05441 Abstracts Collection
Managing and Mining Genome Information: Frontiers in Bioinformatics
— Dagstuhl Seminar —

Jacek Blazewicz[1], Johann Christoph Freytag[2] and Martin Vingron[3]

[1] Politechnika Poznanska, PL
jblazewicz@cs.put.poznan.pl
[2] HU Berlin, DE
freytag@dbis.informatik.hu-berlin.de
[3] MPI für Molekulare Genetik, DE
vingron@molgen.mpg.de

Abstract. From 30.10.05 to 04.11.05, the Dagstuhl Seminar 05441 "Managing and Mining Genome Information: Frontiers in Bioinformatics" was held in the International Conference and Research Center (IBFI), Schloss Dagstuhl. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Biological data management, semantic knowledge, ontologies, large-scale data mining, functional genomics

05441 Executive Summary – Managing and Mining Genome Information: Frontiers in Bioinformatics

This report summarizes the important aspects of the workshop on "Managing and Mining Genome Information: Frontiers in Bioinformatics" which took place October 31st until November 4th, 2005. Twenty five Participants came from six different countries representing various "branches" of the bioinformatics community. The presentations ranged from describing highly theoretical models to presenting prototypes or systems for managing and mining data in bioinformatics.

Keywords: Biological data management, semantic knowledge, ontologies, large-scale data mining, functional genomics

Joint work of: Blazewicz, Jacek; Freytag, Johann-Christoph; Vingron, Martin

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2006/475

SBH - State of the Art and Some Perspectives

Jacek Blazewicz (Poznan University of Technology, PL)

The talk was concerned with DNA Sequencing by Hybridization. Mathematical models for sequencing with isometric libraries were discussed. Spectra without errors, as well as these containing positive and negative errors, were considered. Special attention was paid to the concept of DNA graphs. hen, the new concept of isothermic libraries was introduced. Its features and advantages were later discussed.

Keywords: Sequencing by hybridization

Biclustering of Multiple Gene Expression Data Sets

Stefan Bleuler (ETH Zürich, CH)

Cluster analysis of gene expression data often involves multiple distinct data sets, e. g., multiple time course measurements.

Generally, existing clustering methods work on a single matrix of expression data thus requiring a concatenation of multiple data sets into one data matrix. This can be problematic if the measured expression values cannot be reliably compared across data sets which can be the case if data stemming from different types of experiments, different labs or different microarray technologies are involved. In the present study we investigate the effects of mixing data sets in the context of time course data and we propose a flexible clustering and biclustering framework which facilitates a joint analysis of multiple data sets while keeping them separate. The usefulness of this approach is demonstrated on several time course measurements from Arabidopsis thaliana. A promoter analysis searching for new transcription factor binding sites shows that for homogeneous data sets mixing is slightly beneficial while in the case of diverse data sets mixing is detrimental to the number and significance of the motifs.

Keywords: Biclustering, Gene Expression Data, Evolutionary Algorithms

Two-Phase EA/k-NN for Feature Selection and Classification in Cancer Microarray Datasets

David Corne (University of Reading, GB)

Efficient and reliable methods that can find a small sample of informative genes amongst thousands are of great importance. In this area, much research is investigating the combination of advanced search strategies (to find subsets of features), and classification methods. We investigate a simple evolutionary algorithm/classifier combination on two microarray cancer datasets, where this

combination is applied twice - once for feature selection, and once for further selection and classification. Our contribution are: (further) demonstration that a simple EA/classifier combination is capable of good feature discovery and classification performance with no initial dimensionality reduction; demonstration that a simple repeated EA/k-NN approach is capable of competitive or better performance than methods using more sophisticated preprocessing and classifier methods; new and challenging results on two public datasets with clear explanation of experimental setup; review material on the EA/k-NN area; and specific identification of genes that our work suggests are significant regarding colon cancer and prostate cancer.

