Modeling and Simulation of Gene Regulation and Metabolic Pathways

June 21 - 26, 1998

organized by

Julio Collado-Vides (Cuernavaca, Mexico)
Ralf Hofestädt (Magdeburg, Germany)
Michael Mavrovouniotis (Evanston, USA)
Gerhard Michal (Tutzing, Germany)

Preface

Modeling and Simulation of Gene Regulation and Metabolic Pathways

June, 21-26, 1998. International Conference and Research Centre for Computer Science, Schloss Dagstuhl, Saarland, Germany

Julio Collado-Vides, Ralf Hofestädt, Michael Mavrovouniotis and Gerhard Michal

The second Dagstuhl seminar for *Modeling and Simulation of Gene Regulation and Metabolic Pathways* was held from June, 21 to 26, 1998. It was a multidisciplinary seminar with 59 participants from 15 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 1998 meeting is a sequel to the 1995 Dagstuhl seminar on the same topic. Both were generously supported by grants from the Volkswagen Stiftung and the European Community (TMR Grant). The availability of a rapidly increasing volume of molecular data enhances our capability to study cell behavior. In order to exploit molecular data, one must investigate the link between genes and proteins; the link between protein structure and protein function; and the concerted effects of many proteins acting on, and interacting with, the mixture of small and large molecules within a cell. This last step is the study of gene regulation and metabolic pathways which was the topic of the Dagstuhl seminar. The molecular data must be stored and analyzed. Database systems for genes and proteins (EMBL, GENBANK, PIR, SWISS-PROT) offer access via internet. In the research field of molecular biology this technique allows the analysis of metabolic processes. To understand the molecular logic of cells we must be able to analyze metabolic processes in qualitative and quantitative terms. Therefore, modeling and simulation are important methods. They influence the domain of medicine and (human) genetics - the microscopic level. Today integrative molecular information systems which represent different molecular knowledge (data) are available. The state of the art is shown by P. Karps system EcoCyc, which represents the metabolic pathways of E. coli. For every gene or protein within a specific metabolic pathway, EcoCyc presents the access to all corresponding genes and/or proteins. Moreover, the electronical information system KEGG represents all biochemical networks and allows the access to the protein and gene database systems via metabolic pathways. However, both systems are based on the idea of the statical representation of the molecular data and knowledge. The next important step is to implement and integrate powerful interactive simulation environments which allow the access to different molecular database systems and the simulation of complex biochemical reactions. Molecular information systems for gene regulation and metabolic pathways were one topic of the Dagstuhl seminar. The idea was to discuss the progress of this research field and the

integration of the molecular database systems in combination with simulation tools. The organisers of the seminar invited colleagues, who presented their ideas through 42 talks and computer demos. More than 30 years ago Gerhard Michal started to collect all biochemical reactions. His classification is presented by the Boehringer pathway chart. This data collection was extended by the KEGG research group, which implemented the first electronical representation of this data in 1996. Nowadays all biochemical reactions are available via internet using the KEGG system. KEGG represents links to molecular database systems for genes, proteins, and enzymes, which are elements of metabolic pathways. Thus a link to the EMBL database system represents more information about a specific gene, and a link to the SWISS-PROT system represents more information about the protein (enzyme). Regarding the KEGG system the representation of quantitative data and kinetic data is not available today. Furthermore, additional to the molecular data (genes, proteins, and pathways) the first molecular information systems are available which represent data of the cell signals. Besides the Japanese Cell Transduction Database the GENENET database system is available. Taking regard to both molecular information systems this can be interpreted as the first scientific step in which cell reaction processes are surveyed from the gene regulation process to the cell communication. For molecular biology the phenomena of gene regulation is the main question. The systematic discussion of this question is based on the electronical representation of the molecular knowledge, which allows the complex analysis of this data. For that reason specific database systems are implemented (OperonDB, TRANSFAC and TRRD). These database systems represent all known operons and the transcriptional factors for E. coli (OperonDB) and eukaryotic cells. Today, two research fields based on this data are supported: The prediction of promoter sequences and the modeling of gene regulation. The prediction of promoter sequences is of importance, because the promoter is the starting signal for a structure gene which represents the genetic information. The human genome project will sequence the whole genome until the year 2004 ($64 * 10^9$ base pairs). The next step is to calculate the corresponding genetic map. Therefore, sequence pattern matching algorithms must be developed and implemented. In addition modeling and simulation of gene regulation processes will support the systematic analysis of the metabolic pathways.

John Reinitz opened the seminar. He presented ideas about modeling of genetic factors and analyzed the process of segment determination in Drosophila through numerically inverting a chemical kinetic equation which describes the regulatory circuitry and accounts for the synthesis rate, diffusion, and decay of gene products. The molecular mechanisms of gene regulation were presented by **Edgar Wingender**. During the last decade he has been analyzing the molecular mechanisms of eukaryotic gene regulation and has been collecting all transcriptional factors which can be found using his database system TRANSFAC. The predic-

tion of promoter sequences based on this data was one important topic of the gene regulation session. **Julio Collado-Vides**, **Gary Stormo**, and **Thomas Werner** showed algorithms for the detection of promoter sequences for E. coli and eukaryotic cells. The molecular mechanisms of the cell death were discussed by **Dominique Bergeron**, and **Luiz Mendoza** talked about complex metabolic networks. The modeling of regulatory networks belongs to the topic of Biophysics and Biomathematics. Moreover, discret models are developed using methods of Bioinformatics. At the beginning of that session **Jay Mittenthal** presented the metabolic pathway of the Pentose Phosphat Cyclus. **Gerhard Michal** is the creator of the Boehringer pathway chart which inspired many of us to pursue databases and integrative methods for the study of the metabolism. In his talk he discussed a brief overview of the issues surrounding the development of graphical representations and displays of metabolic pathways and other biological information. In the case of analytic models **Michael Savageau** introduced a model which allows the simulation of complex kinetic effects. Using graph theoretical methods **Michael Kohn** discussed his model for the simulation of metabolic networks. **Stefan Schuster** outlined several powerful methods for determining key features of a metabolic pathway or network. He showed how conservation relations may be identified and how elementary biochemical routes (and hence the spectrum of behaviors of the biochemical network) may be determined. Further he outlined the principles of metabolic control analysis and its extensions. A new grammatical model for the analysis of complex metabolic processes was presented by **Simone Bentolila**. Another topic of the seminar were molecular database systems. At the beginning of this session **Thomas Mück** discussed new topics in the research field of database systems and **Vladimir Babenko** introduced new techniques for the integration of molecular database systems. **Minor Kanehisa** showed the pathway database system KEGG and discussed further applications. **Fedor Kolpakov** demonstrated the database system GENENET, which is similar organized to the Japanese database system for Cellular Signal Transduction, which was presented by **Takako Takai-Igarashi**. **Rolf Apweiler** talked about the SWISS-PROT database, and **Daniel Kahn** demonstrated a new database system for the integration of protein knowledge. One important application of this molecular data is the diagnosis of metabolic diseases. In the case of inborn errors **Manuela Prüss** introduced the database system MDDB. The final topic of the seminar was the integration and simulation of metabolic networks. The first generation of powerful simulation environments for the metabolic network control was discussed. These tools work using the biochemical data and diverse models which were presented in the sessions mentioned before. **Pedro Mendes** demonstrated his simulation environment GEPASI, which allows the analytical modeling of the metabolic processes. A first information system based on the integration of molecular databases and a grammatical simulation environment was introduced by **Uwe Scholz** and **Ralf Hofestädt**. Finally, an expert system for the modeling of metabolic processes was presented by **Jaime Lagunez**.

