The Sixth Annual BioPathways Meeting

Organized by:

Joanne Luciano
Eric Neumann
Aviv Regev
Vincent Schachter

July 29th-30th , 2004

ISMB 2004
Scottish Exhibition and Convention Centre - Boisdale Room
Glasgow
Scotland

Table of Contents

6th BioPathways Meeting Program

Day I – July 29 th		
9:00-9:15	Vincent Schachter *Genoscope & BioPathways Consortium*	Opening remarks
Plenary Session I: Network evolution		
9:15-10:00	Sarah Teichmann *Cambridge University*	Evolution and Dynamics of Transcriptional Regulatory Networks
10:00-10:30	**Coffee break**	
10:30-11:15	Csaba Pal *Eötvös, Loránd University Budapest*	Adaptive and reductive evolution of metabolic networks
11:15-12:00	Jeffrey Chuang *University of California, San Francisco*	Functional Bias and Spatial Organization of Genes in Mutational Hot and Cold Regions in the Human Genome
12:00-13:00	**Lunch (Montrose)**	
Plenary Session II: Regulatory networks		
13:00-13:45	Ruthy Hershberg *Hebrew University, Jerusalem*	Chromosomal organization is shaped by the transcription regulatory network
13:45-14:30	Tommy Kaplan *Hebrew University, Jerusalem*	*ab initio* prediction of transcription factor targets using structural knowledge
14:30-15:00	**Coffee break**	
15:00-15:45	Manuel Middendorf *Columbia University*	Prediction of genetic regulatory response using classification
15:45-16:30	**Open discussion: Network evolution and regulatory networks**	
Contributed Session I: Network Visualization		
16:30-17:00	Gary Bader *Memorial Sloan Kettering Cancer Center, NYC*	Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks
17:00-17:30	Frank Gibbons *Harvard University*	BioGraphNet: A distributed Forum for Heterogeneous Biological Networks
17:30-18:00	Joe Mellor *Boston University*	The VisANT Network and Pathway Workbench
18:00-19:00	**Open discussion: Visualization of biological networks**	

Day II – July 30 th

Plenary Session III & Contributed Session II: Metabolic networks		
9:00-9:20	Michele Green *SRI*	Pathway Tools Software for Creation and Analysis of Pathway/Genome Databases
9:20-9:40	Lars Juhl Jensen *EMBL, Heidelberg*	STRING: Predicting novel metabolic pathways through the integration of diverse genome-scale data
9:40-10:00	Ron Pinter *Technion, Haifa*	A New Tool for the Alignment of Metabolic Pathways
10:00-10:30	**Coffee break**	
10:30-11:15	Tomer Shlomi *Tel Aviv University*	Constraint-based modeling of perturbed organisms : ROOM with a view
11:15-12:00	Peter Kharchenko *Harvard University*	Gap-filling in metabolic networks using expression information
12:00-13:00	**Lunch (Dunkeld)**	
Contributed/Plenary Session II: Protein-protein interactions		
13:00-13:45	Debra Goldberg *Harvard University*	Topology classification for biological networks using maximum likelihood estimation and model selection
13:45-14:30	David Reiss *Institute for Systems Biology*	Predicting protein-protein interactions via a network-based motif sampler
14:30-15:00	**Coffee break**	
15:00-15:30	**Open discussion: Metabolic and protein-protein interaction networks**	
Contributed Session III: Database and exchange languages		
15:30-15:50	Michael Hucka *California Institute of Technology*	Software Infrastructure for SBML, from SBML.org (A Roundup of Software Tools for SBML)
15:50-16:10	Gary Bader *Memorial Sloan Kettering Cancer Center, NYC*	BioPAX - Biological Pathway Data Exchange Format
16:10-16:30	Geeta Joshi-Tope *Cold Spring Harbor Laboratory*	The Reactome Project
16:30-16:50	Ugur Dogrusoz *Bilkent University, Ankara, Turkey*	The PATIKA Project
16:50-17:10	Susie Stephens *Oracle Corp.*	Graph Modeling and Analysis in Oracle DBs
17:10-17:30	Kawashima Shuichi *Kyoto University*	KGML (KEGG Markup Language) for Exchanging the KEGG Graph Objects
17:30-19:00	**Panel discussion:** **Sharing and integration of pathway information**	

Invited Talks

Evolution and Dynamics of Transcriptional Regulatory Networks

Sarah Teichmann
MRC Laboratory of Molecular Biology, Cambridge, England.

The biological characteristics of an organism emerge, in large part, as result of the dynamic inter-play between its gene repertoire and the regulatory apparatus, which includes transcription factors and signal transducers. We have analyzed the possible evolutionary histories and the dynamics of transcriptional regulatory systems from a network perspective. For the transcriptional regulatory network of the prokaryote *E. coli* and the unicellular eukaryote *S. cerevisiae*, we have defined possible duplication scenarios for generating new regulatory interactions between transcription factors and target genes. We show that over one third of the regulatory interactions have evolved by duplication of transcription factors and target genes followed by inheritance of interactions from the ancestral gene. This mechanism is not primarily responsible for the overall network topology or specific topologies of network motifs. Next we turn to network evolution across prokaryotes based on the *E. coli* transcriptional regulatory network as a template, and ask whether regulatory interactions are conserved between organisms. Target genes tend to be more conserved than transcription factors, and there is no bias for conservation of network motifs.

The transcriptional regulatory network has changed in evolution, and it also changes within an organism depending on the external and internal conditions of the cell. We have studied the dynamics of the transcriptional regulatory network in *Saccharomyces cerevisiae* on a genomic scale, by integrating regulatory information and gene-expression data for multiple conditions. Contrary to expectation, we uncover large changes in underlying network architecture between different states. A few TFs serve as permanent hubs while most act transiently during particular conditions. Looking at sub-network structures, we show environmental responses facilitate fast signal propagation (*eg* with short regulatory cascades), whereas the cell cycle and sporulation direct temporal progression through multiple stages (*eg* with dense TF inter-regulation). With these studies, we are shedding light on the ways in which network topology is shaped by evolutionary mechanisms and cellular conditions.