Keywords: Classification methods, evolutionary algorithm

Mining DNA Micro-array data with association rules

Clarisse Dhaenens (Université de Lille, F)

Mining Microarray data has become a great challenge for biologists who are able to produce rapidly a large amount of these data. Classical approaches to mine such data are based on classification and clustering. Therefore data are presented in the "gene table" form. We propose to transpose the gene expression matrix in order to obtain a "treatment table" where genes are columns, lines are experiments and each cell is the expression level of the corresponding gene during the experiment. In this context, relations between genes are looked and we propose to use association rules in the form: IF C THEN P, where C is a Condition and P a Prediction. Such a rule could be:

```
If gene_32 is over_expressed AND gene_584 is under_expressed
   THEN gene_512 is over_expressed.
```

This problem may be seen as a combinatorial optimization problem. Hence the criterion has to be defined. But how can we qualify a "good rule"? Many criteria have been proposed and a statistical analysis of them lead us to propose a multi-objective model. For this problem, a multi-objective parallel genetic algorithm has been developed. As such an approach produces several solutions (rules), it is interesting to help biologists in analyzing them. A vizualization tool has been developed and a post-processing with other data sources is studied. Questions about this post-processing will be raised during the talk.

Keywords: Datamining, microarray, genomic, multiobjective optimization

Joint work of: Dhaenens, Clarisse; Khabzaoui, Mohammed; Talbi, El-Ghazali

Graph Data Management for Molecular and Cell Biology

Barbara Eckman (IBM Life Sciences - West Chester, USA)

As high-throughput biology begins to generate large volumes of systems biology data, there is a growing need for robust, efficient systems to manage metabolic and signaling pathways, chemical reaction networks, gene regulatory networks, and protein interaction networks. Network data frequently is best represented as graphs, and researchers need to navigate, query and manipulate this data in ways that are not well supported by standard Relational DataBase Management Systems (RDBMSs). Current approaches to managing graphs in RDBMSs rely either on external procedural logic to execute the graph algorithms, or clumsy and inefficient algorithms implemented in SQL.

This talk will describe the DB2 Systems Biology Graph Extender, a research prototype that extends DB2, IBMs RDBMS, with graph objects and operations to support declarative SQL queries over biological networks and other graph structures. Supported operations include neighborhood queries, shortest path queries, spanning trees, graph transposition, and graph matching. In a federated database environment, graph operations may be applied to data stored in any format, whether remote or local, relational or non-relational. A single federated query may include both graph-based predicates and predicates on related data sources, such as microarray expression levels, clinical prognosis/outcome, or the function of orthologous proteins in mouse disease models.

Keywords: Networks graphs pathways federation

Joint work of: Eckman, Barbara; Brown, Paul

Multistage Isothermic Sequencing by Hybridization

Piotr Formanowicz (Poznan University of Technology, PL)

Despite the impressive progress in many areas of biological sciences, reading DNA sequences still remains one of the most important problem. One of the methods of DNA sequencing is sequencing by hybridization. Although this method is quite modern, it suffers sensitivity to errors appearing in the biochemical stage of it. This is a motivation for developing some new variants of the method which should be more resistant to errors of these types.

Two of such non-standard approaches are multistage and isothermic sequencing by hybridization. Each of the methods is less sensitive to some types of the errors appearing in the biological stage, in comparison to the standard version of the method. On the other hand, they are more sensitive to the remaining types of errors. However, a combination of the two approaches reduces the sensitivity of the components.

Exploring overlapping data sources

Johann Christoph Freytag (HU Berlin, D)

During data integration users are often faced with differences in the data that come from different sources. To find those differences during data integration and to resolve them efficiently is a major issue.

In this talk I address the problem of how to derive/generate "contradiction patterns" automatically using existing techniques from data mining. Those patterns potentially provide hints and suggestions to the "domain expert" where (and why) differences in the data exist (with what kind of "significance") - it's then up to him/her to decide if those difference are "real" and how to resolve them thus generating integrated data sources without contradictions.

This work is ongoing in my research group in Berlin, specifically work with my Ph.D. student Heiko Mueller and my colleague Ulf Leser.