Concluding remarks

It is not sufficient to know what each protein or gene does in the cell (it usually catalyzes or regulates a biochemical reaction), but one must also decipher what they are all doing together (they form pathways of elaborate transformations and regulatory networks). In order to decipher the metabolic pathways that define the behavior of the cell as a whole, one must use information on single-protein activity. But there is also information flow in the reverse direction: The position and role of an enzyme in the metabolic network provides crucial insights and hypotheses for its genetic regulation and its relationship to other proteins. Genes and proteins are routinely sequenced and stored in database systems. Data on biochemical pathways has been systematically collected for the last three decades (in pictorial and text form), and the accumulation of such data has increased dramatically in recent years (and shifted to computational representations). The systematic use of collected data is also continually making advances. Methods for computational modeling and simulation are made feasible by the availability of data and are driven by the need to understand the behavior of complex biological systems. The integration of information, especially combinations of genes, enzymes, and metabolic pathways will be necessary in the study of biological regulatory structures, which usually involve multiple facets, components, and scales of action. Database systems and powerful models are already available, and the first practical simulation tools are implemented based on powerful theoretical methods. These information-integrative activities will become increasingly shed light on the biochemical mechanism of life. The actual questions of the seminar were focused by the final discussion which concluded that: The number of molecular database systems is increasing. Moreover, these systems are available via internet. The now available accessing techniques are www links to the relevant molecular database systems, which support the navigation through the molecular data. However, this data must be available for further analysis processes. The detection of promoter structures is one actual example, which shows also the algorithmic problems of this research field. Besides the algorithmic analysis, modeling and simulation based on this molecular data are of importance. Different tools are developed and implemented. However, the selection of the model depends on the actual question. The main task for the next years is the integration of the database systems and the simulation environments, which will allow the simulation of complex metabolic networks.

Acknowledgement The organisers thank the Volkswagen Stiftung and the European Community (TMR Grant) for its generous financial support.

Further information about the Dagstuhl seminar:
http://wwwiti.cs.uni-magdeburg.de/iti_bm/dagstuhl/

Julio Collado-Vides is with the Centre for Nitrogen Fixation, National Au-

tonomous University of Mexico (UNAM), Cuernavaca, A.P. 565-A, Morelos, Mexico; **Ralf Hofestädt** is with the Department for Computer Science, Otto-von-Guericke-University Magdeburg, D-39116 Magdeburg, Germany; **Michael Mavrovouniotis** is with the Department of Chemical Engineering and the Council for Dynamic Systems and Control, Northwestern University, Evanston, Illinois 60208-3120, USA; and **Gerhard Michal** was with the Research Department of the Boehringer Company, Kreuzeckstr.19, 82327 Tutzing, Germany.

Contents

Preface **3**

John Reinitz
Solving the Inverse Problem in Gene Expression: Lessons from Drosophila **11**

Gary Stormo
Discovering Regulatory Sites from Expression Data **12**

Edgar Wingender
Database Modeling of Gene Regulatory Pathways **13**

Thomas Werner
Promoter classification by functional organizational models **14**

Andreas Dress
The Family of bHLH-Proteins **15**

Michael Savageau
Development of Fractal Kinetic Theory for Enzyme-Catalyzed Reactions and Implications for the Design of Biochemical Pathways **17**

Michael Kohn
Identifying Sites of Metabolic Regulation by Graph Theoretical Modeling **18**

Wolf-D. Ihlenfeldt
A General System for the Simulation of Organic Reactions **19**

Gerhard Michal
Regulation of Metabolic Pathways **19**

Stevo Bozinovski
Protein Biosynthesis: The Flexible Manufacturing Metaphor **20**

Edgar Wingender
The TRANSFAC-Database **21**

Michael Kohn
A Demonstration of MetaNet Graph Drawing and Analysis Software **21**

Jacky Snoep
A Control of DNA supercoiling in the complex cell **22**

Jay Mittenthal
Designing metabolism: Alternative connectivities for the pentose phosphate pathway **23**

Rolf Apweiler
The SWISS-PROT and TrEMBL Protein Sequence Database as a Tool to Model Regulatory and Metabolic Pathways **23**

Bruno Sobral
A Plant Metabolism Database **26**

Minoru Kanehisa
Regulatory Pathways in KEGG **27**

Dominique Bergeron
The death factors: a combinatorial analysis **28**

Terry Gaasterland
Multi-Genome Views of Whole Genomes with a focus on E.coli global regulatory proteins **29**

Julio Collado-Vides
Regulatory Predictions in the Complete E.coli Genome **31**

Pedro Mendes
Integration of gene expression with metabolism in kinetic simulations **32**

Luis Mendoza
The regulatory network that controls flowering in Arabidopsis **34**

Takako Takai-Igarashi
Cell Signaling Networks Database **35**

Pedro Mendes
Exploring biochemical models with Gepasi, a kinetics simulator **35**

Uwe Scholz
Molecular Database Integration: Analysis of Metabolic Network Control **37**

Thomas Mück
Indexing and Retrieval of Complex Data Sets or Is it a good idea to store metabolic data in an ooDB? **38**

Vladimir Babenko
Databases integration and automatic knowledge acquisition on regulatory regions of eukaryotic genome **38**

Daniel Kahn
Integration of protein data : ProDom, XDOM and genome projects **39**

Fedor Kolpakov
GeneNet: a Database for Gene Networks and its Automated Visualization through the Internet **40**

Hiroyuki Ogata
Integrated Analysis of Metabolic Pathways, Sequence Evolution and Genome Organization **42**

Manuela Prüß
The Metabolic Diseases Database **44**

John Reinitz
The GeNet Database **44**

Thomas Dandekar
Examples on post-transkriptional reguation and metabolic pathways **45**

Klaus-Peter Zauner
Simulation Experiments on the Role of Spatial Arrangement in Enzymatic Networks **45**

Tom Shimizu
E-CELL vs StochSim: System-wide and molecularly-detailed approaches to simulation of cellular processes **46**

Masahiro Okamoto
Towards a Virtual-Lab-System for Metabolic Engineering: Development of Biochemical Engineering System Analyzing Tool-Kit (BEST-KIT) **47**

Simone Bentolila
Modeling signal pathways **49**

Ralf Zimmer
Mapping Metabolic Networks and Gene Expression Data via Protein Structure Prediction **51**

Jacques van Helden
Computer tools for the analysis of yeast regulatory sequences **52**

Edda Klipp
Evolutionary Optimization and Metabolic Control Analysis **53**

Stig Omholt
Why and how to build a conceptual bridge between mechanistic regulatory biology and quantitative genetics **55**

Stefan Schuster
Computer-aided Structural Analysis of Biochemical Reaction Systems **56**

Takayoshi Shoudai
Parallel Knowledge Discovery System for Amino Acid Sequences - BONSAI Garden **57**

Patrizio Arrigo
Application of Conceptual Clustering to the Recognition of the Hierarchical Structure of Transcriptional Control Domains **58**

Jürgen Sühnel
Hydrogen Bonds in Biopolymer Structures - Variations on an Old Theme **58**