Adaptive and reductive evolution of metabolic networks

Csaba Pál

MTA, Theoretical Biology Research Group, Eötvös Loránd University, Pázmány Péter Sétány 1/C, Budapest H-1117, Hungary.

Under laboratory conditions 80% of yeast genes seem not to be essential for viability. This raises the question of what the mechanistic basis for dispensability is, and whether it is the result of selection for buffering or an incidental side product. Dispensable genes might be important, but under conditions not yet examined in the laboratory. Using an metabolic network approach, we show that this is the dominant explanation for apparent dispensability, accounting for 37–68% of dispensable genes, whereas 15–28% of them are compensated by a duplicate, and only 4–17% are buffered by metabolic network flux reorganization. Gene duplicates catalysing the same reaction are not more common for indispensable reactions, suggesting that the reason for their retention is not to provide compensation. Instead their presence is better explained by selection for high enzymatic flux. Using a similar model, we show that reductive evolution of endosymbiotic bacteria is dominated by contingent loss of enzymatic pathways. This result suggests that multiple minimal metabolic reaction sets might exist.

Functional Bias and Spatial Organization of Genes in Mutational Hot and Cold Regions in the Human Genome

Jeffrey Chuang

UC San Francisco - Dept. of Biochemistry and Biophysics
jchuang@genome.ucsf.edu

Genome-wide knowledge of mutation rates can be useful for understanding the evolution of both pathway components and how they are regulated. In human, mouse, and rat, the neutral mutation rate is known to vary widely along chromosomes, and these regional biases are important for the evolution of genes. We provide evidence that categories of functionally-related genes reside preferentially in mutationally hot or cold regions, the size of which we measure. Genes in hot regions are biased toward extra-cellular communication (surface receptors, cell adhesion, immune response, etc.) while those in cold regions are biased toward essential cellular processes (gene regulation, RNA processing, protein modification, etc.). From a selective perspective, this organization of genes could minimize the mutational load on genes that need to be conserved and allow fast evolution for genes that must frequently adapt.

We find that mutation rates in yeast, in contrast, are uniform genome-wide.

We report a method that uses the neutral mutation rate to significantly improve phylogenetic footprinting -- the detection of functional sequences by their conservation across species -- for transcription factor binding sites. The method allows us to estimate the total amount of sequence under purifying selection in yeast promoters. Certain Gene Ontology groupings of genes (e.g. carbohydrate metabolism) have large amounts of highly conserved sequence in their promoters, suggesting complexity in their transcriptional regulation. Others (e.g. RNA processing) have less conserved sequence and are likely to be simply regulated.

Chromosomal organization is shaped by the transcription regulatory network

Ruth Hershberg[(1)], Esti Yeger-Lotem[(1),(2)], and Hanah Margalit[1]

(1) Department of Molecular Genetics and Biotechnology, Faculty of Medicine, The Hebrew University, Jerusalem 91120, Israel
(2) Department of Computer Science, Technion, Haifa 32000, Israel

Transcription regulation, a key step in the control of gene expression, has been the focus of several large-scale studies, yet little attention has been given to its relation to the chromosomal arrangement of transcription units. We study this relationship systematically in Escherichia coli and Saccharomyces cerevisiae using methodologies for network analysis. Our analysis reveals links between transcription regulation and chromosomal organization, suggesting that in both organisms transcription regulation has shaped the organization of transcription units on the chromosome. Differences found between the organisms reflect the inherent differences in transcription regulation between pro- and eukaryotes.

ab initio prediction of transcription factor targets using structural knowledge

Tommy Kaplan (tommy@cs.huji.ac.il)

School of Computer Science & Engineering - The Hebrew University
Jerusalem, 91904, Israel

Current approaches for identification and detection of transcription factor binding sites rely on an extensive set of known target genes. Here we describe a novel structure-based approach applicable to transcription factors with no prior binding data. Our approach combines sequence data and structural information to infer context-specific amino acid-nucleotide recognition preferences. These are used to predict binding sites for novel transcription factors from the same structural family. We demonstrate our approach on the Cys2His2 Zinc Finger protein family, and show that the learned DNA-recognition preferences are compatible with experimental results. We apply these preferences in a genome-wide scan for direct targets of Drosophila melanogaster Cys2His2 transcription factors. By analyzing the predicted targets along with gene annotation and expression data we infer the function and activity of these proteins.

Prediction Of Genetic Regulatory Response Using Classification

Manuel Middendorf, Anshul Kundaje
Columbia University

Inferring the components and structure of gene regulatory networks from high throughput genomic data -- like gene expression data -- has become one of the central problems in computational biology. However, most machine learning approaches to this problem have been essentially descriptive in nature. That is, the outcome of learning is a descriptive model -- such as a graph structure consisting of putative regulatory "edges" between genes, or a set of clusters of potentially co-regulated genes.

We present a new methodology for learning predictive models of gene regulation from gene expression and regulatory sequence data for simple organisms like yeast. The core of our approach is a novel algorithm called GeneClass based on boosting. While descriptive models such as probabilistic graphical models focus on finding structure in the training data, our method is able to make accurate predictions about which genes will be up- or down-regulated in new or held-out experiments (test data). We show how to use GeneClass to identify biologically important regulators and binding motifs for specific regulatory pathways. We can also use Geneclass to study gene specific and condition specific regulation. In this way, we anticipate using predictive models for gene regulatory response to generate biological hypotheses for wet lab testing.

Constraint-based modeling of perturbated organisms : ROOM with a view.

