# Information and Simulation Systems for the Analysis of Gene Regulation and Metabolic Pathways

## Preface

The third Dagstuhl Seminar for Information and Simulation Systems for the Analysis of Gene Regulation and Metabolic Pathways was held from June, 24 to 29, 2001. It was a multidiciplinary seminar with participants from 11 different countries. Schloss Dagstuhl workshops in general emphasize computer science, and we are delighted to focus on the rapidly developing links between biosciences and computer sciences. The 2001 meeting is a sequel to the 1995 and 1998 meeting on the similar topic.

Molecular biology and biotechnology have begun to focus sharply on the problem of gene regulation. This problem is inescapable, because no open reading frame (ORF) will be expressed without the appropriate regulatory sequences. Moreover, some genes code for proteins whose function is to turn other genes on and off. Groups of these genes form networks with complex behaviors. These networks control other genes whose protein products catalyze specific biochemical reactions, and the small molecules which are substrates or products of these reactions can in turn activate or deactivate proteins which control transcription or translation. For that reason, gene regulation can be said to indirectly control biochemical reactions in cellular metabolism, and cellular metabolism itself exerts control on gene expression. For these reasons, the interdependent biochemical processes of metabolism and gene expression can and should be interpreted and analyzed in terms of complex dynamical networks. Hence modeling and simulation are necessary. Two earlier Dagstuhl seminars ( 1995 and 1998) have already dealt with modeling and simulation of biochemical networks. Both sought to bridge two divides by both bringing together scientists in the disciplines of gene regulation and metabolic pathways, and within and across both of these areas bringing together experimentalists and theoreticians. Often there had been little previous contact among these groups, but clearly the integration of metabolic and gene expression models as well as the cooperation of theorists and experimentalists is essential in order to solve these complex problems.

Apart from theoreticians and experimentalists, a third group has emerged since 1995 which is centered around databases and the internet. Many molecular biologists turned towards informatics and systematically collected results relating to specific problems. These data have been and will be stored systematically in specific databases, which nowadays are accessible via the Internet. Recently many firms have been founded which provide data essential for the solution of scientific and industrial problems, and even more importantly the corresponding infrastructure. As a result, there are databases available via the Internet for all known sequenced genes (e.g. EMBL), proteins (e.g. SWISS-PROT, PIR, BRENDA), transcription factors (TRANSFAC), biochemical reactions (KEGG) and signal induction reactions (TRANSPATH, GeneNet). Beyond databases, simulators for metabolic networks which employ most of the currently popular modeling methods are also available via the Internet. In addition to the classical methods of differential equations, discrete methods have become quite important. Examples are the object-oriented approach, rule-based systems, Petri Nets, graphs, and Boolean nets.

These recently implemented tools on the Internet are the basic components of the informatic and analytical infrastructure of biotechnology. Clearly the next evolutionary stage of development will be the implementation of integrated molecular information systems (e.g. SRS). The first step to reach that goal is the integration of databases under a specific biological perspective. The next step will be user-defined molecular information fusion. Up to now, there are no standard tools available in order to successfully separate both methods and databases. Exactly for that reason it is imperative to develop uniform intersections at this stage. To discuss properties of these intersections was one major issue of this seminar.

Ralf Hofestädt
John Reinitz
Nikolay Kolchanov

# Contents

# Final Program

Dagstuhl Seminar "Functional Genomics"
Information and Simulation Systems for the Analysis
of Gene Regulation and Metabolic Pathways

June, 24-29, 2001

**Monday, 25th June**

| | | |
|---|---|---|
| 09:00 | Opening | |
| 9:05 | *Whole Genome & Proteome Analysis*<br>Rolf Apweiler (EBI Hinxton) | p. 11 |
| 10:00 | *Computer-aided design of metabolic networks*<br>Matthias Reuss (University Stuttgart) | |
| 11:00 | Break | |
| 11:15 | *Can bioinformatics bridge the gap between gene expression studies and functional genomics: Case studies and search for algorithmns in aid of an electronic footprinting procedure.*<br>Jürgen Borlak (FHG Hannover) Gene Regulation | |
| Gene Regulation | | |
| 14:00 | *Congruence and Clustering of Transcriptome Expression in E.coli*<br>Julio Collado-Vides (National University of Mexico) | p. 13 |
| 14:45 | *Simulating Linkages between Gene Expression States and Field Phenotypes in Breeding for Drylande Environments*<br>Scott Chapman (CSIRO Plant Industriy) | p. 14 |
| 15:30 | *Sequence-oriented modelling of gene expression*<br>Sabine Arnold and Matthias Reuss (University Stuttgart) | p. 15 |
| 16:15 | Break | |
| 16:30 | *Regulatory Networks in Sea Urchin Embryo*<br>Eric Davidson (CalTech Pasadena) | |
| 17:15 | Break | |
| 20:00 | ***Computer Demos - Open Session***<br>Molecular Databases and Information Systems | |
| | *BClass, Bayesian Self-Organizing Maps for* | p. 16 |

*Classification of Heterogeneous Biological Variables:*
*An Application to Transcriptome and Phylogenetic*
*Profile data*
Luis Arturo Medrano Soto (National University of Mexico)

*Transpath, a database for signal transduction* p. 17
Frank Schacherer (Biobase)

**Tuesday, 26th June**

Molecular Databases and Metabolic Pathways

09:00 *BRENDA Enzyme Information System*
Dietmar Schomburg (Universität Köln)

09:45 *TRANSPATH, a database for signal transduction* p. 17
Frank Schacherer (BIOBASE)

10:30 Break

10:45 *Metabolic pathway analysis* p. 18
Stefan Schuster
(Max Delbruck Centre for Molecular Medicine Berlin)

11:30 *BioPath - Biochemical Pathways Database and* p. 20
*Visualization System*
Falk Schreiber (University Passau)

12:15 Break

Metabolic Pathways

14:00 *From Re-Annotation to Re-Thinking: "Analyzing* p. 21
*Metabolic Pathways in Mycoplasma Pneumoniae*
Thomas Dandekar (EMBL Heidelberg)

14:45 *A Database System to support the Modelling and* p. 22
*Analysis of Biochemical Pathways*
Isabel Rojas (Europ. Media Lab Heidelberg)

15:30 Break

15:45 *Individually Integrated User Databases to support* p. 23
*Modeling and Animation of Regulative Gene Networks*
Andreas Freier (University Magdeburg)

16:30 *Semantic Modelling of Signal Transduction Pathways* p. 24
*and the Connection to Biological Datasources*

Marco Weismüller (DKFZ Heidelberg)

17:15 Break

20:00 ***Computer Demos - Open Session***
Information Systems

*An Informatics Infrastructure for the Analysis of Gene Regulation and Pathways* p. 25
Chris Stoeckert (University of Pennsylvania)

*BioPath* p. 20
Falk Schreiber (Universität Passau)

*MARGBench - Heterogeneous Database Integration for Modeling and Animation of Gene Regulative Networks* p. 26
Andreas Stephanik (Universität Magdeburg)

## Wednesday, 27th June

Gene Regulation Networks

9:00 *Modelling Properties of Gene Networks* p. 27
Mark Cooper (Pioneer Hi-Bred International Inc. - Johnston)

9:45 *GeneNet system: description and modelling of gene networks*
Nikolay Kolchanov (Russian Academy of Sciences)

10:30 Break

11:00 *The Solution to the Problem on Parametric Stability for Ontogenesis Control Gene Networks* p. 28
Rustem Tchuraev (Russian Academy of Sciences, Ufa)

11:30 *In Silico Evaluation* p. 31
Patrizio Arrigo (CNR Genova)