Keywords: Data cleansing, data quality, data integration

Full Paper:
http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/conferences/2004-iqis-mueller.pdf

See also: Heiko Müller, Ulf Leser, Johann-Christoph Freytag:Mining for Patterns in Contradictory Data, Proceedings of the SIGMOD International Workshop on Information Quality for Information Systems (IQIS'04), Paris, France, 2004

Hybryd systems identification - application to cell signaling pathways

Krzysztof Fujarewicz (Silesian University of Technology - Gliwice, PL)

The work is focused on the problem of identification of complex biological systems modeled by means of a set of nonlinear ordinary differential equations (ODE). The complexity of the task comes from the nature of measurements that are assumed to be known only at rare discrete-time moments. An example of such a system may be a model of cell signaling pathway where concentrations of mRNA, proteins and their complexes may be modeled using a set of nonlinear ODEs and these concentrations may be measured using different techniques, for example blotting techniques, electrophoretic mobility shift assays or gene expression microarrays, all giving measurements only at discrete-time moments.

We propose the adjoint sensitivity analysis as a very efficient tool for solving such a problem. Adjoint sensitivity analysis is a technique frequently used in practical optimization problems such as identification or optimization tasks. It calculates the gradient of a performance index very effectively and decreases computational cost comparing to straight (tangent linearized) sensitivity analysis.

In literature related to adjoint systems such systems are defined for continuous-time systems or discrete-time systems and cannot be directly applied to solve the problem of parameter fitting presented above, due to its hybrid, continuous/discrete-time nature. We show rules for creating the modified adjoint systems for hybrid systems and generation of the gradient of the performance index. Results of fitting of a mathematical model of NF-kappaB regulatory module are presented.

This work has been supported by Silesian University of Technology under grant BW/RGH-11/RAu-0/2005.

Keywords: Cell signaling pathways, system identification

Joint work of: Fujarewicz, Krzysztof; Lipniacki, Tomasz; Kimmel, Marek; Swierniak, Andrzej

Medical diagnosis by mass spectrometry - computational methods.

Anna Gambin (University of Warsaw, PL)

Mass spectrometry (MS) is a new high-throughput biotechnology: it allows one to determine the composition of complex protein mixtures, identify the structure of protein complexes, or even study the whole proteome. For bioinformaticians challenging computational problems arise here, like de-novo peptide sequencing or BLAST-like database filtering. However, the most promising application of mass spectrometry for medical sciences is the possibility of diagnosis based on MS data. There is a great hope that some disease markers should be detectable among the small peptides of blood serum or plasma. Identification of such biomarkers will result in fast, cheap and reliable diagnostic tests for many diseases.

In this talk I will focus on the computational problems behind the medical diagnosis from MS data. I would like to sketch the whole process starting from the automated identification of peptides, going through clustering of probes, dimensionality reduction, and ending with classification task and hunting for biomarkers. Our research follows the lines of recently published resutls. For several problems we propose new original solutions, for others we adopt standard techniques. There are however several substantial differences:

- we have decided to perform the complete computational process: we do not concentrate only on the classification but we do also all needed preprocessing steps (like peak-picking, noise reduction, etc.).
- as an answer for criticism of previous approaches, we improve the sensibility of peptide detection procedure by considering two dimensional data (one dimension corresponds to mass/charge ratio and the second to separation of peptides using hight performance liquid chromatography).

We have analyzed several datasets to check the efficiency and quality of our methods. Moreover, our main goal is to design the prognostic test for two possible forms of cystic fibrosis (light and severe). I will conclude with the preliminary result we have obtained till now.

Keywords: Mass spectrometry, HPLC, classification, dimensionality reduction

Joint work of: Gambin, Anna; Dutkowski, Janusz; Kowalczyk, Krzysztof; Kluge, Bogusław; Dadlez, Michał; Tiuryn, Jerzy

Modelling the kinetic behaviour of the MAPK cascade: negative feedback amplifier characteristics

David Roger Gilbert (University of Glasgow, GB)

We present a study of a signalling pathway, the Mitogen Activated Protein Kinase (MAPK) cascade. This pathway responds to a wide range of stimuli and plays a key role in many cellular processes, is deregulated in many diseases and is an important drug target. Specifically, we focus on the Ras-Raf-MEK-Erk pathway, and show how it is a biological amplifier, based on a cascade of three phosphorylation-dephosphorylation loops. Several negative feedback loops are known, including Erk to Sos and the recently identified Erk to Raf. Negative feedback amplifier theory from electronics/control theory predicts that the Erk cascade behaves in such a manner, and that the disruption of the effectiveness of the MEK stage would have a non-linear effect on the behaviour of the output (concentration of phosphorylated Erk) c.f. Sauro et al. 2004, effectively maintaining high levels of Erk-P output as a MEK inhibitor is added until eventually the system output decreases rapidly.