Tomer Shlomi(1), Omer Berkman(2) and Eytan Ruppin(1)
(1) School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel
(2) The Academic College of Tel-Aviv Yaffo, Tel-Aviv, Israel

In the ISMB conference (2004) we describe Regulatory On-Off Minimization (ROOM), a constraint-based algorithm for predicting the behavior of metabolic networks in response to gene knockouts. Going forward we present "ROOM with a view": (i) We present a quantitative score for the linearity of flow in a predicted flux distribution, based on recent findings by Ihmels et al. (2003), showing significantly higher scores for ROOM predictions. (ii) Extending Mahadevan et al. (2003), we present an in-depth treatment of the effect of alternative FBA solutions for the wild-type organism on the accuracy of flux predictions for the knocked-out organism. (iii) Inspecting inaccurate growth rate predictions obtained by Minimization Of Metabolic Adjustment (MOMA) (Segre et al., 2002), we suggest a variant of MOMA that imposes constraints on growth rates and results in improved flux predictions accuracy. (iv) ROOM is based on Mixed Integer Linear Programming (MILP) which is computationally hard. Relaxing integer constraints results in an alternative metric for measuring the distance between the flux distributions of the wild-type and knocked-out organisms, which can be modeled using Linear Programming (LP). Predictions based on this alternative metric correlate well with experimental results.

Gap-filling in metabolic networks using expression information

Peter Kharchenko, Dennis Vitkup
Department of Genetics, Harvard Medical School

The metabolic models of both newly sequenced and well-studied organisms contain reactions for which the enzymes have not been identified yet. We present a computational approach for identifying genes encoding such missing metabolic enzymes in a partially reconstructed metabolic network.

The metabolic expression placement (MEP) method relies on the coexpression properties of the metabolic network and is complementary to the sequence homology and genome context methods that are currently being used to identify missing metabolic genes. The MEP algorithm predicts over 20% of all known Saccharomyces cerevisiae metabolic enzyme-encoding genes within the top 50 out of 5594 candidates for their enzymatic function, and 70% of metabolic genes whose expression level has been significantly perturbed across the conditions of the expression dataset used.

Topology classification for biological networks using maximum likelihood estimation and model selection

Debra S. Goldberg
Harvard Medical School

The large-scale topological structure of a network is thought to provide clues to design principles underlying the development of the network, such as evolutionary constraints or biases of the experimental method used to determine the network. The determination of the form of the degree distribution has often been done by fitting the distribution to a certain class of function and observing that the deviation is small. I will present a principled approach to selecting a model for the degree distribution of a network based on a maximum likelihood analysis.

Four basic models from which the degree distribution can be drawn are considered: power-law (scale-free), exponential, Poisson, and truncated power-law. True positive edges and false positive edges may be drawn from different underlying distributions, so we consider combination models that are the convolution of two basic degree distribution models. We fit these maximum likelihood models to the observed degree distribution of a network, then objectively compare models using the Bayesian Information Criterion.

We applied our method to protein interaction networks from S. cerevisiae, C. elegans and D. melanogaster. Our programs have been incorporated into a tool that will soon be freely available. This is joint work with Giovanni Franklin and Fritz Roth.

Predicting protein-protein interactions via a network-based motif sampler

David Reiss
Institute for Systems Biology

Many protein–protein interactions are mediated by peptide recognition modules (PRMs), compact domains that bind to short peptides, and play a critical role in a wide array of biological processes. Recent experimental protein interaction data provide us with an opportunity to examine whether we may explain, or even predict their interactions by computational sequence analysis. Such a question was recently posed by the use of random peptide screens to characterize the ligands of one such PRM, the SH3 domain.

We describe a general computational procedure for identifying the ligand peptides of PRMs by combining protein sequence information and observed physical interactions into a simple probabilistic model and from it derive an interaction mediated de novo motif-finding framework. Using a recent all-versus-all yeast two-hybrid SH3 domain interaction network, we demonstrate that our technique can be used to derive independent predictions of interactions mediated by SH3 domains. We show that only when sequence information is combined with such all versus all protein interaction datasets, are we capable of identifying motifs with sufficient sensitivity and specificity for predicting interactions. The algorithm is general so that it may be applied to other PRM domains (e.g. SH2, WW, PDZ).

Contributed Talks

Cytoscape: a software environment for integrated models of biomolecular interaction networks

Andrew Markiel, David Reiss, Iliana Avila-Campillo, Larissa Kamenkovich, Paul Shannon, Rowan Christmas, Hamid Bolouri

Institute for Systems Biology, Seattle, Washington 98103, USA Benno SchwikowskiPasteur Institute, Paris, France

Chris Workman, Dan Ramage, Jonathan Wang, Nada Amin, Owen Ozier, Trey Ideker

UCSD Department of Bioengineering, La Jolla, California 92093-0412, USA

Ethan Cerami, Rob Sheridan, Gary D. Bader, Chris Sander

Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY, 10021, USA

Cytoscape (http://www.cytoscape.org) is an open source software project for integrating biomolecular interaction networks with high-throughput expression data and other molecular states into a unified conceptual framework. Although applicable to any system of molecular components and interactions, Cytoscape is most powerful when used in conjunction with large databases of protein-protein, protein-DNA, and genetic interactions that are increasingly available for humans and model organisms. Cytoscape's software Core provides basic functionality to layout and query the network; to visually integrate the network with expression profiles, phenotypes, and other molecular states; and to link the network to databases of functional annotations. The Core is extensible through a straightforward plug-in architecture, allowing rapid development of additional computational analyses and features. Cytoscape provides a powerful visualization engine as well as a number of plug-ins for network-based gene expression analysis, network clustering and network homology detection. The recently released version 2.0 of Cytoscape and the current set of analysis plugins will be discussed.

BioGraphNet, a distributed forum for heterogeneous biological networks

Francis D. Gibbons, Gabriel F. Berriz, and Frederick P. Roth

Dept. of Biological Chemistry and Molecular Pharmacology, 250 Longwood Ave., Harvard Medical School, Boston MA 02115, USA

Biological network information is increasingly abundant. The combination of biological networks may be viewed as a multicolor graph, each color representing a different gene-gene or protein-protein relationship, e.g., protein interaction, sequence homology, correlated expression, transcriptional regulation, genetic interaction (sensu synthetic lethality), or metabolic relationship. Relationship types may be further stratified by type of supporting evidence, by directionality or by confidence measure. Furthermore, each organism has its own collection of networks. Although this information's complexity argues for its maintenance by distributed groups, much of its value is derived through network integration.

BioMOBY has established a 'playground' for distributed services, within which we have developed a 'sandbox' called BioGraphNet. BioGraphNet is a common standard and collection of services for sharing distributed network information. We serve several network data types, and encourage others to participate, using objects we have registered in BioMOBY's ontology.