12:00 Break

14:00 Social Event
Excursion to Trier

20:00 ***Computer Demos - Open Session***
Simulation Tools

*Workbench for Modelling a Dynamics of Gene Expression in Drosophila Clastoderm*
Maria Samsonova (Inst. for High Performance Comput. St. Petersburg)

*Computer-aided design of metabolic networks*
Klaus Mauch and Matthias Reuss (University Stuttgart)

**Thursday, 28th June**

Modelling of Metabolic Pathways

9:00 *Functional Genomics*
John Reinitz (The University at Stony Brook NY)

09:45 *How to construct the integrated atlas of gene expression in situ*
M. Samsonova (Institute for High Performance Computing, St.Petersburg, Russia)

10:15 Break

10:30 *Graph-based analysis of Biochemical Networks* p. 32
David Gilbert (City University London)

11:15 *A Workbench for Modeling and Analysis of Cellular Systems*
Martin Ginkel (MPI Magdeburg)

11:45 *Systems Approaches to Plant-Pathogen Interactions*
Mary C. Wildermuth (Massachusetts General Hospital Boston)

12:30 Break

Metabolic Control

14:00 *Qualitative-Quantitative Simulation of Biological Regulatory Processes* p. 33
Steffen Schulze-Kremer (Resourcenzentrum Berlin)

14:45 *Simulating large-scale biochemical systems: forward and inverse problems*
Pedro Mendes (Virginia Bioinformatics Institute)

15:30 Break

16:00 *Cellular Oscillators in Animal Segmentation* p. 35
Johannes Jaeger (SUNY at Stony Brook)

16:30 *Fast Redundant Dyadic Wavelet Transform in Application to Spatial Registration of the Expression Patterns of Drosophila Segmentation Genes*
Konstantin Kozlov (Inst. for High Performance Comput. St. Petersburg)

17:00 Break

| | | |
|---|---|---|
| 20:00 | Global Discussion | |

**Friday, 29th June**

Virtual Cell

| | | |
|---|---|---|
| 9:00 | *Creating and Analysing in Cilico Bacterial Cells*<br>Serge Dronov (Glaxo Welcome Research Medicine Center, Stevenage) | p. 36 |
| 10:00 | *Limitations of Modelling in Biology*<br>Minija Tamosiunaite (Vytautas Magnus University Kaunas) | p. 37 |
| 10:45 | Final Discussion | |

# Whole proteome analysis

Rolf Apweiler
EMBL Outstation - Hinxton,
European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton,
Cambridge, CB10 1SD,
United Kingdom

It is no longer ludicrous to envisage collecting vast amounts of genomic data, although it remains a massive task. The challenge is in developing the tools and methods required to analyze the data. The SWISS-PROT group at the EBI combines manual annotation and sequence analysis of SWISS-PROT entries with rule-based automatic annotation of TrEMBL entries to provide a comprehensive, reliable and up-to-date protein sequence database. The EBI proteome analysis initiative aims to provide comprehensive, easily accessible information as quickly as possible to the user community. Proteome analysis data has been produced for all the completely sequenced organisms spanning archaea, bacteria and eukaryotes. Complete proteome sets for each organism have been assembled from the SPTR (SWISS-PROT + TrEMBL + TrEMBLnew) database to be wholly non-redundant at the sequence level. This proteome data has been used in the analysis and is easily accessible and downloadable from the proteome analysis pages. The resources with the highest information content are InterPro and CluSTr. InterPro classifies 50-70% of all proteins in a proteome into distinct families. In addition, InterPro provides insights into the domain composition of the classified proteins. The proteome analysis pages make available InterPro-based statistical analysis that includes, among other information:

- General statistics - lists all InterPro entries with matches to the reference proteome. The matches per genome and the number of proteins matched for each InterPro entry are displayed.

- Top 30 entries - lists the top 30 InterPro entries with the highest number of protein matches for the reference proteome.

- 15 most common domains - lists the InterPro entries with the largest number of Pfam and profile matches (defined as domains) for the reference proteome. The matches per genome and the number of proteins matched for each InterPro entry are shown.

WWW links:
CluSTr: http://www.ebi.ac.uk/clustr/
InterPro: http://www.ebi.ac.uk/interpro/
Proteome analysis database: http://www.ebi.ac.uk/proteome/
SP_TR_NRDB: ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/
SWISS-PROT & TrEMBL: http://www.ebi.ac.uk/swissprot/

References:

Apweiler R.;
"Functional information in SWISS-PROT: The basis for large-scale characterisation of protein

sequences."
Briefings in Bioinformatics 2:9-18(2001)

Kriventseva E.V., Fleischmann W., Zdobnov E.M., Apweiler R.;
"CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins."
Nucleic Acids Res. 29(1):33-36(2001).

Apweiler R., Attwood T.K., Bairoch A., Bateman A., Birney E., Biswas M., Bucher P., Cerutti
L., Corpet F., Croning M.D., Durbin R., Falquet L., Fleischmann W., Gouzy J., Hermjakob H.,
Hulo N., Jonassen I., Kahn D., Kanapin A., Karavidopoulou Y., Lopez R., Marx B., Mulder N.J., Oinn T.M., Pagni M., Servant F., Sigrist C.J., Zdobnov E.M.;
"The InterPro database, an integrated documentation resource for protein families, domains and functional sites."
Nucleic Acids Res. 29(1):37-40(2001).

Apweiler R., Biswas M., Fleischmann W., Kanapin A., Karavidopoulou Y., Kersey P., Kriventseva
E.V., Mittard V., Mulder N., Phan I., Zdobnov E.;
"Proteome Analysis Database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes."
Nucleic Acids Res. 29(1):44-48(2001).

Apweiler, R.;
"Protein sequence databases."
In: Advances in Protein Chemistry, Richards F.M., Eisenberg D.S., Kim P.S.(eds.); pp. 54:31-71, Academic Press, New York (2000).

Bairoch A., Apweiler R.;
"The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000."
Nucl. Acids Res. 28:45-48(2000).

# The *E.coli* Paradigm in Microbial Computational Genomics of Gene Regulation

Julio Collado-Vides
CIFN-National Autonomous University of Mexico
A.P. 565-A Cuernavaca, Morelos 62100, Mexico.

*Escherichia coli* is a free-living bacteria that captures a rich legacy of knowledge of years of experimental work in molecular biology. It is a reference organism to many of the tasks in computational biology of microbial finished genomes. We have been gathering for years from the literature, experimental data on transcriptional regulation and operon organization in E.coli K-12 and organizing it in RegulonDB, a relational database available on the web (http://www.cifn.unam.mx/regulondb/). RegulonDB has information for more than 600 transcription initiation sites precisely mapped in the genome, for more than 500 DNA-binding transcriptional interactions involved in transcriptional activation and repression, as well as for close to 300 known operons and 160 transcriptional DNA-binding regulatory proteins.
We have used this knowledge to generate computational methods to predict the missing set in the genome in *E.coli*. These predictions provide a tentative complement to the experimentally supported regulatory elements in the genome. We also take with interest the implementation of these methods, thinking of them as models to understand the interactions of the RNA polymerase with the promoter, as well as of operon organization in bacteria.
We have expanded these analyses to other microbial genomes. We have suggested a common origin in the set of transcriptional regulators in all the microbial organisms, based on the conservation of a motif that extends around the helix-turn-helix DNA binding motif. Four repressor families are suggested as those present before the divergence of eubacteria and archaea. Based on an analysis of pairs of genes located inside operons and the contrasting set of pairs at the boundaries of two operons, we have implemented a method to predict operons in many different microbial genomes. Information on this work can be found through our website at
http://www.cifn.unam.mx/Computational_Genomics/computational_genomics.html

Furthermore, we have used this comprehensive view of operons and transcriptional regulation to analyze global expression profiles in E.coli. These comparisons permit to distinguish sets with a congruent behavior in relation to previous knowledge on one hand, and to initiate an analysis of direct and indirect interactions of subsets of the regulatory network of the cell.