We have developed a number of computational models of the Erk cascade based on Ordinary Differential Equations, and discuss why Mass Action kinetics is preferred to Michaelis Menten kinetics. Our models clearly exhibit behavioural properties predicted by negative amplifier theory. We have validated our models in the wet-lab in-vivo, titrating in an MEK inhibitor to lower the amplifier gain, and using variant Raf constructs to eliminate the negative feedback. These in-vivo models exhibit the behaviour predicted by the computer model. Our work suggests that potential drug targets inside any such negative feedback loop within an amplifier will be poor targets in a live (in-vivo) system.

RO, OS and VV were supported by a grant from the Department of Trade and Industry under the 'Beacon' scheme.

Keywords: Systems biology, biochemical pathways / networks, signalling pathways, dynamic behaviour, negative feedback amplifer

Joint work of: Orton, Richard; Sturm, Oliver; Kolch, Walter; Vyshemirsky, Vladislav

ArrayExpress etc.: Microarray Informatics @ EBI

Misha Kapushesky (EBI - Cambridge, GB)

ArrayExpress is the European public data repository for microarray data. It is one of three MIAME - Minimal Information About a Microarray Experiment - compliant databases in the world (the others are GEO and CIBEX). When publishing microarray experiments in any of a number of major journals (Nature, Cell, Lancet, PLOS, etc.), the data should be submitted to one of the three MIAME repositories. Our team also provides a web-based tool for data submissions, MIAMExpress, which presents a series of pages with a data upload link at the end. Additionally, there are two value-added services we run: the ArrayExpress Data Warehouse, a reannotated, query-optimized database of gene expression data, and Expression Profiler, an online platform for exploratory data analysis and visualization.

Keywords: ArrayExpress, MIAMExpress, ArrayExpress Data Warehouse, Expression Profiler

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

Computational complexity of the Simplified Partial Digest Problem

Marta Kasprzak (Poznan University of Technology, PL)

The Simplified Partial Digest Problem is one of the problems of the restriction mapping of a genome. It bases on the digestion experiments with one restriction enzyme, but - as the advantage over the partial digest approach in the standard version - it needs only two such reactions: the short and the long one. This problem with the additional assumption that the instance contains no errors has been proven to be strongly NP-hard in its search version (although its decision version is easy because the solution always exists), by the transformation from the problem Numerical Matching with Target Sums.

Keywords: Computational complexity, DNA restriction mapping

A Query Language for Biological Networks

Ulf Leser (HU Berlin, D)

Many areas of modern biology are concerned with the management, storage, visualization, compari-son, and analysis of networks. For instance, networks are used to model signal transduction and metabolic pathways, gene regulation, and interaction of molecules in general.

A large number of da-tabases have emerged that collect and provide information on cellular networks and protein interac-tion. However, most users and applications are not concerned with entire databases, but search for specific subsets of the data. For these purposes, it is essential to be able to describe the desired sub-network as specific as necessary and as simple as possible. Despite the increased importance of net-work data in biology, there still exists no proper language for describing and retrieving specific parts of a network.

In this paper, we introduce the pathway query language (PQL) for retrieving specific parts of large, complex networks. The language is based on a simple graph data model with extensions re-flecting properties of biological objects. PQL queries match arbitrary subgraphs in the database based on node properties and paths between nodes. PQL is a powerful language, being able to ex-press graph isomorphism. A specific feature is that the result of a query is de-coupled from the matched subgraph. Thus, a query may require a certain structure in the database to exist, but return a different subgraph. Furthermore, the result of a PQL query itself is a graph and can be used in further queries, which allows for query composition, query nesting, and graph views, features well known from relational databases.