BioTrawler (http://llama.med.harvard.edu/cgi/BioTrawler), our web-based biological network browser, illustrates the use of BioGraphNet. It dynamically discovers suitable distributed data sources within BioGraphNet, integrates selected sources 'just in time', and visualizes the graph neighboring a user-defined set of genes.

The combination of BioMOBY and BioGraphNet represents a distributed network annotation system analogous to the Distributed Annotation System (DAS) for sharing genome annotation.

The VisANT Network and Pathway Workbench

Joe Mellor

Boston University

Tools that help to integrate and analyze the abundant forms of new biological data are important in creating novel understanding of biological systems. Networks are a natural paradigm for describing many biological systems: protein interactions, gene regulation, biochemical pathways, where the underlying living behavior is in a large sense encoded by how pieces are connected. In addition, meta-networks of grouped genes and proteins – for example, regulatory motifs, pathways, expression modules and macromolecular complexes – are layers of information that compound the challenging problem of describing different systems in biology. We've developed the VisANT software tool as a general workbench application for manipulating, analyzing and visualizing biological network information obtained from many different sources. VisANT is a Java-based web tool that can be used for viewing interaction data combined and overlaid with annotation from Gene Ontology, GenBank, SwissProt and KEGG. VisANT is integrated with a database backend for storing and retrieving user-defined interaction data sets, which can then be combined with public data or any amount of visual annotation. The long-range goal of the VisANT project is to integrate different efforts of biological systems research and facilitate the sharing and visualization of newly published data from other sources. VisANT is an open-source project, and is available at http://visant.bu.edu.

Pathway Tools Software for Creation and Analysis of Pathway/Genome Databases

M. Green, P. Karp, S. Paley, J. Pick
SRI (Standford Research Institute)

This presentation will discuss new developments in the Pathway Tools software, which provides an environment for creation, editing, analysis, and Web publishing of organism-specific Pathway/Genome Databases (PGDBs). 70 groups have thus far created close to 40 PGDBs, with 20 more planned.

Given an annotated genome as its input, Pathway Tools creates a new PGDB that describes the genes, replicons, and proteins described in the input file. The software computationally predicts the operons (new) and the metabolic pathways of the organism. A collection of editing tools allows users to update the gene functions, metabolic pathways, and genetic network of the organism. A new ontology of evidence codes allows a PGDB to clearly distinguish computational predictions from experimentally derived information [2]. Users can query PGDB data using new Perl and Java APIs (application program interfaces).

A new pathway hole filling algorithm identifies missing enzymes in an organism's genome to fill pathway holes in the PGDB. A pathway hole occurs when a genome appears to lack the enzyme catalyzing a reaction in a pathway for which other enzymes are present. By identifying enzymes that catalyze these pathway holes, we can not only increase the utility of PGDBs, but also improve the annotation of the corresponding genome.

References:
[1] Karp, P.D., Paley, S. and Romero, P., "The Pathway Tools Software," Bioinformatics 18:S225-32 2002.
[2] Karp, P.D., Paley, S., Krieger, C.J., and Zhang, P., "An Evidence Ontology for use in Pathway/Genome Databases," Proceedings of the Pacific Symposium on Biocomputing 2004.
[3] Green, M.L. and Karp, P.D., "A Bayesian method for identifying missing enzymes in predicted pathway/genome databases," BMC Bioinformatics 5(1): 76.

STRING: Predicting novel metabolic pathways through the integration of diverse genome-scale data

Lars Juhl Jensen
EMBL

The proteins involved in a particular metabolic pathway (or other biological process) can often be predicted. The web-based database STRING (http://string.embl.de) is a large, pre-computed resource that links together functionally associated proteins by integrating a variety of information sources - including de novo predictors. For more than 100 genomes, the tool systematically evaluates i) the conserved chromosomal proximity of genes, ii) gene fusion events, iii) correlations in evolutionary patterns (co-occurrence), iv) large-scale protein interaction data sets, v) gene co-expression, vi) co-mentioning of genes in published literature, and vii) functional annotations in specific databases. Together, these sources of information allow the construction of a high-confidence network of known and predicted functional interactions between proteins, from which functional modules can be extracted using clustering techniques. These cover both novel pathways and allow the extension of known pathways with additional members.

A New Tool for the Alignment of Metabolic Pathways

Ron Y. Pinter, Oleg Rokhlenko, Esti Yeger-Lotem, Michal Ziv-Ukelson

Dept. of Computer Science, Technion - Israel Institute of Technology, Haifa 32000, Israel {pinter, olegro, estiy, michalz}@cs.technion.ac.il

Several genome-scale efforts are underway to reconstruct metabolic networks for a variety of organisms. As the resulting data accumulates, the need for analysis tools increases. A notable requirement is a pathway alignment finder that enables both the detection of conserved metabolic pathways among different species as well as divergent metabolic pathways within a specie. When comparing two pathways, the tool should be powerful enough to take into account both the pathway topology as well as the nodes' labels (e.g. the enzymes they denote), and allow flexibility by matching similar — rather than identical — pathways.

MetaPathwayHunter is a pathway alignment tool that, given a query pathway and a collection of pathways, finds and reports all approximate occurrences of the query in the collection, ranked by similarity and statistical significance. It is based on novel, efficient graph matching algorithms that extend the functionality of known techniques. The program also supports a visualization interface with which the alignment of two homologous pathways can be graphically displayed.

We employed this tool to study the similarities and differences in the metabolic networks of several organisms (as represented in highly curated databases that are available on the World Wide Web). We reaffirmed that most known metabolic pathways common to bacteria and yeast are conserved; furthermore, we present a few cases in which the comparison of metabolic pathways between species exemplifies divergent and putative convergent evolution, and within a specie — exemplifies divergent evolution. We conclude with a description of several biologically meaningful meta-queries, demonstrating the power and flexibility of our new tool in the analysis of metabolic pathways.