# Simulating plant responses to the environment – from crop to gene

S.C. Chapman[1], G.L. Hammer[2], M. Cooper[3], D. Podlich[4]

[1] CSIRO Plant Industry, 120 Meiers Road, Indooroopilly, Queensland 4068, Australia
[2] Agricultural Production Systems Research Unit, Queensland Department of Primary Industries, Tor St, Toowoomba, Queensland 4350, Australia
[3] Pioneer Hi-Bred International Inc., 7300 N.W. 62nd Avenue, P.O. Box 1004, Johnston, Iowa 50131, USA
[1] School of Land and Food Sciences, The University of Queensland, Brisbane, Queensland 4072, Australia

In private and public plant breeding efforts, information technology is beginning to integrate data at 3 levels: gene location and effect; genotype identify and breeding value; phenotype performance. We are developing tools to provide a simulated data platform in which to evaluate better methods of breeding plants with superior adaptation to their environments. The key attributes of this system, cf. genomic applications related to health and disease are the scales associated with organism structure, and temporal and spatial interactions between plants and the environment. Annual crop plants grow for between 30 and 240 days using sunlight, water and nutrients and interact with each other in communities to develop complex organisms. Plants respond to the environment using both coarse, long-term controls (e.g. leaf and root development) and fine, short-response controls (e.g. leaf orientation and stomatal opening). Due to the impossibility of measuring all environmental conditions over the time and spatial scales, it is necessary to construct integrating simulation models that describe the resource acquisition of plants. The sub-components of these models contain parameters that are observable as 'traits' to which we can assign genetic values, and through simulation determine their genetic adaptive values. This final step allows us to create a genotype-environment landscape to be searched using various strategies that represent the process of plant breeding.

Links:
http://www.pi.csiro.au/Research/M-Integrated%20Crop%20Physiology%20and%20Genetic%20Improvement/SubPrograms/MC.htm

http://www.apsru.gov.au/Products/apsim.htm

http://pig.ag.uq.edu.au/qu-gene/

# Sequence-oriented modelling of gene expression

Sabine Arnold, Martin Siemann, Matthias Reuss
Institute of Biochemical Engineering, University of Stuttgart
Allmandring~31, D-70569 Stuttgart, Germany
phone: +49-711-685-5161, fax: +49-711-685-5164,
arnold@ibvt.uni-stuttgart.de

The rates involved in the process of biosynthetic protein production rely critically on the type of protein to be generated and thus on the characteristics of the encoded gene sequence. In an attempt to predict gene expression rates for universal gene products, this study derives a functional context between the genomic sequence itself and some of the key reaction rates involved in the protein synthesis process. At the example of prokaryotic protein expression, it is outlined at which stages the gene sequence serves as a model entry in the formulation of kinetic rate expressions.

Transcription factor-independent mRNA synthesis employing the enzyme T7 RNA polymerase is modelled in terms of the base composition of the mRNA product, as well as parameters related to the recognition sequences of the transcription initiation and termination sites, respectively. The rate of ribosomal protein synthesis considers the impact of codon usage and the interaction among translating ribosomes. Further, translation factor action during translation initiation, elongation, and termination are reflected mathematically. The mRNA degradation model takes into account the common mechanism of both endonuclease and exonuclease activities. The impact of mRNA secondary structure on both translation efficiency and mRNA stability is shown to be crucial in determining gene expression levels.

An integrative approach combining these modelling units into one model is particularly useful in the optimum design of recombinant protein production and the modulation of enzyme activities within synthesis pathways. Additionally, such a model may serve to provide in a preliminary estimation the parameters needed for black-box modelling of gene expression, where the rates of protein synthesis and mRNA degradation are typically described by simple first-order kinetics.

# BClass, Bayesian Self-Organizing Maps for Classification of Heterogeneous Biological Variables: An Application to Transcriptome and Phylogenetic Profile Data

Arturo Medrano-Soto, Andres Christen-Gracia and Julio Collado-Vides

A number of computational methods have been applied to cluster transcriptome data alone, in order to find common patterns of gene expression (Eisen et al, 1998; Tamayo et al, 1999; Claverie, 1999; and references therein), however it would be desirable to add another related biological variables to the clustering analysis that could help to extract additional knowledge from transcriptome data.

We present here a Bayesian approach to cluster heterogeneous types of biological variables (BClass), and its application to classify all genes in Escherichia coli considering the following variables: 5 transcriptome experiments and protein phylogenetic profiles for 17 bacterial genomes. BClass is based mainly on mixture models, statistical distributions and avoids the definition of a „distance“ measure to cluster observations together. Instead of assigning a given observation exclusively to a single cluster, we report the posterior probabilities of each observation to belong to each one of the elements in the mixture model.

(1) Eisen MB, Spellman PT, Brown PO, Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. PNAS 95(25):14863-8

(2) Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. PNAS 96:2907-2912

(3) Claverie JM.(1999). Computational methods for the identification of differential and coordinated gene expression. Hum Mol Genet 8:1821-1832

# Transpath - A Database for Signal Transduction

Frank Schacherer
Biobase
Halchtersche Strasse 33
D-38304 Wolfenbüttel
Germany

TRANSPATH is an information system on signal-transduction networks. It focuses on pathways involved in the regulation of transcription factors. Molecules and reactions are the nodes in a signaling graph. They are stored in an object-oriented database, together with information about their location, quality, family relationships and signaling motifs. Also stored are links to other databases and references to the original literature. TRANSPATH differentiates between the states of a signal molecule, and can adequately describe the reaction mechanisms of signaling interactions.

TRANSPATH is free for academic use and available over the web (http://transpath.gbf.de). Pathway query mechanisms and several kinds of display are provided for the database in addition to text-based queries and information on single entries.

TRANSPATH professional is an enhanced version of TRANSPATH which contains more data and additional visualisation tools. It can be obtained from Biobase GmbH (http://www.biobase.de).

Commented Bioinformatiks links:
http://www.schacherer.de/frank/bookmark/Bioinformatics.html

# Metabolic pathway analysis

Stefan Schuster
Max Delbruck Centre for Molecular Medicine
D-13092 Berlin-Buch, Germany

Metabolic network analysis, alternatively called pathway analysis, has manifold applications in functional genomics, metabolic engineering and medicine.

We present the concept of elementary flux modes, which has turned out helpful in pathway analysis. Roughly speaking, an elementary mode is a minimal number of enzymes that can operate at steady state with all irreversible reactions used in the right direction. The enzymes are weighted by the relative flux they carry. Elementary modes can be interpreted as pathways.

The mathematical fundamentals of this concept are explained, in particular the properties of convex cones. It is compared with related concepts introduced by other authors. The routine for computing all elementary flux modes in a given system has been implemented in computer programs by several authors. These programs have been written in Pascal, C, Maple and JAVA. The applicability of the concept for the prediction of molar yields in metabolic engineering is outlined. The method is illustrated by several biochemical examples such as the metabolism of Mycoplasma pneumoniae. Finally, we comment on the evolution of different pathways in energy metabolism, in particular fermentation and respiration. As organisms degrading sugar by fermentation exhibit a high ATP production rate but a low yield, they use the resources in a wasteful manner. This is known in game theory as the „tragedy of the commons." We show under which conditions organisms using the more efficient respiratory pathway can win the competition against fermenters.

References:

S. Schuster, D. Fell, T. Dandekar: A General Definition of Metabolic
Pathways Useful for Systematic Organization and Analysis of Complex
Metabolic Networks. Nature Biotechnol. 18 (2000) 326-332.
T. Pfeiffer, S. Schuster, S. Bonhoeffer: Cooperation and Competition in
the Evolution of ATP Producing Pathways. Science 292 (2001) 504-507.