Jaime Lagunez-Otero
The Cell as an Expert System **60**

Falk Schreiber
Visualization of Biochemical Pathways **60**

Solving the Inverse Problem in Gene Expression: Lessons from Drosophila

John Reinitz, Brookdale Center for Molecular Biology, New York
In collaboration with C. Alonso, K. Chu, Y. Deng, D. Kosman, and D. H. Sharp

This talk describes recent progress in a long term project devoted to solving fundamental problems in animal development. We use the process of segment determination in the fruit fly Drosophila melanogaster as a model system. Segment determination is the formation of a stable chemical blueprint for the segmentation pattern of the animal. The segments are determined to a resolution of one cell in about 45 minutes, so this process is both rapid and accurate. Segment determination takes place during a period of time in which the embryo is composed of a hollow shell of nuclei: cells have not yet formed. There are 4 classes of segmentation genes: in the talk I am chiefly concerned with gap genes like Kruppel, which are expressed in one or two broad domains. Other genes, of the pair-rule class, are expressed in seven stripes. Most pair-rule genes require input from both gap genes and other pair-rule genes to make stripes, but one pair-rule gene, called eve, can make stripes from gap gene input alone. We can ask why eve is the only pair-rule gene with this property using a four-fold approach. 1. Formulate a theoretical model. The rate of change of the concentration of the product of gene a in nucleus i is given by a kinetic equation with three terms. The first term on the right hand side of the equation describes gene regulation and protein synthesis, the second describes exchange of gene products between neighboring cell nuclei, and the third represents the decay of gene products. The parameters in the equation must be determined from data. To do that we 2. Generate gene expression data. We raise antibodies to the segmentation genes, scan them with a confocal microscope, and use image segmentation and computer vision methods to obtain a quantitative numerical dataset at cellular resolution. Then we take this data and 3. Do large scale fits. We fit the solutions of the equations to data. We do the fits by simulated annealing. I describe a new algorithm for simulated annealing based on two basic principles. First, all statistics concerning energy and variance are pooled among all processors. Second, a periodic mixing step is performed in which a given processor takes on the state of another processor with Boltzmann probability. We show that with appropriate mixing intervals, the algorithm performs at 100% parallel efficiency for up to 50 processors and and 80% parallel efficiency for 100 processors. 4. Validate the model. That is, we use it to learn new biology. First, the property that gap genes can encode pair-rule stripes only in the eve position is demonstrated to be an implicit property of the model, which had only gene expression patterns as explicit input. I explain how this property follows from the arrangement of gap domains in the embryo. This

analysis shows that:

1. Pattern forming information is transmitted from gap to pair-rule genes by means of a non-redundant set of morphogenetic gradients and

2. The stripe forming capability of the gap genes is constrained by the arrangement of these gradients, and also by the fact that each gap domain consists of a pair of correlated gradients.

I close with an inference about evolutionary development. We argue that the constraints on gap gene architecture are a consequence of selective pressures that minimize the number of gap genes required to determine segments in long germ band insects.

Discovering Regulatory Sites from Expression Data

Gary Stormo, University of Colorado, Dept. of MCD Biology, Boulder, USA

Expression arrays, or "DNA chips", provide a means of identifying sets of genes that are co-regulated. Such a set implies that there should be regulatory proteins, such as transcription factors, that control the set, and there should be sites occuring in the adjacent DNA sequences for those factors to bind. Therefore we should be able to apply some pattern recognition methods to identify what the common sites are that are responsible for the co-regulation. Several methods have been developed to help identify such sites when the appropriate data exist. We have used models for protein-DNA interaction that are embodied in a "weight matrix". This is a simple matrix with a weight associated with each possible base at each position in the binding site. Under appropriate conditions those weights can be made proportional to the free energy contribution of the base at the position; while this model is simple it has been shown to be a reasonable approximation in several cases. So the problem we're interested in reduces to finding the most statistically significant weight matrix that is in common to the set of genes. Several years ago we described a greedy algorithm to accomplish that task, and in recent years it has undergone a number of refinements and improvements. An Expectation-Maximization (EM) algorithm has also been developed for this problem, originally by Lawrence and colleagues and then by several other groups. Recently we have explored the use of simple neural networks for this problem. Because these interactions are often well modelled by a weight matrix, out neural network can be a simple perceptron with on hidden

layer; each weight of the perceptron corresponds to one weight in the matrix. The objective function of our net is to maximize the specificity of the protein. Given a set of weights we can predict the binding energy to any sequence. So given a complete genome we can compute the partition function for that genome. The object is to find a set of sites upstream of the co-regulated genes, and a set of weights describing their binding energies, such that they have high probabilities of being bound by the protein. The probability takes into account the partition function in the natural way, so this method uses as its objective function a good approximation to the equilibrium thermodynamics of the system. If we make some simplifying assumptions about the genome we can calculate the partition unction analytically. And under these assumptions the weights that maximizing the binding probabilities is the same as the weights used in the "information content" analysis of the sites. That is, under those assumptions the neural network method has exactly the same objective function as the greedy and EM algorithms, but its approach to the solution is much different. Therefore we often use all three methods as a check to see whether we have obtained suboptimal solutions. The neural network method has an advantage over the other approaches because we are not forced to make assumptions about the genome. That is, we can calculate the partition function exactly, or approximate it closely, and then maximize our function by some optimization method. This has been shown to be useful in several cases. In particular, we can use as the "background" sequences that we wish to discriminate against a particular subset of sequences rather than the whole genome. For example, we may know that a large set of genes have some regulatory factors in common, but also can be divided into distinct subsets that have different behaviors. Then we can use one set as the "positive" set and the other as the "negative" set and find the patterns that are both in common to the positive set and also serve to distinguish it from the negative set. Enhancements of this approach can allow us to find common elements in RNA sequences, where the important information is a combination of sequence and structure constraints. Other refinements, to allow the patterns to have gaps, as in general Hidden Markov Models of aligned sequences, can be put into the same general framework without too much additional difficulty.

Database Modeling of Gene Regulatory Pathways

Edgar Wingender, Molecular Bioinformatics of Gene Regulation, GBF, Braunschweig, Germany

Important components of the basic bioinformatics infrastructure for genomics and proteomics projects are databases such as EMBL/GenBank/DDBJ, SwissProt and PIR providing sequence data along with some basal annotation as static bio-

logical objects. As an important step towards "functional genomics", we need now databases which model biological mechanisms. One example is the TRANSFAC database whose major goal is to model specific DNA-protein interactions which are of regulatory importance. It also includes data about the regulation of transcription factor activities as well as information about their expression profiles. Presently, attempts are being made which aim at modeling these data in specific database modules to assign to each regulator (transcription factor) a certain "expression matrix" in a multi-dimensional spatio-temporal-conditional space. The conditional "dimension" is modeled as an object-oriented database about signal transduction pathways (TRANSPATH), which will be developed further in close cooperation with CSNDB (see contribution of T. Takai-Igarashi, NIHS, Tokyo). The time axis is given as one table of defined stages of (human) embryonic development. Also for the human system, a relational database system about cell/ organ/ tissue types has been established which together with the integrated time table enables us now to systematically map expression patterns. While these systems allow to represent cell- and stage-specific signaling pathways, loops are going to be implemented for the regulation of those genes which by themselves encode regulators (i. e. transcription factors, more upstream components of signal transduction pathways or extracellular inducers). This will enable us to model regulatory networks from the contents of the databases described above.