PQL is easy to learn for everybody with a basic knowledge of SQL. It is implemented on top of a relational database. A query is compiled into a stored procedure which returns the resulting graph in temporary tables. All computation is performed by relational queries, thus exploiting the capabilities of modern database systems in terms of query optimization and memory management. The code is for free available from the author.

Querying biological networks

Ulf Leser (HU Berlin, D)

Many areas of modern biology are concerned with the management, storage, visualization, compari-son, and analysis of networks. For instance, networks are used to model signal transduction and metabolic pathways, gene regulation, and interaction of molecules in general. A large number of da-tabases have emerged that collect and provide information on cellular networks and protein interac-tion. However, most users and applications are not concerned with entire databases, but search for specific subsets of the data. For these purposes, it is essential to be able to describe the desired sub-network as specific as necessary and as simple as possible. Despite the increased importance of net-work data in biology, there still exists no proper language for describing and retrieving specific parts of a network.

In this paper, we introduce the pathway query language (PQL) for retrieving specific parts of large, complex networks. The language is based on a simple graph data model with extensions re-flecting properties of biological objects. PQL queries match arbitrary subgraphs in the database based on node properties and paths between nodes. PQL is a powerful language, being able to ex-press

graph isomorphism. A specific feature is that the result of a query is de-coupled from the matched subgraph. Thus, a query may require a certain structure in the database to exist, but return a different subgraph. Furthermore, the result of a PQL query itself is a graph and can be used in further queries, which allows for query composition, query nesting, and graph views, features well known from relational databases.

PQL is easy to learn for everybody with a basic knowledge of SQL. It is implemented on top of a relational database. A query is compiled into a stored procedure which returns the resulting graph in temporary tables. All computation is performed by relational queries, thus exploiting the capabilities of modern database systems in terms of query optimization and memory management. The code is for free available from the author.

Keywords: Databases, query language, bioinformatics, systems biology

Full Paper:
http://www.informatik.hu-berlin.de/wbi

Tabu search strategy in HP protein model

Piotr Lukasiak (Poznan University of Technology, PL)

Understanding protein functionality would mean understanding the basics of life. This functionality follows a three-dimensional structure of proteins. Unfortunately till now it is not possible to obtain these structures artificially. HP-model is one of the most successful and well-studied simplified lattice model of protein folding, that use mathematical abstraction of proteins for hiding many aspects of folding process and works as hypothesis generator. Due to NP-hardness and NP-completeness results of the protein folding problem many computer scientists have tried to find fast approximation algorithms and meta-heuristics for searching enormous number of protein states. Tabu search (TS) strategy is one of the most successful meta-heuristics that has been applied for large number of optimization problems.

The condition that native conformation should be stable is not the only one, the protein has to be able to find this conformation in a short time, starting from a denaturated state characterized by a random population of unfolded conformations. A folding pathway is built into the structure by a natural selection. The method proposed, tries to follow the nature in the artificial environment and gives impressive results for sequences containing up to 100 amino acids.

The proposed algorithm has a good performance and finds low energy conformation values for subsequences of protein chains therefore compares well with the other heuristic approaches and due to its low computation time can be used as a complementary tool for an analysis of the three-dimensional protein structures.

Keywords: Tabu search, Meta-heuristic, Hydrophobic-hydrophilic lattice model, Protein structure prediction

Joint work of: Lukasiak, Piotr; Maciej, Milostan

What the plant does: Mass Spectrometry and Bioinformatics for Metabolomics

Steffen Neumann (IPB - Halle, D)

In the last years Metabolomics has emerged as an important technology used to solve functional genomics challenges. Community-wide accepted data standards for interchange are currently emerging, such as mzData (psidev.sf.net/ms) and ArMet (www.armet.org).

We present an infrastructure to support the use of these data standards combining XML I/O, database persistence and Editor functionality. Both mzData and ArMet have been created and described through UML Models, allowing for an implementation under the "Model driven Architecture (MDA)" paradigm. The Information contained in these models is sufficient to create Java Model implementation, XML I/O, Editor functionality, JSF Web Frontend and an object-relational persistence mapping for each of them using the Open Source Eclipse Platform with the EMF and JDO Plugins. The MDA approach allows to recreate the neccessary code basis and backend database with minimal manual coding, since the data standards can be expected to be evolving rapidly.