S. Schuster, T. Dandekar, D.A. Fell: Detection of Elementary Flux Modes in
Biochemical
Networks: A Promising Tool for Pathway Analysis and Metabolic Engineering,
Trends
Biotechnol. 17 (1999) 53-60.

T. Pfeiffer, I. Sanchez-Valdenebro, J.C. Nuqo, F. Montero, S. Schuster:
METATOOL: For Studying Metabolic Networks. Bioinformatics 15 (1999)
251-257.

C.H. Schilling, S. Schuster B.O. Palsson, R. Heinrich: Metabolic Pathway
Analysis: Basic Concepts and Scientific Applications in the Post-genomic
Era. Biotechn. Prog. 15 (1999) 296-303.

S. Schuster: Studies on the Stoichiometric Structure of Enzymatic Reaction Systems, Theory Biosci. 118 (1999) 125-139.

T. Dandekar, S. Schuster, B. Snel, M. Huynen, P. Bork: Pathway Alignment: Application to the Comparative Analysis of Glycolytic Enzymes, Biochem. J. 343 (1999) 115-124.

S. Schuster: Use and Limitations of Modular Metabolic Control Analysis in Medicine and Biotechnology. Metab. Engng. 1 (1999) 232-242.

S. Schuster: Use and Limitations of Modular Metabolic Control Analysis in Medicine and Biotechnology. Metab. Engng. 1 (1999) 232-242.

WWW-Links:
http://www2.bioinf.mdc-berlin.de/metabolic/

# BioPath - A Biochemical Database and Visualization System

Falk Schreiber,
Universität Passau

Metabolic processes in organisms can be modelled as large networks consisting of reactants, products and enzymes with multiple interconnections representing reactions and regulations. An example is given by the well known Boehringer poster. BioPath - a joint project of research groups at the universities of Erlangen, Mannheim, Passau and Spektrum Verlag - provides convenient electronic access to the content of the poster and to the growing information of biochemical reactions. BioPath supports browsing through biochemical pathways, the computation of reaction networks between given substances and the access to additional information about substances and reactions.

One key feature of BioPath is the automatic visualization of pathways. The state of the art in the visualization of biochemical reaction networks are manually produced drawings, as they appear in biochemical textbooks, on the poster or in electronic information systems like KEGG. As biochemical pathways are represented by networks the placement of objects and the routing of their connections is a typical graph drawing problem. However, existing algorithms are not sufficient to draw these networks according to the established conventions of biology and chemistry.

Therefore we developed a new algorithm for the visualization of pathways. This algorithm produces hierarchical layouts of directed graphs taking into account node sizes and layout constraints. A special treatment of node sizes leads to compact placements. Restricted by layout constraints the reaction components are placed in the established drawing style of biochemistry. Distinguished paths are drawn differently (e. g. the urea cycle as a real cycle) and the mental map of the user is preserved in sequences of related drawings. In graph drawing the term „preserving the mental map“ is used to express that small changes of the data shall imply only small changes of the picture.

For more information about BioPath see
http://biopath.fmi.uni-passau.de.

# From Re-Annotation to Re-Thinking: Analyzing Metabolic Pathways in Mycoplasma Pneumoniae

Thomas Dandekar
(EMBL, Heidelberg and Institute for Molecular Medicine, Freiburg)

The seminar discusses first the results of a major re-annotation effort now five years after the original sequence has been published (Dandekar et al., 2000). The effort combined theoretical analysis including more sensitive software with new proteomics and gene expression data. This allowed a clear gain in annotation. The total number of ORFs was increased from 677 to 688 (10 new proteins were predicted in intergenic regions, two further were newly identified by mass spectrometry and one protein ORF was dismissed) and the number of RNAs from 39 to 42 genes. A number of new biologically findings emerged, too including a small 200 nt RNA, several pathogenicity factors, new sugar utilization pathways and a protein secretion system similar to that from *E.coli*.

These results are now developed further in the seminar to general insights emerging from such a project such as:
ˉ Protein domain architecture remains a source of annotation inaccuracy and increases complexity of pathway predictions.
- The use of a controlled vocabulary, or, in larger projects, an onthology for genome annotation or re-annotation is highly recommended.
- Data integration, in particular from proteomics and gene expression data, but including also genetics and biochemical data is very helpful to improve genome annotation and will stay a major challenge during the next years.
- Genome annotation is a continuous task: what we see in a genome is always only a current interpretation.

This web-pointer gives detailed data on the Mycoplasma re-annotation project:
http://www.bork.embl-heidelberg.de/Annot/MP/

Labpointers:
http://www.embl-heidelberg.de/~dandekar/
http://www.ukl.uni-freiburg.de/immz/molmed/ag_dandekar/ag__dandekar.htm

reference:

Dandekar T, Huynen M, Regula JT, Ueberle B, Zimmermann CU, Andrade MA, Doerks T, Sanchez-Pulido L, Snel B, Suyama M, Yuan YP, Herrmann R, Bork P.
Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. Nucleic Acids Res. 2000 Sep 1;28(17):3278-88.

# A database system to support modelling and analysis of biochemical pathways

I.Rojas, L. Bernardi, E. Ratsch and R. Kania
Scientific Databases and Visualisation Group
European Media Laboratory, Heidelberg, Germany

To provide support for the analysis of biochemical pathways a database system based on a model that represents the characteristics of the domain is needed.  This domain has proven to be difficult to model with the use of conventional data modelling techniques. In part this is due to the difficulty of formalising even what are considered to be „basic“ concepts, such as „genes“. The data model must also be flexible enough to support the evolution of relationships and of properties characterising objects (and thus the classification of objects), common in molecular biology and biochemistry. Exceptions, generalisations, complex relations, and behavioural constraints have also to be taken into account.

We are building an ontology for biochemical pathways, which acts as the basis for the generation of a deductive database on the same domain, allowing the definition of complex queries and complex data representation. The generation of a relational database based on this ontology is also intended.  The ontology is used as a modelling and analysis tool, which allows the expression of complex semantics based on a first-order logic representation language. The induction capabilities of the system can help the scientist in extracting and testing research hypotheses that are difficult to express with the standard relational database mechanisms. An ontology representing the shared formalisation of the knowledge in a scientific domain can also be used as data integration tool clarifying the mapping of concepts to the developers of different databases. Other applications include the use of the ontology as the basis for information extraction programs and data curation of biological databases.

Web Links:

https://projects.villa-bosch.de/sdbv/

http://www.ims.uni-stuttgart.de/projekte/GenIE/

# IIUDB – Individually Integrated User Databases To Support Modeling and Animation of Regulative Gene Networks

Andreas Freier
University of Bielefeld
afreier@techfak.uni-bielefeld.de

Today, more and more information in the area of molecular biology and bioinformatics is stored in freely available internet databases. Each of these databases has been designed for the efficient storage and representation of data taking into account a specified view on the biological cell. Regarding regulative gene networks, the access to information of different views stored in specific databases and information systems is necessary. Common integrative databases and database systems (e.g. the Sequence Retrieval System, SRS) provide a unique interface to a set of existing databases. Hence, the heterogeneity of the access interfaces and database models has been overcome and a homogeneous access interface is available. But still, without accomplishing data integration the problem of the different biological database views still exists. Additionally, no remote access to the databases is provided. In our MARGBench project*, we are solving this task by integrating database access, database schemes and database content at the same time. Semi-automatic access adapters are responsible for the remote database access, while user-defined integrated schemes describe the user biological destination view. The system, called MARGBench, in detail the component BioDataServer is responsible for online database access, data integration and restructuring the users integrated scheme.