Promoter classification by functional organizational models

Thomas Werner, Institute of Mammalian Genetics, GSF-National Research Center for Environment and Health, Neuherberg, Germany

Due to the enormous amount of new genomic sequences it is mandatory to preselect candidate sequences by computerized analysis prior to experimental functional analysis. This includes prediction of exons and introns as well as the identification of potential regulatory regions which usually encompass multiple regulatory elements that exert their regulatory function only within the correct context. Last year, we reported our approach to this problem and presented sucessful identification of a new LTR as an example. We have now extended our work aiming at the prediction of inherent tissue and/or cell specificity of such regions. Actins comprise one of the most commonly expressed gene families in mammalian tissues. Yet there are specialized actin genes which are either preferentially or exclusively expressed in all or only subsets of muscle cells. These expression patterns are mostly controlled at the level of transcription as is known

from Jim Ficketts work (and his excellent web-site) about muscle-specific gene expression. Therefore, the muscle-specificity of particular actin genes is most likely encoded in their promoter sequences although the most prominent muscle-specific transcription factors MEF2 and MyoD are apparently not crucial in this case although present in some of these promoters. Here, we present a study focusing on the specific recognition of actin promoters in general as well as muscle-specific actin promoters. We developed a general actin promoter model starting from a general analysis of the correlation of transcription factor binding sites (TF-sites) with these promoters and identified candidates for crucial TF-sites. Our model consists of 6 different elements and was developed on a training set of 11 sequences. This training set was already to heterogeneous in sequence to allow identification by FASTA analysis. The model could be refined by addition of another SRF binding site and this muscle-actin specific model does not recognize most of the other muscle-specific promoters indicating that there are several independent ways to achieve muscle-specificity of a promoter. This demonstrates that specific promoter recognition against a vast background of anonymous sequences is pricipally possible and that tissue specificity can be achieved by minor changes in a more general promoter structure.

The Family of bHLH-Proteins

Andreas Dress, Fakultät für Mathematik, Universität Bielefeld, Germany

Many biological processes are spatially and temporally controlled at the level of transcription. To understand the transcriptional regulation of gene expression, one needs to decipher the molecular modes of differentiation and development of eukaryotic cells. Transcriptional control is mediated by complex interactions between regulatory transcription factors with their various enhancer elements giving rise to sequence-specific multiprotein complexes that control gene expression at multiple control points. Hence, it is crucial that we understand the structure of the various components of these transcriptional complexes, are able to classify their components into well-defined categories, and understand their origin and evolution. Transcription factors are structurally complex proteins containing distinct functional components associated with DNA binding, protein oligomerization, phosphorylation, activation and other activities. As a consequence, functionally heterogeneous proteins are often classified based upon small, highly conserved amino acid domains which are discrete connected parts of proteins that can be equated with a particular function. Thus, transcription factors are generally grouped into families like zinc fingers, helix-turn-helix, helix-loop-helix or basic leucine-zippers because the relevant proteins share a particular, short do-

main associated with DNA binding, oligomerization or other activities. Several problems are inherent to evolutionary classifications based on domains. First, the domains are often short and highly conserved so that the amount of information contained within them that can be used for classification, may be small. Complicating the issue is the fact that outside the conserved domain, these proteins may exhibit considerable sequence dissimilarity to the point of being apparently unrelated. Second, these domains are associated with a limited number of functions like DNA binding or oligomerization. Mechanistically, there may be only a few ways to solve a particular problem. As a consequence, convergent evolution often can not be excluded, particularly for structurally simple domains, e.g., the structurally equivalent E-box and G-box domains involved with DNA binding, or the leucine zipper oligomerization domain. Third, the definition of the domains in terms of primary sequences are not well understood so that determining whether a particular protein should be included in one of these families is sometimes difficult (e.g., zinc finger proteins). Consequently, detailed analyses are needed to characterize rigorously the structure and function of these important domains and to deduce their origin and evolution. Such studies require large amounts of divergent data to better elucidate their structural and functional limits as well as to explore the constraints regarding their evolution. In the lecture, we examine some structural aspects of the basic helix-loop-helix domain (bHLH) which defines an important group of transcription factors. bHLH proteins are characterized by highly conserved bipartite domains for DNA binding and protein-protein interaction. Proteins containing the evolutionarily conserved helix-loop-helix domain are an important class of regulatory components in transcriptional networks of many developmental pathways. They are involved in regulation of neurogenesis, myogenesis, cell proliferation and differentiation, cell lineage determination, sex determination and other essential processes in organisms ranging from plants to mammals. These various proteins can be grouped into clades and groups reflecting their evolutionary history. Since the bHLH domain was first described, a large number of helix-loop-helix proteins have been identified. Most are classified as bHLH transcription factors based on overall sequence similarity with existing bHLH proteins. Several important questions exist regarding the structure of the domain and sequence variability in bHLH proteins.

1. What primary sequence structure identifies a helix-loop-helix protein and how does this structure vary among related proteins?

2. How much sequence variability is permitted while still preserving the necessary helix-loop-helix configuration?

3. Which sites are most highly conserved?

4. What dependencies exist between the amino acid distribution observed at variable sites and clade membership, loop length, and the existence of a leucine zipper?

5. Are there significant associations between the function(s) of these residues and the extent of their evolutionary conservation and/or coevolution?

Consequently, the goal of our analyses is to examine the extent of primary sequence variability in a large number of functionally diverse bHLH proteins, to suggest a short hypothetical motif that will serve as a predictive model for identifying putative bHLH proteins, and to explore the goodness of fit of this motif to a wide variety of known and of previously unrecognized bHLH proteins.

Development of Fractal Kinetic Theory for Enzyme-Catalyzed Reactions and Implications for the Design of Biochemical Pathways

Michael Savageau, Department of Microbiology and Immunology, Michigan, USA

Recent evidence has shown that elementary bimolecular reactions under dimensionally-restricted conditions, such as those that might occur within cells when reactions are confined to two-dimensional membranes and one-dimensional channels, do not follow traditional mass-action kinetics, but fractal kinetics. The power-law formalism, which provides the context for examining the kinetics under these conditions, is used here to examine the implications of fractal kinetics in a simple pathway of reversible reactions. Starting with elementary chemical kinetics, we proceed to characterize the equilibrium behavior of a simple bimolecular reaction, derive a generalized set of conditions for microscopic reversibility, and develop the fractal kinetic rate law for a reversible Michaelis-Menten mechanism. Having established this fractal kinetic framework, we go on to analyze the steady-state behavior and temporal response of a pathway characterized by both the fundamental and quasi-steady state equations. These results are contrasted with those for the fundamental and quasi-steady state equations based on traditional mass-action kinetics. Finally, we compare the accuracy of three local representations based on both fractal and mass-action kinetics. The results with fractal kinetics show that the equilibrium ratio is a function of the amount of material in a closed system, and that the principle of microscopic reversibility has a more general manifestation that imposes new constraints on the set of fractal kinetic orders. Fractal kinetics in a biochemical pathway allow an increase in flux to occur with less accumulation of pathway intermediates and a faster temporal response than is the case with traditional kinetics. These conclusions are obtained regardless of the level of representation considered. Thus, fractal kinetics provides a novel means to achieve important features of pathway design.