Software Infrastructure for SBML from SBML.org

Michael Hucka, Andrew Finney, Benjamin J. Bornstein, Bruce E. Shapiro, Sarah M. Keating, Akira Funahashi, Maria J. Schilstra, Benjamin L. Kovitz, Joanne Matthews

California Institute of Technology

The Systems Biology Markup Language (SBML) is a widely-used XML-based exchange format for computational models of biochemical networks. As part of the SBML project, we have been developing software infrastructure to help developers and modelers work with and use SBML. In this presentation, I will summarize the following free software tools we make available today as resources for the computational systems biology community: libSBML, a software library providing C, C++, Java, and Python interfaces for reading and writing

SBML files and data streams; MathSBML, a Mathematica package for working with models in SBML format; SBMLToolbox, a Matlab package for working with SBML models; KEGG2SBML, a tool for converting KEGG metabolic models to SBML; CellML2SBML, a tool for converting CellML models to SBML; and the online facilities available on our website. More information about all of these is available at http://sbml.org.

BioPAX - Biological Pathway Data Exchange Format

BioPAX Workgroup (Gary D. Bader, Michael P. Cary, Erik Brauner, Robert Goldberg, Chris Hogue, Peter Karp, Teri Klein, Joanne Luciano, Debbie Marks, Natalia Maltsev, Eric Neumann, Suzanne Paley, John Pick, Aviv Regev, Andrey Rzhetsky, Vincent Schachter, Imran Shah, Jeremy Zucker, Chris Sander)
Memorial Sloan Kettering Cancer Center

BioPAX (http://www.biopax.org) is a community-based effort to develop a technical recommendation for a biological pathway data exchange format. Level 1, recently released, supports metabolic pathway data and is implemented in OWL. The BioCyc, WIT and PharmGKB databases are committed to making their pathway data available in this format. Subsequent releases of BioPAX will add support for protein-protein interactions (e.g. PSI-MI compliant databases), signal transduction pathways (e.g. TransPath, BioCarta, STKE), genetic interactions (e.g BIND, GRID), transcription and translational regulation (e.g. TRANSFAC) and other pathway data types.

The BioPAX data exchange format should eventually be able to represent most of the data available in the over 100 pathway related data resources currently in existence (http://www.cbio.mskcc.org/prl). In order to provide a useful format as quickly as possible, BioPAX is being developed in a leveled approach in which each level will support a greater variety of pathway data.

The BioPAX workgroup is coordinating their work with that of other pathway related standards initiatives, such as SBML, CellML, and PSI-MI, with the intent of designing BioPAX so that it is compatible with these standards in the areas where they overlap. The BioPAX format and any associated software developed by the BioPAX group are open source and freely available to all under the GNU LGPL license.

The Reactome project

Lincoln Stein, Geeta Joshi-Tope, Gopal Gopinath, Guanming Wu, Marc Gillespie, Peter D'Eustachio, Lisa Matthews , Adrian Arva, Marcela Tello-Ruiz (alumnus)
Cold Spring Harbor Laboratory

Ewan Birney, Imre Vastrik, Esther Schmidt, Bijay Jassal, Bernard de Bono
European Bioinformatics Institute

Suzanna Lewis
Gene Ontology Consortium

Reactome (formerly Genome Knowledgebase) is a curated database of biological processes in humans, with cross-references to PubMed, SwissProt, LocusLink, Ensembl, and GO. It is a dual-purpose project, usable by the general biologist as an online biology textbook, or by bioinformaticists for data mining and for analysis of pathways. While Reactome is targeted at human pathways, it also includes electronically inferred reactions from other organisms, presently rat, mouse, pufferfish and zebrafish. This makes the database relevant to researchers who work on model organisms. All the information in Reactome is backed up by its provenance: either a literature citation or an electronic inference based on sequence similarity. Our data model is event-centric; the basic unit of Reactome is a reaction.

The website (http://www.reactome.org) sports a javascript powered user interface in which the entire set reactions known to the database are represented as a series of constellations in a "starry sky." This serves as a navigation tool, and a visualization tool for pathways known to the database. Reactome provides a qualitative framework, on which quantitative data can be superimposed. Interfaces for such overlays are in development.

Our data and software are distributed under an open source license that allows unrestricted reuse and redistribution.

The PATIKA Project

E. Demir, U. Dogrusoz, O. Babur, A. Ayaz, E. Giral, Z. Erson, G. Nisanci and G. Gulesir
Bilkent University, Ankara, Turkey

Lately, there has been an enormous amount of effort for creating ontologies, standards and tools to facilitate modeling and analysis of biological pathways. Still, current bioinformatics infrastructure is far from coping with inherent complexity of data produced by state-of-the-art biology techniques. The PATIKA Project aims to provide scientific community with an integrated environment composed of a central database and a visual editor, built around an extensive ontology and an integration framework. It also features tools for analyzing microarray data and inference of pathways.

PATIKA has been designed mainly for distributed research communities, who wish to build large-scale integrated pathway models in a collaborative manner. Users may retrieve desired parts of the model stored in the database via a querying interface for standard and graph based queries. Pathways constructed from query results or built from scratch by the user can be modified and submitted to the server. These changes are then integrated back into the central model, after resolving potential conflicts including those due to concurrent modifications of others.

In this talk, we will give an overview of PATIKA project and discuss latest improvements to its ontology, including bioentity level interactions and attributes and a better structured cell model. We will also present features of our new editor, PATIKApro, such as ability to represent varying levels of abstractions through nested pathway drawings, multiple pathway views, facilities for analysis of gene expression data including a new method for inference of pathway activity using gene expression data. Finally we will talk about our efforts for populating the PATIKA database by automated pathway inference.

http://www.patika.org

Graph Modeling and Analysis in Oracle Database 10g

Dr. Susie Stephens
Principal Product Manager, Life Sciences, Oracle Corporation

The Oracle Spatial Network Data Model (NDM) feature enables graph modeling and analysis in Oracle Database 10g. NDM explicitly stores and maintains connectivity (nodes, links, and paths) within networks and provides network analysis capability such as shortest path and connectivity analyses. NDM includes a PL/SQL API Package for Network Data Query and management, and a Java API for network representation, network analysis, and network creation
and editing. This presentation will give an overview of the architecture of NDM. Examples as to how customers have used NDM to represent and analyze biochemical pathways and protein-protein interaction data will also be given.