Additionally, we used metabolite data from seeds of two different Arabidopsis thaliana lines to train a bayesian network model to represent relationships between the data. For the structure learning of the bayesian networks we used the Sparse Candidate approach of Friedman et. al combined with maximisation methods based on greedy hill climbing and simulated annealing. We are able to represent extra information such as the type of the tissue or experiment treatments in additional nodes.

Keywords: Metabolomics, Mass Spectrometry, Databases, Model Driven Architecture, MDA, Clustering, Bayesian Networks

Weighted HMMs and Applications to Fragment Statistics for Peptide Mass Fingerprinting

Sven Rahmann (Universität Bielefeld, D)

Peptide mass fingerprinting is a technique that allows to identify a protein from its fragment masses obtained by mass spectrometry after enzymatic fragmentation: An experimental mass fingerprint is compared with or aligned to reference fingerprints obtained from protein databases using in-silico digestion. Recently, much attention has been given to the questions of how to score such an alignment of mass spectra and how to evaluate its significance; results have been developed mostly from a combinatorial perspective. In particular, existing methods generally do not (or only at the price of a combinatorial explosion) capture the fact that the same amino acid can have different masses because of, e.g., isotopic distributions or variable chemical modifications.

We offer several contributions: We introduce the notions of a probabilistically weighted alphabet, where each character can have different masses according to a specified probability distribution, and the notion of a random weighted string as a fundamental model for a random protein. We then develop a general computational framework, which we call weighted HMMs for various length and mass statistics of cleavage fragments of random proteins. We obtain general formulas for the length distribution of a fragment, the number of fragments, the joint length-mass distribution, and for fragment mass occurrence probabilities, and special results for so-called standard cleavage schemes (e.g., for Trypsin). Computational results are provided, as well as a comparison to proteins from the SwissProt database.

Keywords: Weighted HMM, mass spectrometry, peptide mass fingerprinting, convolution

Joint work of: Rahmann, Sven; Kaltenbach, Hans-Michael; Böcker, Sebastian

Facts from Text – information extraction online

Dietrich Rebholz-Schuhmann (EBI - Cambridge, GB)

Scientific literature is an essential provider of new scientific facts and hypotheses. Publications are increasingly available short after acceptance through the publisher. This talk will focus on information extraction methods available from the EBI. This includes the integration of large terminological resources into the text mining technology and linking from the text to the bioinformatics databases. As part of this talk EBI's new service EBIMed will be presented.

Keywords: Text mining, information extraction, bioinformatics, terminological resources

Joint work of: Rebholz-Schuhmann, Dietrich; Kirch, Harald; Gaudan, Sylvain

Mining Heterogenous Data with Mixture Models

Alexander Schliep (MPI für Molekulare Genetik, D)

Mixture models are an statistical framework for dealing with data which cannot easily be partitioned into groups, as it is frequently the case for biological applications. Heterogeneous, rich data as currently is available can be represented in appropriately complex models allowing for statistical queries. Time-courses models based on Hidden Markov Models are one example.

While the integration of heterogeneous data is formally quite easy, a lot of important details have to be addressed. For example, partially supervised learning allows to integrate data of wildly different abundance, for example expression values for all genes versus experimentally confirmed transcription factor binding. We report on our method and sample applications which stress the importance of judicious selection of data sets to integrate.

Identifying active transcription factors from expression data using Pathway Queries

Florian Sohler (Universität München, D)

Microarrays provide a unique way of obtaining a snapshot of a cell's state as they can measure mRNA levels on a genome-wide scale. The analysis of the obtained data still poses challenges to biologists, statisticians and computer scientists. After a basic analysis, detailed biological questions usually appear, like the following examples:

1. What is the mechanism that causes the observed regulation?
2. What is the effect of the observed regulation, e.g. on the metabolism?

To answer such biological questions, it is necessary to make use of biological background knowledge.