Identifying Sites of Metabolic Regulation by Graph Theoretical Modeling

Michael Kohn, Laboratory of Computational Biology and Risk Analysis, National Institute of Environmental Health Sciences, USA

Many children are born with defects in metabolism owing to inheritance of a deleterious mutation in a gene for a particular enzyme. Modeling the affected pathways of intermediary metabolism can yield insights into regulatory mechanisms that can assist in the design of effective therapies for such individuals. Because the parameters of a kinetic model may not be known with sufficient precision, a useful modeling strategy must be robust with respect to uncertainties in parameter values and qualitatively indicate sites of regulation. A graph theoretical representation similar to familiar metabolic pathway flowcharts has been developed. Nodes of the graph, joined by directed arcs, represent the chemical species and their reactions. Formal operations on the graph identify feedback cycles. The set of reaction steps with the fewest members whose deletion would simultaneously sever all the feedback loops identifies the critical feedback chemicals. The enzymes which make or consume the feedback chemicals set the concentrations of those intermediates and, hence, control the influence of the feedback regulators. If estimates of binding constants, enzyme activities, and chemical concentrations are available, the contributions of the feedback chemicals and controlling enzymes can be ranked in order of their effectiveness. This strategy is illustrated by a graph model of the urea cycle and associated amino acid metabolism in human liver. This pathway is the major route for elimination of nitrogen derived from breakdown of protein. The model identified glutamate as a major controller of urea production and suggested that increasing the availability of the acceptor for glutamate nitrogen would have the greatest effect on urea production. Indeed, clinical observations indicate significant improvement in patients deficient in one of the cycle enzymes by providing extra citric acid cycle substrate in the diet.

A General System for the Simulation of Organic Reactions

Wolf-D. Ihlenfeldt, Computer Chemistry Center, University of Erlangen-Nürnberg, Germany

We have recently finished a new version of our reaction prediction system EROS. Besides numerous improvements in the internal representation of chemical structures, it allows the simulation of different ways of running a reaction: from laboratory batch reactions, degradation of compounds in the environment, all the way to the modeling of the reactions occurring in a mass spectrometer. This could be achieved by introducing very general and versatile concepts of reactors, phases and kinetic modes for running a reaction. Building on these concepts, reactions in various kinds of reaction vessels, including cellular compartments, can be handled Processes such as the events in the uptake, release and metabolism of pharmaceuticals adminstered at certain intervals including the pharmacokinetics can be modelled. It allows the integrated study of several different, but linked processe such as the generation of the products of combinatorial chemistry experiments with the concomitant simulation of the mass spectra of all products. A variety of examples is given, including models of combined enzymatic and spontaneous reactions as occurring in soil chemistry.

Regulation of Metabolic Pathways

Gerhard Michal, Tutzing, Germany

A short survey of the various systems of metabolic regulation is given. Regulation can proceed via changes of the enzyme activity or via changes of the amount of enzyme. Examples for such situations and their kinetic treatment are presented. Most simple is the effect of moderate substrate concentration on the reaction velocity: The ratio of turnover by 2 parallel enzymes depends on the kinetic constants of the enzymes. Enzyme inhibitors can act competitively or non-competitively in a reversible way. This causes different responses to changes in the substrate concentration. Irreversible inhibitors reduce the amount of active enzyme, but the kinetic constrants of the remaining enzyme are not changed. Enzyme control by allosteric mechanisms can be homotropic or can be effected by activators or inhibitors. Beyond a strictly phenomenological description, the symmetry and the sequential models allow a more detailed discussion, altrough the actual situation often containts elements of both. Covalent modification of enzymes is frequently used in biological systems to adapt enzyme activities to

environmental changes or to variations in supply and demand. Frequently cascades of such regulation systems exist in order to potentiate the effects or to modify and fine-tune the responses. While these are usually reversible systems, enzyme activation by cleavage of precursors is one-way process. Variation of the amount of enzymes can be achieved by regulation of their degradation as well as of their synthesis. The latter can take place at all levels of protein synthesis: by enhancing or repressing transcription, by influencing the stability of mRNA or by regulating the translation. In reaction chains, usually the first committed step is the target of regulation mechanisms. The kinetic properties of consecutive enzymes of the pathway allow a suitable response to changes in the metabolic flux. In branched pathways, different systems exist which coordinate the fluxes to the various end products.

Protein Biosynthesis: The Flexible Manufacturing Metaphor

Stevo Bozinovski, Laboratory of Intelligent Machines and Bioinformation Systems, Electrical Engineering Department Liljana Bozinovska, Laboratory of Neurophysiology, Institute of Physiology, Medical Department Sts Cyril and Methodius University, Skopje

In order to understand molecular genetics, several metaphors have been used, the oldest and most prominent being the linguistic metaphor. It uses basic terms as transcription and translation to describe what is going on during the protein biosynthesis process. Considering the linguistic metaphor, we believe that the processes can be described using the concept of Turing machine. One example is the translation process where the ribosome can be viewed as a two tape Turing machine. Gradually, in the protein biosynthesis process, a terminology of manufacturing has been adopted. We proposed (Bozinovski and Bozinovska, 1987) the metaphor of flexible manufacturing as an appropriate metaphor for describing the protein biosynthesis. We are looking for analogy between the protein biosynthesis and modern concepts of Computer Integrated Manufacturing (CIM) and Flexible Manufacturing Systems (FMS). First we considered the translation-I process. In that process we see ARS-ase as loading station, t-RNA molecules as AGVs, and ribosomes as FMS cells. Viewing that way, we proposed a tree-structured genetic code as the most informative representation of the genetic code computation. Further we developed a conceptual model of the FMS that covers both protein biosynthesis and human made FMSs. It includes multilevel regulatory pathways from event recognition system to the event related protein production, with the corresponding feedbacks. A simulation system has been developed on the basis of

such a concept, which emphasizes analogies between the protein biosynthesis and human made FMS. The system has been modeled as network of communicating agents. The main agents of the system are Polymerase, ARS-ase, and Ribosome, but other regulatory agents are also modeled.

The TRANSFAC-Database

Edgar Wingender, Molecular Bioinformatics of Gene Regulation, GBF, Braunschweig, Germany

The TRANSFAC database contains information about eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles. The TRANSFAC server
http://transfac.gbf.de
provides access to a flat file version of the database which is converted "on the fly" into html format. The contents are arranged in six flat files, the most important ones are SITE and FACTOR. Active hyperlinks allow to navigate between these and the other TRANSFAC tables (GENE, CELL, CLASS, and MATRIX) as well as to eleven external data sources. The MATRIX table compiles positional weight matrices derived from experimentally characterized and aligned transcription factor binding sites. Accompanying programs allowing sequence analysis for potential transcription factor binding sites are PatSearch and MatInspector, the latter being a joint development with the group of T. Werner at the GSF
http://www.gsf.de/biodv

A Demonstration of MetaNet Graph Drawing and Analysis Software

Michael Kohn, Laboratory of Computational Biology and Risk Analysis, National Institute of Environmental Health Sciences, USA

MetaNet is a pair of programs written in C++ for the dynamic definition of a graph model of metabolic or gene expression systems and for the analysis of the resulting topology to identify potential regulatory sites. The graph drawing program provides a palette of tools for creating nodes that represent either the chemical constituents of the pathway (chemnodes) or the elementary reaction steps for their interconversion (relnodes). Chemodes are placed by the user and are automatically joined via relnodes by "click and drag" mouse movements. The

structure of the graph can be edited at any time. Nodes can be inserted, deleted or reconnected at will. Arcs joining pairs of nodes are initially defined as straight lines but can be edited to follow an arbitrary path. Double clicking on a node raises a property panel containing a form for input or editing of the numerical values of the parameters associated with each node. Selecting "Analyze" from the "Run" menu launches the analyzer program. The minimal size cut set of relnodes is highlighted in red. If there is no unique solution, all of the degerate solutions are identified. The program provides a tree view of the nodes for chemical species, reactions, cut set relnodes, and controlling enzymes.