Our aim is to provide metabolic together with gene regulation networks in order to support tools in the area of the analysis and simulation. The second MARGBench component, the Individually Integrated Database Server (IIUDB) has the capability to automatically implement databases, where the integrated data can be stored in the users biological view. Storing integration data in between, enables the user to check the database content for consistency, enrich or remove data. With the object-oriented IIUDB system, we switched from the raw relational database view (BioDataServer) to a view of object networks. Because the system is specified to support analysis and simulation of gene networks, a unique interface for network search and extraction exists. Together with the IIUDB Network Tool, any dynamically generated IIUDB database, dynamically filled with integrated data, can be searched for networks. E.g. it is possible to extract metabolic pathways interactively by path navigation and furthermore to search automatically for alternative pathways. The resulting network, which is a subset of the integration database, can be provided in common formats like ordinary biochemical rules and the Graph Markup Language (GML). As a reference application, our simulation environment MetabSim/Vis is directly linked with the IIUDB. We will provide a demo version of the network and the simulation tool as soon as possible. Currently, several demos for the BioDataServer are available at
http://www-bm.cs.uni-magdeburg.de/iti_bm/marg.

* This work is sponsored by the Research Council (DFG)

# Semantic Modeling of Signal Transduction Pathways and the Linkage to Biological Data Sources

M. Weismueller and R. Eils
Div. "Intelligent Bioinformatics Systems",
German Cancer Research Center (dkfz), 69120 Heidelberg, Germany

Signal transduction (ST) is the mechanism a cell reacts on a stimulus coming from outside the cell. ST alters gene transcription in the nucleus, therefore changing protein synthesis and the behavior of the cell. ST can be described as an information flow from outside into the cell mediated by biochemical reactions of signal molecules.

ST pathways play a major role in the field of cancer research. Several pathways have been identified to be responsible for cancer development by over- / underexpression of genes or by functional modification of signal proteins caused by alterations of their sequence.

Quantitative data and measurements of signal molecules are not yet available for a comprehensive number of pathways. Several model systems have been studied in detail including concentration and activity measurements of signal molecules. But these attempts are not sufficient to allow a study of whole signal transduction networks of cells. Therefore the idea is to reduce the view on signal information flow in the cell to a state based one: A protein is active (mediating information) or inactive (not mediating information). Additional information is incooperated when available from biological experiments: Protein X binds to protein Y, protein X phosphorylates protein Y etc.

One way to describe this non-quantitative view on signal transduction of cells is the qualitative modeling of signal information flow through the cell. The information flow is an abstract view on biochemical reactions of signal molecules. These interactions of molecules are semantically modeled describing the biochemical interaction not in numerical equations like differential equations, but in abstracted biochemical reaction descriptions.

The aim is to use the information of a ST database - TRANSPATH (http://transpath.gbf.de) - to build up a comprehensive model of ST pathways in the computer. One should be able to answer biological questions about the interaction or alteration of pathways under certain conditions and formulate hypotheses, which might be tested in experiments. To model ST pathways the pi-calculus (http://www.lfcs.informatics.ed.ac.uk/reports/89/ECS-LFCS-89-85/) is used to represent parallel interactions of proteins. This notion was adapted from the BioPSI project (http://www.wisdom.weizmann.ac.il/~aviv/). The pi-calculus is a kind of programming language. Protein interactions of ST pathways can be programmed and simulated using the information of TRANSPATH. These simulations result in an output, which has to be interpreted under the posed conditions.

The first step of the work is to model an important ST pathway: the ERK-MAPK pathway. In this model a protein is interpreted as a computational unit having an input layer, a computational layer and an output layer. This pathway model implementation exemplifies how ST pathways can be built up in a computer.

# An informatics framework for the analysis of gene regulation and pathways.

Chris Stoeckert
stoeckrt@SNOWBALL.pcbi.upenn.edu

Understanding gene regulation as it pertains to tissue specificity or developmental regulation requires a foundation of information on gene sequence annotation and gene expression experiments. We have developed two closely linked databases  for this purpose: Genomics Unfied Schema (GUS) and RNA Abundance Database (RAD). GUS integrates sequences and their annotation through tables representing genes, RNAs, and proteins.  RAD follows the guidelines of the Microarray Gene Expression Database (MGED) group in representating array (and non-array) experiments. Both databases are generic systems out of which project-specific views can generated. These views include a human and mouse gene index, AllGenes, a central resource for the malarial parasite Plasmodium falciparum, PlasmoDB, and a site supporting the Functional Genomics of the Developing Endocrine Pancreas Consortium, EPConDB. These sites integrate sequence and expression data. Future work will utilize tools such as the Transcription Element Search Software (TESS) to mine this data.

Lab home page: http://www.cbil.upenn.edu
MGED: http://www.mged.org
RAD: http://www.cbil.upenn.edu/RAD2
AllGenes: http://www.allgenes.org
PlasmoDB: http://plasmodb.org
EPConDB: http://www.cbil.upenn.edu/EPConDB
TESS: http://www.cbil.upenn.edu/tess

# A system for the integrated retrieval of molecular biology data

Matthias Lange, Andreas Stephanik, Uwe Scholz
(Otto-von-Guericke-University of Magdeburg)
{mlange|stephani|uscholz}@iti.cs.uni-magdeburg.de

**Andreas Freier**
(Bielefeld University)
afreier@techfak.uni-bielefeld.de

Abstract

The internet is developing into the most powerful medium for information retrieval. This fact is consequently reflected in molecular biology. With the automated high throughput experimental technologies that have been developed in the last years, a large number of research projects are currently producing an exponentially increasing amount of data. Only the Human Genome Project by itself produces 3*109 base pairs and mapping data. In order to take advantage of the potential of valuable databases it has to be considered that there is a requirement for standardized, integrative and homogeneous access to data from a wide range of sources using powerful query languages e.g. SQL. Thus, the integration of databases and the offering of a declarative query language can help scientists to detect new information and coherence to find correlations across the spectrum from genomics via proteomics up to drug design.

The idea of a mediator based database integration approach, called BioDataServer (BDS), has resulted in an architecture, which is realized as a client-server system. Where the BDS is the server and any molecular biology application, including database import modules, could act as clients. The homogeneous database access is realized by specific adapters. The BDS are able to process read-only SQL queries. On this basis a JDBC with XML capabilities and an ODBC driver have been implemented. The attachable software ranges from simple analysis tools via various molecular information systems up to complete frameworks for complex problems like simulations. Further information, applications, demos and drivers are available using the URL http://integration.genophen.de.

Selected References

Freier, A. et al, MARGBench - An Approach for Integration, Modeling and Animation of Metabolic Networks, in Proceedings of the German Conference on Bioinformatics (GCB '99), Hannover, Germany, 1999, pp. 190-194.

Hofestaedt, R. et al., Information Processing for the Analysis of Metabolic Pathways and Inborn Errors, BioSystems, 1998, vol. 47, pp. 91-102.

Köhler, J. et al., Logical and Semantic Database Integration, in BIBE 2000: IEEE International Symposium on Bio-Informatic & Biomedical Engineering, Washington, D.C., U.S.A.: IEEE Computer Society, 2000. pp. 77-80.

**Links**

BDS online Demos (phpMetatool, XML Browser, BioDataServer Demo Applet):
http://integration.genophen.de/

# Quantifying gene-to-phenotype relationships: Integrating across levels of organization in the genotype-environment system

M. Cooper[1] and S.C. Chapman[2]

[1] Pioneer Hi-Bred International Inc., 7300 N.W. 62nd Avenue, P.O. Box 1004, Johnston, Iowa 50131, USA
[2] CSIRO Plant Industry, 120 Meiers Road, Indooroopilly, Queensland 4068, Australia

Integrating information across levels of organization within a genotype-environment system is a major challenge in computational biology. However, if we are to be able to understand and predict phenotypes given knowledge of the genome this is a fundamental requirement in modeling biological systems. Organisms are consequences of this integration and it is a major property of biological systems that underlies the responses that we observe when we investigate components of these systems. We discuss the *E(N:K)* model as a framework for investigation of gene-to-phenotype relationships and prediction of system properties at different levels of organization.