Pathway Queries constitute a method to extract relevant biological context according to user-specified queries from prior knowledge given in the form of networks and functional annotations. Based on this method, a scoring function is developed to identify active transcription factors or other regulators, thus making a first step toward explaining the measured expression data. Applied to a public data set the method yields results that are in good concordance with the biological literature.

Keywords: Expression data, networks, pathways, query language, gene regulation

Ontology based data analyis

Andrea Splendiani (Institut Pasteur - Paris, F)

We address the problem of the functional evaluation of high-throughput data in molecular biology, focusing on the use of ontologies in the analysis of microarray data. The microarray technology allows the measurement of the "activation" (expression) of all genes of an organism in relation to a condition or a stimulus. This is a key information to understand the dynamics of a living system. A first step to understand this is to determine how the observed patterns of gene activation are related to previous knowledge of gene functions. This cannot be done on a human-expert basis given the size and complexity of a genome-level analysis. It is common practice to associate terms describing functions, defined in a coherent ontology, to genes, and to evaluate how this terms are related to features extracted from the microarray data, like sets of coherently expressed genes or significantly expressed genes (terms enrichment). However this approach considers functions only as attributes, while in general functions can be represented as a network of related concepts. The most used ontology, gene ontology, defines a set of terms to describe gene functions, and relates these terms in a dag structure through part-of and subclass-of relations. At present, the relations among

concepts are rarely used in functional evaluation of gene expression data. This is particularly relevant for genes involved in well known mechanisms (pathways) that are described through highly structured ontologies as bioPAX. We present a software platform and strategies to enable a richer exploitation of the information content of ontologies in the analysis of gene expression data, focusing on pathway ontologies (bioPAX). We introduce some key concepts of the semantic web, a framework in which bioPAX and other ontologies are represented, then we present a software platform that unifies a graph based analysis and visualization software widely adopted in the life-science community (Cytoscape), and a library to handle ontologies in the semantic web framework (Jena). We show how our platform can be used to browse and query the content of ontologies through an intuitive interface. We present how through inference it is possible to derive new functional properties associated to genes from knowledge explicitly represented. In particular, we show how this can be used to support a user-driven research of possible functional relations.

Then we discuss how to extend some term scoring functions to more general relation scoring functions, allowing an assessment of the relevance of terms, relations and attributes describing gene functions, as compared to simple terms enrichment. Finally we discuss how this significance evaluation can be used to enrich the content of ontologies.

Keywords: Bioinformatics , ontologies, semantic web, pathways, biopax, microarrays

Semantic browsing of pathway ontologies and biological networks with RDFScape (working paper)

Andrea Splendiani (Institut Pasteur - Paris, F)

Studying biological organisms at the systems level is a complex task. Computational approaches require structured representations of existing biological knowledge. This necessity has prompted the development of formal representations of specific areas of knowledge, resulting in ontologies such as Gene Ontology and BioPAX. However, only part of this formalized knowledge is exploited for the interpretation of experimental data. Specifically, it is common to use the association between entities and annotations, like genes and functions, while the structure of the annotation is not considered beyond some common features as inheritance. This is partly due to a lack of tools and methods that bridge resources related to ontologies to the ones related to data analysis. Here we present a platform that merges a semantic web toolkit with a widely adopted modular tool for systems biology investigation. We demonstrate how in this environment it is possible to query ontologies not only as a list of annotations but as a knowledge base from which new information can be derived. We also show how this knowledge can be integrated with biological data.

Keywords: Pathways, Semantic Web, Ontologies, Microarrays

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2006/474

On some computational problems arising in RNA 3D structure determination process with NMR

Marta Szachniuk (Poznan University of Technology, PL)

Recognition of the importance of RNA in many biological processes has increased dramatically in recent years. The discoveries of non-coding regulatory RNA, RNA catalytic activity and a phenomenon of RNA interference have involved a broad line of disciplines. Consequently, we can observe growing interest in the research on RNA molecules, including structural analysis, molecular dynamics, etc. Regarding RNA functional variety as well as its quick degradation under in vitro conditions, studying the structures of these molecules proved to be more difficult than the examination of proteins and DNA. An elucidation of RNA tertiary structure in solution has become possible owing to the development of NMR spectroscopy. Structure determination procedure using NMR is composed of two general stages: experimental, where multidimensional correlation spectra are acquired and computational, where spectra are analyzed and structure is determined.