A Control of DNA supercoiling in the complex cell

Jacky Snoep, Coen C. van der Weijden, and Hans V. Westerhoff, Mol Cell Physiology, Free University Amsterdam, Dept. of Molecular Physiology, The Netherlands; Heidi W. Andersen, and Peter R Jensen, Microbiology, Technical University of Denmark

DNA isolated from the prokaryotic cell is usually negatively supercoiled, i.e. the linking number of covalently closed DNA molecules is lower than it would be in the relaxed state. There are at least two enzymes which have the potential to control the level of DNA supercoiling: topoisomerase I, a type I topoisomerase, which relaxes negatively supercoiled DNA, and DNA gyrase, a type II topoisomerase, which introduces negative supercoils in the DNA by coupling the reaction to ATP hydrolysis. The genes encoding topoisomerase I (topA) and DNA gyrase (gyrA and gyrB) are among the genes that respond to changes in the level of DNA supercoiling: the expression of the DNA gyrase is highest when the level of DNA supercoiling is low and the expression of topoisomerase I, is stimulated by high levels of negative supercoiling. This feedback on the level of gene expression may contribute to a homeostatic control of DNA supercoiling. Homeostatically controlled systems have not been widely studied in terms of control analysis. In traditional Metabolic Control Analysis the levels of enzymes are considered to be parameters of the system i.e. fixed. If the control exerted by the topoisomerases on DNA supercoiling is attenuated through genetic feedback loops, then the concentrations of these two enzymes will not remain constant. Hierarchical Control Analysis, is the extension to Metabolic Control Analysis that does accept variations of enzyme concentrations as regulatory mechanisms. The concentration of DNA gyrase was modulated in growing E.coli cells, and the extent DNA gyrase controls the steady state level of DNA supercoiling was determined. Furthermore, using Hierarchical Control Analysis, we show the relationship between

direct metabolic control (with constant enzyme levels) and hierarchical control (which does include regulation through transcription).

Designing metabolism: Alternative connectivities for the pentose phosphate pathway

Jay Mittenthal, University of Illinois, Dept. of Cell and Structural Biology, Urbana, USA; Ao Yuan, Bertrand Clarke, Alexander Scheeline

We present a method for generating alternative biochemical pathways between specified compounds. We systematically generated diverse alternatives to the nonoxidative stage of the pentose phosphate pathway, by first finding pathways between 5-carbon and 6-carbon skeletons. Each solution of the equations for the stoichiometric coefficients of skeleton-changing reactions defines a set of networks. Within each set we selected networks with modules; a module is a coupled set of reactions that occurs more than once in a network. The networks can be classified into at least 53 families in at least 7 superfamilies, according to the number, input-output relations, and internal structure of their modules. We then assigned classes of enzymes to mediate transformations of carbon skeletons and modifications of functional groups. The ensemble of candidate networks was too large to allow complete determination of the optimal network. However, among the networks we studied the real pathway is especially favorable in several respects: It has few steps, uses no reducing or oxidizing compounds, requires only one ATP in one direction of flux, and does not depend on recurrent inputs.

The SWISS-PROT and TrEMBL Protein Sequence Database as a Tool to Model Regulatory and Metabolic Pathways

Rolf Apweiler, European Bioinformatics Institute, Cambridge, UK

SWISS-PROT, established in 1986 and maintained collaboratively, since 1987, by the University of Geneva and the EMBL Data Library (now the EMBL Outstation - The European Bioinformatics Institute (EBI)), is the most widely used protein sequence database since it distinguishes itself from other sequence databases by three essential criteria: MINIMAL REDUNDANCY - Many se-

quence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. In SWISS-PROT we try as much as possible to merge all these data so as to minimise the redundancy of the database. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding entry. INTEGRATION WITH OTHER DATABASES - It is important to provide the users of biomolecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialised data collections. SWISS-PROT is currently cross-referenced with 30 different databases. Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT. ANNOTATION - One of SWISS-PROT's leading concepts from the very beginning was to provide far more than a simple collection of protein sequences, but rather a critical view of what is known or postulated about each of these sequences. In SWISS-PROT each sequence entry consists of the sequence data, the citation information (bibliographical references), the taxonomic data (description of the biological source of the protein), and the annotation which describes the following items: - Function(s) of the protein - Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc. - Domains and sites. E.g. calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc. - Secondary structure - Quaternary structure - Similarities to other proteins - Disease(s) associated with deficiencie(s) in the protein - Sequence conflicts, variants, etc. In SWISS-PROT, annotation is mainly found in the comment lines (CC), in the feature table (FT) and in the keyword lines (KW). We use a controlled vocabulary whenever possible; this approach permits the easy retrieval of specific categories of data from the database. We include as much annotation as possible in SWISS-PROT. To obtain this information we use, in addition to the publications that report new sequence data, review articles to periodically update the annotations of families or groups of proteins. We also make use of external experts, who have been recruited to send us their comments and updates concerning specific groups of proteins. However, due to the increased data flow from genome projects to the sequence databases we face a number of challenges to our way of database annotation. The attachment of biological knowledge abstracted from publications to the sequences is a skilled and labour-intensive task. Maintaining the high quality of sequence and annotation in SWISS-PROT requires careful sequence analysis and detailed annotation of every entry. It is the rate-limiting step in the production of SWISS-PROT. The ever-increasing rate of determination of new sequences requires new approaches if SWISS-PROT is to keep up. While we do not wish to relax the high editorial standards of SWISS-PROT, it is clear that there is a limit to how much we can speed the annotation procedures. On the other hand, it is also vital that we make new sequences available as quickly as possible. To address this concern, we introduced in 1996 TrEMBL (Translation of EMBL nucleotide se-

quence database). TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences (CDS) in the EMBL database, except for CDS already included in SWISS-PROT. SWISS-PROT + TrEMBL represent the most complete and up-to-date protein sequence database with the lowest degree of redundancy and the highest standard of annotation publicly available today. However, to cope with the flood of sequence and functional data new techniques to speed up sequence analysis, information acquisition and data integration into SWISS-PROT + TrEMBL need to be developed. Most of the sequence data nowadays is coming from genome projects and lacks biochemical evidence to provide hard data on the function of the protein. The prediction of functional information from primary sequence information is a comparative problem based on a set of general rules and relationships derived from the current set of known proteins. Modern sensitive database search algorithms find already characterised sequences similar to new sequences and enable us to annotate new sequences by analogy to old sequences. Secondary pattern and profile databases are used to enhance TrEMBL entries by adding information about the potential functions of proteins, metabolic pathways, active sites, cofactors, binding sites, domains, subcellular location, and other annotation. We are automating the similarity and motif searches to accelerate the upgrading of TrEMBL entries to SWISS-PROT standard. The annotation task, whether automated or carried out by database curators, can proceed far more quickly if large groups of related proteins, such as families of sequences sharing a similar motif, can be annotated together. A collaborative environment of so-called "agents" has been implemented which enables the investigation of different possibilities to store, share and deduce biological data. We embedded in this environment software to automate and combine similarity searches, motif searches, special sequence analysis tools, and the parsing of verified information from related biomolecular databases. This serves as a framework for the automation of annotation and takes advantage of a rule-based system to analyse sequences by comparison to the biochemically characterised and well-annotated entries in SWISS-PROT to predict in a standardised way the functional properties of TrEMBL entries. The rule-based system consists of a growing number of rules and hierarchical classifications of the annotation content of SWISS-PROT entries, where all nodes in these hierarchical trees are linked to certain annotation. The rules consider the sequence analysis results to decide which node(s) in the classification tree(s) are sufficiently similar to the query sequence and lead subsequently to the incorporation of the appropriate annotation (linked to the node) in the TrEMBL entry. The incorporated annotation is flagged as annotation based on sequence analysis methods. We only add information based on our automatic analysis to TrEMBL entries, if we are convinced that the computer-generation creates correct annotation in more than 99% of the cases. The tools currently in place enable us to add information about the potential function of the protein, metabolic pathways, active sites, catalytic activity, cofactors, binding sites, domains, subcellular location and other annotation to

