to impose order on this variety, the current collections (SCOP, CATH, FSSP) are all based on the pairwise comparison of structures. Using this approach, the decision to group proteins can often be arbitrary, or more cautiously, not made at all.
The approach taken in this talk is to represent proteins as secondary structure vectors and to compare these to idealised stick structures (referred to as Forms). The Forms are derived from a minimal basis set which can be expanded through regular symmetry operations to cover arbitrarily large structures. Any given

protein can then be matched to a single instance of a Form or to a combination (tiling) with cah instance typically corresponding to a domain (or sub-domain) of the protein structure.
The ability of the stick-matching algorithm to find solutions up to but not beyond the core of the protein opens the possibility of using them simultaneously for domain definition and classification. The series of nested sub-solutions also represents a 'history' of the growth of the structure through the addition of secondary structure elements. The optimal path through these solutions (which can be extracted using dynamic programming) then represents a possible evolutionary path for the growth of the structure.
The matching of all proteins against all Forms will produce a classification of proteins – not unlike the Periodic Table of elements (although the proteins will require three or four dimensions) – which will allow the relationships between all proteins to be visualised.

Detecting networks of protein interaction

Alfonso Valencia
Centro Nacional de Biotecnologa, Madrid, Spain

The considerable amount of information available about individual protein components in the form of genome sequences, protein structures, and functional genomics (gene expression patterns) demands further work in integrating all the up to now disperse information. Protein interactions are the obvious next step in this direction. We present here three complementary computational efforts for the study of protein-protein interactions. The first approach is based on the study of the patterns of variation in multiple sequence alignments, in the search for the possible signals left by evolution for the process of compensatory substitutions and co-adaptation. We have previously demonstrated that the so called correlated mutations was enough to single out the right inter-domain docking solution amongst many wrong alternatives [1]. These predictions have been tested in different experimental systems [2-3]. The extension of this method to the detection of interacting partners in large collections of multiple-sequence alignments shows quite promising results in terms of the number of interactions predicted in complete genomes and quality of the predicted interactions when compared with known molecular complexes [4]. The second approach is based on the application of information extraction techniques [5-6] to the retrieval of information about protein interactions directly from the scientific literature (Medline abstracts). Our current system is able to detect automatically networks of functional interactions, by identifying automatically protein names and the actions liking them

[7]. The results of its application to different complex biological systems show interesting possibilities and a number of still problematic areas that require further development. Finally, the application of clustering techniques [8] and text retrieval methods [9] to the available expression array data leads to a new avenue for the discovery of relations between genes, that can be considered as complementary information to the predicted and detected protein interactions.

References

[1] Pazos et al., J. Mol. Biol. 272:1-13 (1997).
[2] Gdssler et al., Proc. Natl. Acad. Sci. USA. 95:15229-15234 (1998).
[3] Azuma et al., J. Mol. Biol. 289:1119-1130 (1999).
[4] Pazos et al., submitted (2000).
[5] Andrade, Valencia, ISMB 5:25-32 (1997).
[6] Andrade, Valencia, Bioinformatics 14:600-607 (1998).
[7] Blaschke et al., ISMB 7:60-67 (1999).
[8] Herreros et al., Bioinformatics (2000), in the press.
[9] Blaschke et al., Functional and Integrative Genomics (2000), in the press.

Pattern matching and pattern learning algorithms for TOPS diagrams

Juris Viksna and David Gilbert
City University London, UK

The objectives of this work was to develop algorithmic techniques for protein comparison and classification from the protein TOPS diagrams (secondary structure descriptions containing information about the secondary structure elements (strands, helices) and their relations (h-bonds, chiralities)) and to evaluate the biological relevance of these comparison/classification techniques. Protein comparison in TOPS formalism is based on finding a maximal "pattern" in a given set of proteins and the required methods can be reduced to subgraph isomorphism and maximal common subgraph problems in a special kind of graphs.
An efficient pattern matching (i.e. subgraph isomorphism) algorithm has been developed; protein comparison (i.e. maximal common subgraph) algorithm is based on repeated pattern extension and matching against all elements in a given set of proteins. The method works well for TOPS formalism, although it is doubtful whether this comparison method can be applied for much larger structures.
The comparison results have been compared with CATH classification. Results suggest that larger patterns characterize protein sets quite consistently with

CATH. As a next step it is planned to compute database of "good" TOPS patterns and to set up a web page offering protein comparison based on these patterns.