Computational part of the process starts from assignment of the observed NMR signals to the corresponding protons. This procedure, unlike other computational steps, is performed manually, thus, being a bottleneck of the RNA structure elucidation process. Therefore, it has been of a great need to facilitate NMR structural analysis of biopolymers by an introduction of automatic procedures at this level.

I have proposed theoretical model of the assignment problem and described the main aspects of its computational complexity analysis. The three algorithms for automatic signal assignment, enumerative, evolutionary and tabu search, have been shown and the results of their application to the real data presented.

Distribution of paralog families in genomes

Jerzy Tiuryn (University of Warsaw, PL)

We propose a probabilistic model of genome evolution and provide a mathematical analysis of the asymptotic distribution of paralog families in the model. It is proven that under quite general conditions the distribution is close to the logarithmic distribution. The experimental results (mostly bacterial genomes, but also several yeast species) support this theoretical prediction reasonably well.

Keywords: Genome evolution, Markov chain, asymptotic distribution, paralogs

Joint work of: Rudnicki, Ryszard; Tiuryn, Jerzy; Wojtowicz, Damian

Statistics for detecting overrepresentation of genes in Gene Ontology categories

Martin Vingron (MPI für Molekulare Genetik, D)

Gene functions are categorized in the Gene Ontology hierarchy (rather, it is a directed acyclic graph). The problem arises to determine, whether genes of a particular category are enriched among a given set of genes. To this end, traditionally a hypergeometric distribution is used to determine the significance of the overlap beween a category and the given set of genes. Here, we propose an alternative statistic. It focuses on how the genes of interest in a parent node are distributed over the child-nodes. We evaluate the alternative statistic by simulation and show a real example.

Joint work of: Grossmann, Steffen; Vingron, Martin

Expert knowledge without the expert - Automatic derivation of network contexts from expression data and the biomedical literature

Ralf Zimmer (Universität München, D)

The interpretation of expression data without appropriate expert knowledge is difficult and usually limited to exploratory data analysis such as clustering and finding differentially regulated gene sets. Thus, manual analysis by experts is required to obtain confident predictions about the involved biological processes.

We present an algorithm to derive interpretations of expression measurements together with biological hypotheses from bio-medical publications without involving manual intervention by an expert neither in specifying prior knowledge nor during the analysis procedure. The approach is applied to a juvenile arthritis expression dataset, where a number of clusters and accompanying concepts are identified as an interpretation of the data. These clusters are both more sensitive and more specific than GeneOntology categories detected on the same data.

Joint work of: Zimmer, Ralf; Küffner, Robert

Petri Nets for Bioinformatics

Ralf Zimmer (Universität München, D)

The talk gave a brief introduction to Petri Nets and their possible applications in Bioinformatics. We use Petri Nets for formal representations of biochemical networks, for formulating pathway hypotheses, for specifying queries for networks and pathways (pathway queries) and for simulating dynamical processes in those networks

Keywords: Petri nets, pathways, biochemical networks

Colored Petri net modeling and simulation of signal transduction pathways

Ralf Zimmer (Universität München, D)

The talk resented a methodology for quantitatively analyzing the complex signaling network by resorting to colored Petri nets (CPN). The mathematical as well as Petri net models for two basic reaction types were established, followed by the extension to a large signal transduction system stimulated by epidermal growth factor (EGF) in an application study. The CPN models based on the Petri net representation and the conservation and kinetic equations were used to examine the dynamic behavior of the EGF signaling pathway. The usefulness of Petri nets is demonstrated for the quantitative analysis of the signal transduction pathway. Moreover, the trade-offs between modeling capability and simulation efficiency of this pathway are explored, suggesting that the Petri net model can be invaluable in the initial stage of building a dynamic model.

Keywords: Colored Petri nets (CPN), Petri net representation, Signal transduction networks, Systems biology, Epidermal growth factor (EGF)

Joint work of: Zimmer, Ralf; Lee, Dong-Yup; Lee, Sang Yup; Park, Sunwon