more than 20% of all new TrEMBL entries in a highly reliable way. With this annotation concept of SWISS-PROT + TrEMBL we try to combine the strengths of annotation carefully done by human experts with biological knowledge and after consultation of the relevant literature and thorough sequence analysis with the power of automation of sequence analysis and computer-generation of annotation. Since predicted annotation assignments and assignments based on hard experimental evidence are clearly distinguishable, we present in TrEMBL highly reliable although putative functional predictions, without lowering the high editorial standards of the standard SWISS-PROT entries. SWISS-PROT + TrEMBL's comprehensiveness and high degree of integration with other databases, as well as the combination of clearly distinguishable experimental and predicted data in SWISS-PROT + TrEMBL makes this protein sequence database a central tool to model regulatory and metabolic pathways.

A Plant Metabolism Database

Bruno Sobral, Agricultural Genomics, National Center for Genome Resources, Santa Fe, USA

The bioinformatics components of agricultural genomics should enable the exploration of various different hypotheses concerning the relationship between genotype and phenotype. Rather than monolithic data repositories, agricultural genomics needs information systems that enable rather than restrict users. These will be required to move forward to the next levels of genomics: understanding (the fundamental question) proteins expressed by genes, the protein's role in the traits (phenotypes) of interest, and the variations in genes and gene expression patterns in populations. To tackle this tough question requires the integration of various types of data through the creation and public deployment of an agricultural genomics information system. Information systems can be usefully described in terms of nouns (databases, providing storage of data) and verbs (analytical or other methods that do things with the stored data). A dictionary of nouns is not particularly useful without the verbs. Nouns, in this regard, could also be called data types. Some examples of important types of data needed to effectively build an agricultural genomics information system are: phylogenies, protein motifs, genetic maps, physical maps, traits (phenotypes), DNA and protein sequences, and multidimensional gene expression data. Some of the verbs that could act on such data are: enter data (to the system), assemble (a physical map), locate and show (a genetic map), relate (different types of data), compare and contrast (among organisms), and, most importantly, suppose (enabling exploratory queries)! The design and construction of an effective agricultural genomics information system requires the user community (data producers and data

users). Through collaborative projects with the community, it is possible to identify the major requirements of the system: which queries, analyses and outputs? Building such a system requires a focussed, concentrated group of professional software developers, tied into distributed organismal information resources. The system should be open so that others, especially those developing new analytical approaches, can develop tools that work with the system. In addition, biologists need training in accessing and using data. New analytical approaches are sorely needed because we are trying to do extremely sophisticated things; in particular, hooking up many types of data multi-dimensionally. Worse, we are faced with connecting technologies and software deliberately designed to be independent. Knowing that our problems are complex is not a discouraging factor, however: it simply means that we must learn to manage complexity. We must understand the multiple and various parts that make up biological systems, but we must also understand how they fit together to produce viable organisms. Finally, we must not continue to build systems for single users! Biological research is changing. Public information and biological reagent repositories are a decentralizing and democratizing force in research. In the future, this may allow any scientist to have direct and rapid access to information and reagents without needing to build a large-scale operation. Hopefully, if this occurs, biological scientists will be able to once again focus on biological questions and be rewarded for creativity instead of fund-raising. In the 21st century, genome projects will generate large, homogeneous, top-down data sets. Rapid information/reagent access will require repositories serving many users and biologists will need: hardware (high-speed connectivity to internet-accessible information systems), software (fully integrated, exploratory toolkits), and brainware (skills to access and manipulate the information).

Regulatory Pathways in KEGG

Minoru Kanehisa, Institute for Chemical Research, Kyoto University, Japan

Molecular biology has been a discipline dominated by the reductionistic approach, where starting from a specific functional aspect of a biological organism the genes and proteins that are responsible for the function are searched and characterized. In contrast, the genome sequencing projects have made it possible to take an alternative approach, which may be called a synthetic approach, toward functional reconstruction of a biological organism from the complete set of building blocks. While it is unlikely that the reductionistic approach alone can cover the entire aspects of biological functions, the synthetic approach has a potential to provide a complete picture of how the biological system works. KEGG (Kyoto Encyclopedia of Genes and Genomes) is an effort to make links from the gene cat-

alogs generated by the genome sequencing projects to the biochemical pathways that may be considered wiring-diagrams of genes and molecules. Specifically, the objectives of the KEGG project are: (i) to computerize all aspects of cellular functions in terms of the pathway of interacting molecules or genes, (ii) to maintain gene catalogs for all organisms and link each gene product to a pathway component, (iii) to organize a database of all chemical compounds in the cell and link each compound to a pathway component, and (iv) to develop computational technologies for pathway comparison, reconstruction, and analysis. The current knowledge of metabolic pathways, especially on the intermediary metabolism, is already well represented in KEGG. The next question is how to organize divergent sets of regulatory pathways. We are collecting data from published literature on various aspects of cellular functions, such as signal transduction, cell cycle, and developemental pathways. However, the existing literature is the result of the traditional reductionistic approach in molecular biology, which probably represents only a fragmentary portion of actual regulatory pathways in the cell. It is therefore necessary to design new systematic experiments, for example, on the gene expression profiles using the microarray technology. KEGG provides the reference dataset and the computational tools to uncover underlying gene regulatory networks in such experimental data. KEGG is publicly made available as part of the Japanese GenomeNet service
http://www.genome.ad.jp/

The death factors: a combinatorial analysis

Dominique Bergeron, Laboratoire de Retrovirologie Humaine, Departement de microbiologie et immunologie, Universite de Montreal, Canada and Anne Bergeron, Paul Geanta, LACIM, Universite du Quebec a Montreal, Canada

We develop theoretical and computational tools to understand how a small group of proteins can modulate signals to trigger two opposite cellular responses. The key point is in recognizing the basic modular properties of the proteins, and their ability to form molecular clusters whose characteristics can be statistically analyzed. The current project focuses on the death factors. These proteins are known to participate in the first step of a process that can signal a diverse range of activities, including cellular proliferation, or death by apoptosis. Our goal is to understand, qualitatively and quantitatively, how minor variations among this group of proteins can generate opposite effects, even in the presence of similar stimuli. The basic hypothesis can be summed up as: a regulatory protein is characterized by the set of its binding domains; these domains are used to construct clusters of different compositions and properties; cellular response depends on

the characteristics of the population of possible clusters. We developed a virtual laboratory that generates clusters of proteins using combinatorial tools and that provides statistical analysis of clusters among the population generated. This laboratory can be used to study any process involving cluster formation. The experimenter must provide the description of the proteins involved, such as: enumeration of binding domains on each factor, quantity of factors, rules of interactions between domains (affinities). In order to predict cellular response, we simulated cluster formation during the first step of signalization and monitored enzymatic activity among protein clusters. Starting with a medium composed of several copies of each death factors and their relatives, we computed the expected number of clusters exhibiting enzymatic activity for both kinase and protease. We compared levels of kinase and protease in population of clusters generated, respectively, following stimulation of TNFR1, TNFR2 or FAS, three members of the tumor necrosis factor (TNF) family. The use of computational tools can provide guidelines to experiments on the cellular response upon receptor stimulation in different cellular contexts. Although presently limited, our virtual laboratory could be improved to include parameters such as geometrical localization of the protein domains, and the relative affinity of domain interaction based on experimental values.