Links: http://pig.ag.uq.edu.au/qu-gene/

# The Solution of the Problems on Parametric Stability for Ontogenesis Control Gene Networks

Tchuraev R.N., Galimzyanov A.V.

It is evident that each living system must possess the property of stability, in the broad sense, that affords the process of self-reproduction. Specifically, *ontogenesis* as a set of intrasystem processes resulting in self-reproduction of living systems must be stable both to external influences and internal fluctuations. The ontogenesis stability is achieved to a great extent by Ontogenesis Control System, which may be conceived as *a control gene network*.

Ten years ago one of the authors put forward a method of generalized threshold models – GTM-formalism to analyse the dynamics of control and controlled molecular genetic systems [1]. This method taking into account the peculiarity of control processes at the molecular level makes it possible to get both qualitative and quantitative pattern of the dynamics of gene networks.

In terms of GTM-formalism, using the method of generalized threshold models a number of mathematical models for both metabolism control prokaryotic systems (tryptophan, arabinose) [2] and two actual eukaryotic control gene networks, namely, the morphogenesis control subnetwork for *Arabidopsis thaliana* flower (MCS-AT) and the early ontogenesis control subnetwork for *Drosophila* (Dr-CGN) [3] were constructed. Also, models of λ-phage development control system (λ-PDCS) have been constructed by means of a special version of this formalism [4].

Any conclusions relying on research of strict mathematical models of actual systems can be trusted only when they are proved through research of their *structural stability*, in particular, *parametric stability* [5]. When constructing pattern models of control gene networks for the purposes of prediction of their dynamics one has to come up against the fact that most of the kinetic parameters (for example, unit intensities of synthesis and degradation of molecular components of a system) are not calculated experimentally. Because of this, on the strength of indirect evidence some verisimilar sets of parameters are to be selected, these being set, together with the structure of a system, in numerical calculations of dynamic equations. The plausibility of predictions necessitates an investigation on parametric stability of the models, when "sensitivity" of functioning regimes is tested for random changes of the parameters in a sufficiently wide range of values.

***Research on "sensitivity" of the regimes of the λ-phage development control system (λ-PDCS) for random changes of the parameters with the aid of machine experiments using the threshold model.*** Based on qualitative analysis of the model it was conjectured that ontogenesis regimes are not very sensitive to random changes in the parameters of the system. The essence of original approach to substantiate this statement is as follows. In computer models the values for the parameters (the frequencies of RNA polimerase and ribosome binding to the initiation sites on templates, thresholds) are chosen randomly from intervals of given length. The chosen set of parametric values is used to test the ability of the system to function in each of the two regimes within a fixed time interval of 40 minutes. The system was repeatedly tested by this procedure. The computer estimated the number of cases when there were realized either (a) lytic regime, (b) lysogenic regime or (c) both regimes. The ratio of these numbers to the total numbers of random choices characterizes the "sensitivity" of the regimes.

Two types of experiments were computer-simulated. In the first experiment the values of parameters were chosen at random over the intervals of ($x_i$ – 0.2) to ($x_i$, $x_i$ + 0.2 $x_i$), where $x_i$ – a parameter belonging to a "good" set, in other words, to such a set where both ontogenesis regimes could be realized. The ability of the system to function in the lytic (lysogenic) regime was estimated in the following way: it was necessary to attain concentration thresholds of O, P, Q and tof proteins ($C_2$, $C_3$ и $C_I$ proteins, except for Q protein). A program was developed to be calculated for two systems of intervals involving two “good” initial sets of parameters. The computer made random choice of the parameters 500 times for each of the two systems of intervals. The results are given in  the table 1 [4].

Table 1 - Results of the first computer experiment

| Set of parameters, № | Percentage of the cases when the system is capable of functioning in | | |
|---|---|---|---|
| | Lytic regime | lysogenic regime | both regimes |
| 1 | 78 | 12 | 12 |
| 2 | 94 | 61 | 61 |

The second computer experiment differed from the first one in that the computer chose sets of parameters at random in much greater intervals, some of their limits being set from physical reasons. In these intervals the computer made 1000 random choices of the sets of parameters and assessment for system functioning. It was of importance that as parameters changed in such a wide range the system turned out to be capable of functioning in 37 percent cases.

Thus, the results of both computer simulations offer evidence in favour of the suggested *stability, in the broad sense* of the word, *of λ-phage development control system*. This system was found to be insensitive to parameters fluctuations.

***Results of computer experiments in testing parametric stability for MCS-AT-2 model.*** Some verisimilar intervals of values for kinetic parameters were found on the basis of experimental data. For example, it is known that eukaryotes growth factor mRNAs do not exceed 30 minutes.

**Three types of computer experiments based on AGENDY software [6] were carried out. In the first experiment the values of parameters were chosen at random over the intervals of ($x_i$ – 0.3) to ($x_i$, $x_i$ + 0.3 $x_i$), where $x_i$ – a parameter from the set with the value less than the down-boundary of an appropriate verisimilar interval. In the second experiment the values of parameters were chosen at random within the 30 percent range around "good" set of parameters. Here, a reference “good” set of parameters was a set, in which values of all the variables were in the middle of appropriate permissible intervals. In the both experiments protein threshold concentrations were chosen much less than the expected protein stationary concentrations. In the third experiment the values for parameters were chosen in the same manner as in the second experiment, and protein threshold concentrations were much greater in their values so as to approximate protein stationary concentrations**

**In the first, second and third experiments the computer made 1000, 300 and 1000 random choices respectively. The ability of the system to change to the stationary regime and stay in it was evaluated at the end of a chosen time period of 96 hours long by a total number of switched-on genes in the normal stationary state, and also by analysing the distribution of gene switch-on in time. The results of computations are given in the table 2.**

Table 2 – Results of the computer experiments to test parametric stability of changes-over to stationary regimes for MCS-AT-2 model

| *Experiment, №* | Normal regimes (in percent) in whorls | | | | Normal flowers (in percent) |
|---|---|---|---|---|---|
| | sepals | petals | stamens | carpels | |
| 1 | 99.5 | 96.7 | 84.6 | 93.2 | 74 |

| | | | | | |
|---|---|---|---|---|---|
| 2 | 100 | 100 | 100 | 100 | 100 |
| 3 | 63.3 | 25.3 | 47 | 33.8 | - |

*Interpretation.* (1) Under conditions of the first and second experiments the system has high parametric stability. At the qualitative level, with one and the same behaviour, MCS-AT-2 can have strongly different quantitative characteristics (two to ten-fold differences of the concentration level of molecular components). (2) The results of the second experiment show that at the same initial data MCS-AT-2 is stable in the course of transitions to stationary states (stability of the system in stationary states is quite evident here). (3) Under conditions of the third experiment the system has not parametric stability. The results show that it is impossible for the given system to have the proposed ratio between coefficients of gene products synthesis and threshold values for proteins. (4) Comparison of the results of these three experiments show that as threshold values approach protein stationary concentration values the system stability decreases.

Thus, *the method of generalized threshold models makes it possible to study parametric stability of both prokaryotic and eukaryotic control gene networks*. The adequacy of the models constructed with the help of this method is supported by the results of computer experiments showing a high level of their parametric stability.