KGML (KEGG Markup Language) for Exchanging the KEGG Graph Objects

Shuichi Kawashima, Susumu Goto, Toshiaki Katayama, Mika Hirakawa, Minoru Kanehisa

Kyoto University

The KEGG pathway maps are graphical image maps representing networks of interacting molecules responsible for specific cellular functions. There are two types of KEGG pathways:

- reference pathways which are manually drawn and
- organism-specific pathways which are computationally generated based on reference pathways.

The KEGG Markup Language (KGML) is an exchange format of the KEGG graph objects, especially the KEGG pathway maps that are manually drawn and updated. KGML enables automatic drawing of KEGG pathways and provides facilities for computational analysis and modeling of protein networks and chemical networks.

In KGML the pathway element specifies one graph object with the entry elements as its nodes and the relation and reaction elements as its edges. The relation and reaction elements indicate the connection patterns of rectangles (gene products) and the connection patterns of circles (chemical compounds), respectively, in the KEGG pathways. The two types of graph objects, those consisting of entry and relation elements and those consisting of entry and reaction elements, are called the protein network and the chemical network, respectively. Since the metabolic pathway can be viewed both as a network of proteins (enzymes) and as a network of chemical compounds, another distinction of KEGG pathways is:

- metabolic pathways viewed as both protein networks and chemical networks and
- regulatory pathways viewed as protein networks only.

Each KGML file may be acquired from the KGML top page:
http://www.genome.ad.jp/kegg/xml/

Contributed Abstracts

BioUML - open source extensible workbench for systems biology

Fedor Kolpakov
Design Technological Institute of Digital Techniques, Novosibirsk, Russia

BioUML - Biological Universal Modeling Language - is open source extensible Java workbench for systems biology.

BioUML's core is a meta model that provides an abstract layer for comprehensive formal description of wide range of biological and other complex systems.
Content of databases on biological pathways, SBML and CellML models can be expressed in terms of the meta model and used by BioUML workbench. Plug-in based architecture provides the workbench extensibility and possibility of seamless integration with other tools for systems biology (there are plug-ins for integration with MATLAB and SBW/SBML). The workbench consists from Eclipse platform runtime that supports plug-ins registry and a set of plug-ins for database access, diagram editing and biological systems simulation. Query engine plug-in allows user to find interacting components of biological systems and show them as a graph that can be edited by a user.
The module concept allows the developer to define new diagram types and incorporate databases on biological pathways into the workbench. There are modules for GeneNet, TRANSPATH and KEGG/pathways databases. BioUML technology is also used for development of new databases: Cyclonet - database on cell cycle regulation and database on molecular mechanisms of chronic lung diseases.
Availability: http://www.biouml.org.

ELM - The Eukaryotic Linear Motif resource

EMBL Biocomputing Unit (linding@embl.de)

In the post-genomic era, analysis of cellular regulatory networks and systems are of increasing importance. Yet we do not know the role of the topologies, modules and different regulatory protein interaction networks play in a future complex model of cellular systems. Hitherto, protein function models have been primarily described in terms of functional modules known as globular domains.

However an only now catalogued large group of functional sites is found primarily in unstructured parts of proteins. These linear system modules encompass ligand sites such as 14-3-3, SH3 and Cyclin ligands as well as post-translational modification sites and targeting signals. This work describes the first comprehensive and proteomic level analysis of such system modules.

Linear system modules are short and co-linear in both sequence and structure space, i.e. they behave as linear peptide motifs, which make them difficult to detect from sequence alone. Experimentally they are often neglected because they reside in unstructured parts of proteins which are often recombinantly removed during protein expression. We have created the largest and most comprehensive computational resource:

The Eukaryotic Linear Motif resource, ELM [http://elm.eu.org], for finding these functional sites. The ELM resource is knowledge based and stores contextual profiles for linear functional sites annotated from the scientific literature. Contextual profiles and logical filters reduce the overprediction of short linear motifs. Linear system modules are as important for protein function models as are globular domains and we show how one can infer novel functional networks by considering both types as modules. We propose how they can be integrated into the modular model of protein function. Several proteome interaction datasets are analysed for interactions between a linear and globular module, to show molecular details about verified protein interactions.

Differential Network Expression During Drug and Stress Response

Laurence Cabusora Harvard University
Andy Fulmer Procter & Gamble
Christian V. Forst Los Alamos National Laboratory

The application of microarray chip technology has led to an explosion of data concerning the expression levels of the genes in an organism under a plethora of conditions. One of the major challenges of systems biology today is to devise generally applicable methods of interpreting this data in a way that will shed light on the complex relationships between multiple genes and their products. The importance of such information is clear, not only as an aid to arenas of research like drug design, but also as a contribution to our understanding of the mechanisms behind an organism's ability to react to its environment.

We detail one computational approach to using gene expression data to identify response networks in an organism. The method is based on the construction of biological networks given different sets of interaction information and the reduction of said networks to important response sub- networks via integration of the gene expression data. As an application, expression data of known stress responders and DNA repair genes in M. tuberculosis is used to construct a generic stress response network. This is compared to similar networks constructed from data obtained from subjecting M. tuberculosis to various drugs; we are thus able to distinguish between generic stress response and specific drug response. We anticipate that this approach will be able to accelerate target identification and drug development for tuberculosis in the future.

GenMAPP and MAPPFinder 2.0: Tools for Viewing and Analyzing Genomic Data Using Gene Ontology and Biological Pathways

Kam D. Dahlquist, Scott W. Doniger, Nathan Salomonis, Kristina Hanspers, Karen Vranizan, Lynn Ferrante, Alexander C. Zambon, Jeff C. Lawlor, Steven C. Lawlor, and Bruce R. Conklin

GenMAPP (Gene MicroArray Pathway Profiler) is a free, stand-alone program designed for viewing and analyzing genomic data on MAPPs representing biological pathways and other functional grouping of genes. A MAPP is produced with the graphics tools in GenMAPP and depicts the biological relationship between genes or gene products. MAPP files are small database files that store identifiers for genes and vector coordinates for all objects on the MAPP. GenMAPP automatically color-codes the genes on the MAPP according to data and criteria supplied by the user. The accessory program MAPPFinder relates the gene expression data and the user’s criterion for a meaningful gene expression change to the Gene Ontology hierarchy, returning a list of biological processes, cellular components, and molecular functions overrepresented in the data. Improved features in GenMAPP and MAPPFinder 2.0 include an expanded and flexible gene database and the ability to export sets of MAPPs to HTML for display on web sites. Pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) have now been translated into the GenMAPP MAPP format. Future plans for pathway exchange will be discussed. GenMAPP and MAPPFinder are available at http://www.GenMAPP.org.