Promoter finding on a genomic scale

Thomas Werner
GSF National Research Center for Environment and Health, Neuherberg and Genomatix, Munich, Germany

The human genome sequencing project provided an unprecedented amount of data outside protein coding regions including all regulatory sequences. For several reasons promoters are among the most important regulatory sequences in the genome. All transcriptional signals finally have to be integrated at the gene promoters in order to influence the transcription of the genes including the action of enhancers or signal transduction pathways. Therefore, analysis of the promoter regions of the genes in the human genome will yield a wealth of information about the organization of life.
Unfortunately, human polymerase II promoters are not directly evident from the genomic sequence and knowledge of the cDNAs is of limited help as promoters may be located tens of kb upstream of the coding regions. Previous attempts to locate promoters by in silico prediction in genomic sequences met with limited success and are clearly not suitable for whole genome analysis. The genome sequencing era also brought along a change of paradigm in bioinformatics of sequence analysis. Whereas the main focus used to be sensitivity followed by experimental removal of false positive predictions whole genome analysis now requires specificity in the first place. The amount of data precludes large scale experimental evaluation of predictions turning false positive matches into real killers for the whole analysis.
We have designed a completely new approach to locate promoters in genomic sequences (working well at least for all mammalian species and many vertebrates). PromoterInspector is based on the concept of a common promoter context rather than internal common promoter features. The method uses feature extraction techniques to compile libraries of IUPAC words to be used by three different classifiers: Promoters are separately classified against exons, introns, and 3'-untranslated regions. Only 24 consecutive predictions of all three classifiers (window size 100 nucleotides, step 4 nucleotides) are accepted as a valid promoter region. Tests on shorter genomic regions as well as on whole human chromosomes 21 and 22 showed the method to be extremely specific and provided support by other annotation for at least 40% of the predictions. Thus, PromoterInspector is capable of high quality promoter annotation completely independent of any other

annotation. We finished the whole human genome in the mean time predicting about 30,000 promoters. Given the sensitivity of PromoterInspector is 50% this would bring our gene count up to 60,000 genes in the human genome.
Reliable promoter prediction may also help to significantly improve gene predictions as one of the weakest points of all gene finding programs is finding the first exon, especially if it is short and noncoding. Since this is exactly where the promoter is located combination of promoter prediction with gene prediction appears to be very promising.

Gene expression data analysis

Alexander Zien
GMD National Research Center for Information Technology,
Sankt Augustin, Germany

The first part of this talk relates to the pathway scoring method which was introduced by Ralf Zimmer in the previous talk. The goal is to identify (metabolic) pathways that are realized in (some of) the examined cells. We start by extracting subgraphs from a generic network that form plausible hypotheses. Each of these putative pathways is evaluated by scoring it with respect to the available gene expression data. Three scoring functions are presented in detail: The first one quantifies, individually for each gene, the amount of regulation. The second measures average correlation of the genes involved in the putative pathway. The third function incorporates both features by relating the covariance to the measurement error. Evidence is given that supports our view that this method is more appropriate for identifying distinct pathways than clustering.
In the next part, a novel method for the normalization of gene expression data is presented. First, for each pair of N measurements (eg, chips or filters) a relative scaling is estimated without relying on the questionable assumptions that common methods are based on (globalization, housekeeping). In the second step, an N-dimensional normalization vector is estimated that minimizes the inconsistencies with the pairwise scalings. The inconsistencies arise because the system is overdetermined by $N(N-1)/2$ erroneous pairwise estimates. The resulting normalization is shown to be more robust with respect to the selection of genes than conventional schemes.
Finally, a variant of the rank sum test (by Wilcoxon) is suggested that accounts for measurement errors. Because this test is distribution-free, it can be used to estimate the significance of differential expression of a gene in two classes of probes. The new idea is to sum probabilities of inversions instead of counting them. Further work is needed to explore properties and utility of this heuristic.

Structure based target finding exploiting context information

Ralf Zimmer
GMD National Research Center for Information Technology,
Sankt Augustin, Germany

Gene expression data measured with DNA chips and EST sequences currently are promising data for finding possible drug target proteins via bioinformatics. Structure based annotation methods can enhance more conventional sequence analysis and annotation approaches in order to detect more distant homologues and more clues towards possible functions and cellular roles of proteins. In order to be successful it is more and more important to make efficient use of the context information available on protein sequence and structure families, conserved sequence motifs and structural sites, secondary structure information, distance constraints, functional annotations, and biochemical network contexts.
We present a new approach to apply homology based structure prediction methods (threading) which exploits context information as pre- and post-filter and during the threading process. Context information is specified and stored in a formal language called ProML (Protein Meta Language), which is a variant of XML. Thus, a range of generic XML browsers and editors can be used to specify and inspect context information for protein structure prediction. In addition, the context information can directly be used in threading and clustering algorithms. Another type of context information are biochemical networks, which are used as functional constraints for large scale structure prediction tasks for differential expression measurements.
Biochemical (metabolic and regulatory) network knowledge is represented via Petri nets. This allows to define a notion of complete pathways to formalise the notion of cellular role in the target finding context. These complete pathways serve as biological hypotheses, which could be realised in a specific system state and whether they are is evaluated using the expression data at hand. Together with an appropriate statistical scoring function this allows for a new evaluation method for proteomics and gene expression measurements.