1. **Tchuraev R. N. (1991). A New Method for the Analysis of the Dynamics of the Molecular Genetic Control Systems. I. Description of the Method of Generalized Threshold Models. J. theor. Biol.,** 151**, 71-87.**

2. **Prokudina E.I., Valeev R.Y., Tchuraev R.N. (1991). “A New Method for the Analysis of the Dynamics of the Molecular Genetic Control Systems. II. Application of the Method of Generalized Threshold Models in the Investigation of Concrete Genetic Systems”. J. theor. Biol.,** 151**, 89-110.**

3. **Tchuraev R.N., Galimzyanov A.V. (2001) Modeling of Actual Eukaryotic Control Gene Subnetworks with the Method of Generalized Threshold Models as the Base. J. Mol. Biol. In press.**

4. **Ratner V.A., Tchuraev R.N. (1978). Simplest Genetic Systems Controlling Ontogenesis: Organization Principle and Models of Their Function. In: “Progress in Theoretical Bioilogy”, Acad. Press N.Y. et al.,** 5**, 81 – 127.**

5. **Arnold V. (1997). Personal communication.**

6. **Galimzyanov A.V. (2000). Software Automated Package for Analyzing the Dynamics of Control Gene Networks. In: "Proceedings of Second International Conference on Bioinformatics of Genome Regulation and Structure",** 1**, 233-234.**

**Links:**

- National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/
- Arabidopsis Information Resource (TAIR) http://www.arabidopsis.org/
- The TIGR Arabidopsis thaliana Database http://www.tigr.org
- Munich Information Center for Protein Sequences http://mips.gsf.de

# In silico determination of potential antisense target sites for beta haemoglobin variants

P.Arrigo
CNR Istituto Circuiti Elettronici,Via De Marini 6   16149 Genova
e-mail:arrigo@ice.ge.cnr.it

The completion of the rough primary genomic sequence of several organism has greatly impacted the approach to the drug discovery. The knowledge discovery in the huge amount of information available in the databases require a subsequent determination of the biological functionality of a specific gene. HTS method enhance the capability to screen the gene function. In order to design efficient drugs we need to investigate the involvement of a specific molecular target in a pathological process; only after the target validation it is possible to start the drug design.

There are many different approaches for gene functionalization or target validation, one of the more promising is based on the antisense technology.The antisense philosophy is rather simple, any tract of a mRNA can be potentially used for the hybridization; unfortunetly many constrains limits the capability to recognise the high efficient antisense domains.

At the present there is not a standardised procedure for the optimal target selection, a bioinformatic approach can help their identification. We have developed a procedure that can support the antisense design. The method combine an unsupervised SN (Sensory Network) with mechanical statistic approach and biophysical parameter. We have applied this methodology to the oligonucleotide design for site directed mutagenesis for unstable beta haemoglobin variants. These abnormal haemoglobin are characterised by a single nucleotide substitution and by a great variability in the phenotype expression. This approach is able to detect subtle modification induced by a single base variation in the potential profile of the sequence.

## Graph-based analysis of Biochemical Networks

David Gilbert
drg@soi.city ac.uk

Biochemical networks, e.g. metabolic pathways and gene regulation networks, or signalling pathways, can be represented as graph data structures, in fact directed graphs which may contain cycles. These graphs are characterised by partial information (indirect interactions, missing steps), negative results and partial interactions. They permit network navigation - how many pathways/steps from A to B, all pathways containing/lacking compounds/processes (constraints on pathways); highlighting significant parts of pathways; effects of turning on/off genes. even ignoring stochiometric and flux information we can perform network analyses - comparison of networks (inter, intra-organism), representation at different levels of resolution, discovery of recurrent 'motifs' (topological patterns, templates), finding positive or negative cycles.

We have been working in collaboration with the PFBP group at EBI led by Professor Wodak, who have created a database of biochemical interactions and associated curation tools. We have developed prototype computational systems to perform network navigation using constrained queries, and some graph layout facilities for metabolic pathways. We have also coded subgraph extraction algorithms which permit the identification of subgraphs associated with "seed nodes" - genes identified as being co-associated by gene expression experiments (array data). The current system comprises a Prolog image of the PFBP database and algorithms written in Prolog. This work was performed in collaboration with Jacques van Helden (ULB, Belgium) and Lorenz Wernisch (Birkbeck College, UK). The system can be accessed via www.ebi.ac.uk/research/pfmp

My bioinf resource list
http://www.soi.city.ac.uk/~drg/bioinformatics/resources.html

reading list
http://www.soi.city.ac.uk/~drg/bioinformatics/reading.html

the bioinf group at City
http://www.soi.city.ac.uk/~drg/bioinformatics/

# GermanyAbstract Part I: RZPD

Dr. Steffen Schulze-Kremer
CIO

RZPD Deutsches Ressourcenzentrum
für Genomforschung GmbH
Heubnerweg 6
D-14059 Berlin

The Resource Center (RZPD) has been established in 1995 as the central infrastructure of the German Human Genome Project to provide standardized biological reference materials and biological information for German and international research teams in the field of genome research. It is funded by the German Federal Ministry for Education and Research (BMBF). With its components at the Max-Planck-Institute for Molecular Genetics, Berlin, and the Deutsches Krebsforschungszentrum, Heidelberg, the Resource Center was turned into a nonprofit private limited company, RZPD GmbH in July 2000 (www.rzpd.de). The new organisational form creates the prerequisites to sustain and expand the comprehensive clone and data collection of the Resource Center for the advancement of genome research at favourable conditions that has attracted to date more than 7.000 scientists worldwide.

The main objectives of the RZPD are (a) providing information on, (b) generating and distributing of high quality biological materials to the scientific projects of the German Human Genome Project and other academic and industrial laboratories world wide. All information generated with our source material are collected in the Primary Database, which serves as an open platform for the exchange of information and results. The parallel generation of material and collection of data lead to valuable synergy effects that are of great advantage to many research projects. RZPD also provides molecular biological services like library spotting, rearraying, library screening.

RZPD staff conducts university lectures (e.g. Bioinformatik, WS 2000, Freie Universität Berlin) and specialised laboratory courses (e.g. EMBO course; expression data analysis course).

## Abstract Part II: Qualitative Simulation

Traditionally, biochemical systems are modelled using kinetics and differential equations in a quantitative simulator. However, for many biological processes detailed quantitative information is not available, only qualitative or fuzzy statements about the nature of interactions. In a previous paper we have shown the applicability of qualitative reasoning methods for molecular biological regulatory processes. Now, we present a newly developed simulation environment, BioSim, that is written in Prolog using constraint logic programming techniques. The simulator combines the basic ideas of two main approaches to qualitative reasoning and integrates the contents of a molecular biology knowledge base, EcoCyc. We show that qualitative reasoning can be combined with automatic transformation of contents of genomic databases into simulation models to give an interactive modelling system that reasons about the relations and interactions of biological entities. This is demonstrated on the glycolytic pathway.

References:

K. R. Heidtke, S. Schulze-Kremer (1998). Design and Implementation of a Qualitative Simulation Model of lambda Phage Infection. Bioinformatics, 14, pp. 81-91.

K. R. Heidtke, S. Schulze-Kremer (1998). BioSim - A New Qualitative Simulation Environment for Molecular Biology. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, pp. 85-94.

Heidtke, K. R. and Schulze-Kremer, S. (1999). Improving semi-quantitative reasoning by landmark approximation. In Proc. 13th Int. Workshop on Qualitative Reasoning About Physical Systems, Scotland.

Heidtke, K. R. and Schulze-Kremer, S. (1999). Deriving simulation models from a molecular biology knowledge base. In Proc. 4th Workshop on Engineering Problems for Qualitative Reasoning of the 16th Int. Joint Conf. on Artificial Intelligence.