GEST: a pathway editor for hierarchical structures

Ken Ichiro Fukuda

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST)

Toshihisa Takagi

Graduate School of Frontier Sciences, University of Tokyo

As functional-genomics data become available in an ever increasing rate, proper accumulation and dissemination of background knowledge about biological functions of higher order, such as signal transduction pathways, become indispensable.
Although graph-analogical representations are typically observed for pathways in the biological literature, a simple graph definition is not sufficient to annotate these hand-drawn pictures or descriptions in the form of free text. Due to the diversity of topics that the term ``signal'' covers, the constituents of biological pathways are highly diverse. Besides, implicit sub-structures are typically embedded in a pathway diagram (e.g., MAPK cascade). And another diagram may refer the sub-structure as a single bio-process (e.g., “activates the MAPK cascade”).

To support the accumulation processes of this type of knowledge, a pathway editor that displays and manipulates heterogeneous pathway components and enables the user to annotate implicit substructures of pathways is necessary.

This presentation describes a fully operational implementation of a compound graph based pathway editor for textual knowledge based curation.

The system is designed to support curation tasks for signal transduction pathway and other functional knowledge that is buried in biological literatures. The system supports manipulation of attributed compound graphs. The software should be available to the open public from http://www.inoh.org/GEST/.

The Stochastic Master Equation for modeling intracellular processes

H.S. Booth, C. Burden, M. Hegland, L. Santoso and S.R. Wilson

We investigate approaches to modeling genetic regulatory networks in the context of cellular functions. We propose a framework in which the system is viewed in terms of states and transitions. A state may comprise, for example, the species as index to the state vector, the number of molecules and their physical locations. At any given state, there is a set of propensities pertaining to a set of state transitions. This represents the likelihood that transcriptions, translations, reactions, or bonding, to name a few examples, would occur. The expression of the change of the probability that at time t, the system is at state x, is what we call the Stochastic Master Equation (SME). The stochastic and discrete state nature of the SME accounts for its appropriateness for a system where concentrations of species are low, and molecular processes are not deterministic. We examine how approximations of this equation lead to other frameworks. For example, considering only the binary variables (e.g. gene on/off) leads to Boolean networks, and taking the average of the distributions leads to continuous kinetic rate equations, placing the Master Equation in the centre of the framework.

Representation and Querying of Biological Pathways as Graphs

Barbara A. Eckman (1), Paul G. Brown (2), Julia E. Rice (2)

(1) IBM Life Sciences, 1475 Phoenixville Pike, West Chester, PA USA 19380 baeckman@us.ibm.com
(2) IBM Almaden Research Center, 650 Harry Road, San Jose, CA USA 95120 pbrown1@us.ibm.com, julia@almaden.ibm.com

As high-throughput biology generates large volumes of data about the "parts list" of living organisms, there is a growing need for robust, efficient systems to represent and manage such structures as metabolic and signaling pathways, gene regulatory networks, and protein interaction networks. Pathway data frequently is best represented as graphs, and researchers need to navigate and manipulate this data in ways not well supported by standard database tools.

We present a prototype from IBM Research to extend DB2, IBM's RDBMS, with graph objects and operations to support systems biology. Graph operations are executed as function calls from within SQL queries. In a federated database environment, graph operations may be applied to data stored in any format (remote or local, relational or non-relational). Supported operations include neighborhood queries, shortest path queries, spanning trees, graph transposition, and subgraph isomorphism.

Real-world examples demonstrate the usefulness of this approach: "To disrupt the activity of disease-related protein A, find all proteins upstream of A within a path length k, considering only a certain type of interaction, whose weights/confidence values exceed a certain threshold." "Predict pathways in Halobacterium based on phylogenetic fingerprinting, domain fusion, genome co-location, and/or protein-protein interactions between orthologous proteins in yeast, E. coli, etc."

LacplantCyc: a reference Pathway Database for Lactic Acid Bacteria with *Lactobacillus plantarum* WCFS1 as model

Frank H.J. van Enckevort1,2, Christof Francke1,3, Bas Teusink2,3 and Roland J. Siezen1,2,3 (Frank.van.Enckevort@cmbi.kun.nl)

(1) Centre for Molecular and Biomolecular Informatics, University of Nijmegen;
(2) NIZO food research, Ede; 3Wageningen Centre for Food Sciences, Wageningen; The Netherlands.

Lactobacillus plantarum is a versatile and flexible lactic acid bacterium (LAB) that is important in industrial food fermentation processes. After careful annotation of the genes encoded by the genome of L. plantarum WCFS1 (1), a reconstruction of the metabolic network has been carried out.

LacplantCyc is a pathway / genome database (PGDB) generated from the annotated genome sequence and the MetaCyc database (2), using the Pathologic software of Pathway Tools (3). The predicted metabolic network is being manually validated and adjusted with literature and experimentally verified data.

L. plantarum will serve as a model organism for pathway/genome annotations and comparisons with other LAB. We wish to accelerate the reconstruction of the metabolic potential of other LAB, e.g. Lactococcus lactis, by making use of Pathologic software and an in-house database of orthologous genes. Annotation data describing the function of genes, comments of curators and additional information are stored in an in-house developed, web-interfaced, mySQL database, with links to LacplantCyc (http://www.lacplantcyc.nl) and to a tool to visualize the gene context of all public bacterial genomes, the Microbial Genome Viewer (http://www.cmbi.kun.nl/MGV)(4). Visualization of the data sets in different levels of detail is extremely important to help interpreting these data from a biological viewpoint.

Acknowledgements:
Douwe Molenaar, Jos Boekhorst, Robert Kerkhoven, Richard Notebaart are acknowledged for their work on the in-house databases and visualization tools.
Peter Karp and the Bioinformatics Research Group, SRI International, USA are acknowledged for Pathologic software and support.
Supported by the Netherlands Organisation of Scientific Research (NWO) BioMolecular Informatics Programme, grant 050.50.206.

References:
1. PNAS USA 2003; 100:1990-1995
2. Nucleic Acids Research 2004; 32:D438-42
3. Bioinformatics 2002; 18:S225-32
4. DOI: 10.1093/bioinformatics/bth159

metaSHARK: a database of automated metabolic reconstructions derived from genomic DNA sequence

John Pinney

School of Biochemistry and Molecular Biology, University of Leeds

The metabolic SearcH And Reconstruction Kit (metaSHARK) is an online resource for the visualisation, navigation and analysis of fully automated metabolic reconstructions for a variety of organisms.

For each enzyme in our database, the metaSHARK reconstruction software uses a combination of profile-based methods to search genomic DNA for regions with significant similarity to a set of model sequences. This approach offers significant advantages over other fully automated tools for metabolic reconstruction that require a set of predicted protein sequences, particularly in cases where accurate gene models are unavailable. Results for the genome of the apicomplexan parasite Eimeria tenella show that metaSHARK is able to identify approximately twice as many enzymatic functions as a recent protein sequence-based method.

Visualisation of the results is achieved with a Java applet which represents the metabolic network as a Petri Net structure. The user is free to navigate the whole network, overlaying the metabolic reconstruction results for a single genome or comparing reconstructions for two different genomes.

metaSHARK is available online at http://bioinformatics.leeds.ac.uk/shark/. We welcome requests for the analysis of additional genomes.

ROSPath: Pathway visualization at multiple levels

Kiyoung Choi1, Eunok Paek2, Kong-Joo Lee3

(1) kchoi@azalea.snu.ac.kr, School of Electrical Engineering and Computer Science, Seoul National University
(2) paek@uos.ac.kr, Dept. of Mechanical and Information Engineering, University of Seoul
(3) kjl@ewha.ac.kr, Div. of Molecualr Life Sciences and College of Pharmacy, Ewha Womans University

We have developed pathway visualization software that allows automatic visual inspection of stored signaling pathway information. It dynamically generates a graphic view of signaling pathway data stored in ROSPath database by communicating with ROSPath web and database servers. Different types of signaling entities and signaling interactions have distinct graphical representation for quick and easy visual discrimination. It can also process the graph-theoretic queries and present the search results as highlighted subgraph within the context of the target pathway.

Our work is unique in that it renders two mutually related graphical representations of a signaling pathway, one as a "signaling graph," and the other, as an "interaction graph."

Another feature of our application is its ability to draw hyper graphs. Signaling graphs are general directed graphs, while interaction graphs are hyper graphs. Most graph generation tools avoid the problem of drawing hyper graphs by indicating signaling interactions as nodes, rather than as edges, and by connecting related entities with so called “interaction nodes.”

Tool URL: http://rospath.ewha.ac.kr/viewer.jsp

Understanding omics results, improved maps for functional pathway mapping

Rachel I.M. van Haaften1, Marjan J. van Erk2, Rob H. Stierum2 and Chris T.A. Evelo1

(1) BiGCaT Bioinformatics, Technical University Eindhoven and Maastricht University, The Netherlands
(2) TNO Nutrition and Food Research, Zeist, The Netherlands. All authors are members of the Work Package 7 (bioinformatics) of the European Nutrigenomics Organisation NuGO.

Understanding transcriptomics or proteomics results is far from trivial. Genome wide gene expression results may sound fancy; dealing with expression changes for over 20.000 genes may very well turn into an awkward task. Typical approaches to study such data sets include 1) just look at the genes that show the largest fold changes (can hardly be called omics), 2) only check the researchers favorite genes (certainly not omics at all, and you could often have used cheaper approaches) and 3) look for patterns of genes with the same kind of behavior over different experimental conditions (PCA, clustering etc). Although the latter does or at least can use all the data in a dataset it also just transfers the problem. First you don't understand the behaviour of the large number of genes, and now you don't understand why so many genes are in a certain pattern.

One way to solve this problem is to try to visualize the expression results within understandable biological pathways. This can for instance be attained with the help of the free computer application GenMAPP (Gene MicroArray Pathway profiler. http://www.genmapp.org). Maps and tables for expression visualization can be specifically designed or be derived from databases like GO and KEGG. A useful map typically consists of a visible level showing the actual relationships between genes, proteins and metabolites and a back page containing information about these biological entities plus their IDs in well-annotated databases like Swiss-Prot. The latter are needed for data coupling. In practice maps are often not satisfactory for the expert scientists working with them: genes are missing, whole pathways were not mapped yet, or problems exist for back page information. NuGO together with the Dutch nutrigenomics research program on gut health launched an initiative to improve pathway maps. Fatty acid metabolism maps will be used as a first set of maps to gain experience.

As a first step existing maps and data available in e.g. GO and KEGG will be combined and forwarded to a limited group of experts to verify the main layout (do we need whole new maps, do we need to split maps, are there any gross errors or obsolete maps?) and for identification of the most important focuses in the maps. After processing of the expert responses we will use the suggested focuses to start text mining procedures (for this we plan to cooperate with the Dutch text mining initiative from BioRange). The maps (with new main layouts where applicable) and the summarized text mining results will again be forwarded to a limited group of experts this time asking for more detailed responses (what sub-paths are missing, are all genes present?). We will then start collecting the back page information and develop the actual maps to be used. For this a new program and operating system independent map format will be developed in XML with converters to mainstream mapping tool formats including but not limited to GenMAPP. In parallel with this the new drafts (without the back page updates) will be forwarded to a larger group of experts within NuGO. The resulting maps will be tested using actual transcriptomics and proteomics (and possibly metabolomics) results and made available to the scientific community.