URLs:

www.rzpd.de - RZPD Deutsches Ressourcenzentrum fuer Genomforschung GmbH

igd.molgen.mpg.de - Steffen Schulze-Kremer's Bioinformatics Pages at the Max-Planck-Institute for Molecular Genetics, Berlin

# Cellular Oscillators in Animal Segmentation

Johannes Jaeger
Graduate Program in Genetics
Dpt. of Molecular Genetics and Microbiology
SUNY Stony Brook
Stony Brook, NY 11794-5222
USA
tel (lab): ++1-631-632-9031
fax (dpt): ++1-631-632-9797
email: yoginho@usa.net

In my talk, I present a coarse-grained simulation of periodic pattern formation in multicellular organisms based on cellular oscillators (CO). An oscillatory process within cells serves as a developmental clock whose period is tightly regulated by cell-autonomous and non-autonomous mechanisms. A spatial pattern is generated as a result of an initial temporal ordering of the cell oscillators freezing into spatial order as the clocks slow down and stop at different times or phases in their cycles. When applied to vertebrate somitogenesis, the CO model can reproduce the dynamics of periodic gene expression patterns observed in the presomitic mesoderm. Different somite lengths can be generated by altering the period of the oscillation.  There is evidence that a CO-type mechanism might also underlie segment formation in short-germ-band insects. This suggests that the dynamical principles of CO-type mechanisms might be conserved throughout evolution although most of the genes involved in segment formation differ between distant phyla. Course-grained models, like the CO model presented here, could prove helpful to study the evolution of developmental dynamics in the vast majority of biological systems for which little or no molecular data is available.

# Creating and Analysing in Silico Bacterial Cells

Igor Goryanin
Cell Simulations & Pathway Modeling, GlaxoSmithKline

Dr. Igor Goryanin
Head, Cellular Simulations and Pathway Modelling
GlaxoSmithKline,
Stevenage,
SG1 2NY
e-mail iig2468@gsk.com
phone 44-1438-764197
fax 44-1438-764918

A unified representation for the cellular networks has been designed to model the total cell behaviour.

For genetic networks we have developed original genetic regulation classification and compiled expression tables for transcription regulation in prokaryotes. Special algorithm has been developed to convert expression tables boolean logic to the genetic stoichiometric matrix . This matrix is used for storing all elementary genetics events. A protein interaction Stoichiometric Matrix contains information about all protein/protein interactions, and Metabolic Stoichiometric Matrix comprises all metabolic events. The whole cell model can be obtained by merging these three matrices. Four original DTD schemas have been designed to store information required for simulations, i.e.interactions in genetic, metabolic, protein interactions networks (including signal transduction) and metabolites thesaurus.

We have chosen well known E.coli as a first organism for total simulation, collected all available information about E.coli interactions, created corresponding XML files, and then downloaded these files to a new version of Dbsolve 6 software. This approach allows us to simulate three types of the cellular networks simultaneously or/and independently. We are able to apply the mathematical methods and technique developed earlier for metabolic networks to other types of cellular networks. It includes connectivity analysis, concentration balance analysis, FBA, MCA, transient behaviour, steady state simulations, bifurcation analysis and pathway optimisation.

http://website.ntl.com/~igor.goryanin
ftp://ftp.cds.caltech.edu/pub/goryanin/
http://emp.mcs.anl.gov

# Limitations of Modeling in Biology: In Search of a Free Lunch

Minija Tamosiunate
Vytautas Magnus University, Kaunas, with J. Rimas Vaisnys

In Informatics one mainly transforms information from one form to another, and only in exceptional circumstances generates new information. This must be the explanation for the emphasis displayed at the sessions on joining and reorganizing the various data bases describing different aspects of genes, proteins, and organisms. Such efforts can influence the ease with which certain work is carried out by an order of magnitude, and are indeed worthwhile. But such efforts are at best of indirect use in two primary concerns of biologists: how to understand the systems which are of interest, and what to do when the needed information is either missing or is actually incorrect.

Understanding is really a state of mind of the person doing the understanding, but it is fostered by certain formal constructions, often called models, which describe the behaviors of interest. In a systems approach to describing the modeling processes one introduces certain entities. The first is the system of particular interest, characterized not only by the observed behavior (usually classified into outputs $y(t)$ and inputs $u(t)$) but also by the state function $s(t)$, which serves to relate the behaviors of the system in a specific way: $s'(t)=f[s(t),u(t)]$, $y(t)=g[s(t),u(t)]$. The second is the environment surrounding the system, which provides the input seen by the system and also receives the outputs generated by the system. The goal of the person engaged in modeling is to organize the empirical information about a biological system into the above stated form: one must choose the input and output variables, thereby implicitly identifying the system and the environment, and then identify the state variables and the two functions $f[\ ]$ (the state development function) and $g[\ ]$ (the output function). The latter are often specified in parameterized form, with different classes for different types of systems. For the sake of completeness we point out the two main activities of the modeling process: The first is to take information drawn from observations about the behavior of the system and evaluate the needed functions as specified above. This is technically an inverse problem. The second is to take models so constructed and to explore the consequences of joining such models to either novel environments or to other similar systems, or both. This is technically a direct problem, either in analysis or synthesis. Observe that only open system models are interesting in this sense. But in all the activities it must be noted that the required information is of a very specific kind - just any information will not do. A major reason for introducing a specific modeling framework is that the phrasing of the model defines what and how information is to be used in solving the biological problem.

One problem is how to detect "incorrect" data, and what to do with it. For purposes of this discussion we characterize "incorrect" data as being of two kinds: problems arising from the measuring, recording and transforming processes, and problems arising from the system and system model themselves. With regard to the first problem, the best approach is to admit that such errors are possible and to include appropriate verification and consistency checks. With regard to the second problem one would like to apply the gold standard in science: reproducibility of results when the same observational process is repeated. This is expensive, so that often one substitutes comparisons between data extrapolated in different ways, and this in turn already depends on assumptions which are best phrased in terms of the models that are being considered; the data themselves rarely tell their own story. One of the most fundamental assumptions is with regard to the nature of the variables and parameters: are they deterministic or are they probabilistic; if the latter, reproducibility must refer to their distribution functions, not to their realizations. In any event, for open systems, it is crucial that reproducibility include the state of the environment; there is an unfortunate trend to assume that when environmental variables are difficult to measure they must be constant (and take on values most convenient to the scientist). If one fails to appropriately account for the environment, or has misidentified system functions and states, one will observe data that appear to be "incorrect".

Biological systems are large and complex: in terms of the number and kinds of variables required to describe their behavior, in terms of the time and size scales involved, in terms of the types of entities involved. The sheer size, in the number of variables involved, raises problems both for modeling and for observational data collection. There are many methods for doing the computations required to both fit models to data and to explore the behavior of the resulting models, but most of them have been tested and found adequate only with small problems. Demands both on the amount of observational information and its accuracy also increase sharply with size. To make observations and modeling possible one is forced to partition the systems into subsystems, to organize the processes in a hierarchical way, and to do this without the benefit of understanding how the systems function. In such a situation trial-and-error modeling can probably be shown to be optimal, so that there is no systematic recipe. Very often the choice of hierarchical levels can be guided by time and space scale arguments, but these do not provide unique specifications. If the lower level models can have similar structures and behaviors (described by perhaps many parameters, but of similar classes), then the higher level models may benefit from reduced sensitivity to lower level model details. It is advisable to define subsystems that minimize the number of input/output interactions between them or with the environment, provided the observations are sufficient to define the state variables. The modeling process must also reflect the goals of the process. For example, if one is simply seeking the simplest description of a subsystem, it may very well be best to include feedback processes within the subsystem. If one is seeking to find ways of controlling a system, it may well be best to have feedbacks represented in subsystem inputs and outputs.

The aim of the talk was to further discussion about the possibilities and constraints of modeling in biology, especially those arising when passing from the smaller to the larger scale phenomena (in the direction from genotype towards phenotype). The talk tried to structure questions and ideas raised in the presentations and discussions of other participants in the course of the seminar.