Multi-Genome Views of Whole Genomes with a focus on E.coli global regulatory proteins

Terry Gaasterland, Argonne National Laboratory, Mathematics and Computer Science Div., Argonne, USA

Genome interpretation is an on-going iterative process in which each successive pass incorporates previously gathered data into a new decision process. In genome annotation, every potential coding region in a genome must be compared with each protein sequence in public curated databases, including all other fully sequenced genomes. Similarity at the sequence level translates into putative function assignments. To reinforce sequence alignment information, DNA patterns, e.g. promoter and terminator sites, can be deduced and associated with coding regions. However, no functional assignment is sure until it has been confirmed through biological experimentation. A system that carries out automated genome analysis must be capable of reasoning about the genomic data in the context of this uncertainty. An important part of such a reasoning process is to reinforce putative and even suspected assignments based on subsequent deductions. Just as important is the visual presentation to users of evidence about decisions made by the systems. The MAGPIE system (joint work with Christoph Sensen, IMB-

NRC, Halifax, Canada) has been designed to meet these requirements. To compare genomes, every coding region in a genome is aligned with every coding region in every other fully sequenced genome. We have devised a system to parse the alignment data into genomic and phylogenetic signatures for every coding region in every genome. Collectively, those signatures provide a phylogenetic overview for an entire organism. If we consider the phylogenetic and genomic signatures for a functionally defined subset of coding regions from multiple genomes (e.g. all gene products involved in energy metabolism or all gene products categorized as global regulatory proteins), we can deduce allowable losses, gains, and alterations of function. As with annotation, visualization of comparative genomic data helps users to gain insights and intuition about the genomes. We have used the MAGPIE system as the data collection engine to gather cross-genome analysis data for 23,971 open reading frames (ORFs) in 10 genomes (Aquifex aeolicus, E.coli, H.infl.,M.genit.,M.pneu., Synechocystis sp, M.janna., M.thermo., S.solfataricus, and S.cerev.). The amino acid sequences from each coding region in each genome were compared via BLAST, FASTA, CLUSTALW, and PHYLIP coordinated via MAGPIE. We used a new suite of programs (joint work with Mark Ragan, IMB-NRC, Halifax, Canada) to generate genomic signatures and derive cross-genome analyses from the collected data. We use the genomic signatures to further define the following concepts: genomically universal proteins (proteins that have a counterpart in every fully sequenced genome); proteins characteristic of phylogenetic subsets, including proteins characteristic of bacteria (proteins that have a counterpart in every fully sequenced bacterial genome and NO detectable counterpart in any other genome), of archaea, of prokaryotes, of both bacteria and eukaryote, of both archae and eukaryote. Highlights of the results include the following: - We deal explicitly with 310 mitochondrial biogenesis ORFs in the - yeast nuclear genome. Their profile is more bacterial than that of - non-mitochondrial biogenesis yeast ORFs. Only 15% of yeast ORFs shared - with bacteria but not archaea are involved in mitochondrial biogenesis. - Likewise only 12% of yeast ORFs shared with bacteria and archaea are involved in mitochondrial biogenesis. We profile the ORFs shared universally between each pair of phylogenetic domains but not the third. A number of ORFs are shared between bacteria and yeast and between archae and yeast at each level. However, only 1 ORF is characteristic of prokaryotes (present in all bacteria and archae but absent in yeast) at level 1. We notice that functions related to replication and transcription are indeed over-represented among ORFs that have counterparts only in both archae and eukaryotes; however, many other unrelated cell-processes functions are also present. We argue for monophyly of archae based on the fact that matching Sulfolobus is a better predictor of matching another archae than is matching a bacterium a predictor of matching an archae. We profile the difference between two methanogenic genomes. ORFs that have been lost in one methanogen but not the other almost all have counterparts in bacteria. For these genomes, specialization has occured by losing prokaryotic ORFs. We make several observations about the functional

categories represented by the proteins that occur in all genomes. First, they are almost exclusively functions that are considered 'necessary' to an autotrophic organism's survival. Second, they do not encompass all such functions. Thus, they generally comprise a necessary but not sufficient set of protein functions. This study lays a foundation for systematic comparison of multiple whole genomes. It also demonstrates how to include partially sequenced genomes in 'cross-genome' profiles. With 10 microbial genomes, our system has confirmed and qualified common observations from the literature. It has also led to new insights into genomic evolution at a protein level. Future work will include the next cohort of fully sequenced genomes, which include pathogens, non-archaeal extremophiles, and a putatively ancient bacteria. It will also include the available predicted protein coding regions from C. elegans and human.

Regulatory Predictions in the Complete E.coli Genome

Julio Collado-Vides

Julio Collado-Vides, Heladia Salgado, and Araceli M. Huerta, Centro de Investigation sobre Fijacion de Nitrogeno. Universidad Nacional Autonoma de Mexico, Cuernavaca, Mexico

The complete genome sequence of E.coli has been recently completed (Blattner et al., Science (1997) 277: 1453-1462). The work here presented summarizes the analysis and predictions of operon organization, and regulatory signals such as promoters and binding sites for proteins regulting the initiation of transcription. This work was done in collaboration with the laboratory of Fred Blattner. This analysis is based on RegulonDB, a database of regulation of transcription initiation, as well as operon organization in E.coli that has been built in our laboratory. RegulonDB is a relational database available on the web, currently with around 300 known operons, roughly a similar number of promoters, and around 500 regulatory interactions. This can be found in:
http://www.cifn.unam.mx/Computational_Biology/regulondb/
Based on a large body of known promoters and sites for the binding of regulatory proteins, we performed a global analysis of regulatory features of the E.coli genome in collaboration with the laboratory of Dr. Fred Blattner. The main goal of this work is to make use of the incomplete knowledge of gene regulation in order to develop algorithms that can make reasonable predictions on the complete genome. The distribution of known promoters show that they can be located up to 200 bp upstream from the beginning of the gene. Furthermore, it is known that promoters can vary considerably in their strength. These properties were taken

into account in an algorithm we developed to find promoter candidates within upstream regions of plausible operon regions in E.coli. We initially searched for potential promoter sites using the weight matrices for the conserved -35 and -10 regions, and in a second phase the different candidates within a given regulatory region were filtered based on a comparison or competition of the different candidates in a given region. The distribution of around 400 regulatory sites for 56 different regulatory proteins show that the majority of them occur within 250 bp upstream from the point of initiation of transcription. Therefore, we looked for potential sites for these different proteins in a region 450bp upstream of the beginning of genes within operons. These were searched by means of weight matrices constructed with at least four experimentally characterized sites for a given protein, as well as with a subsequent string-match filter that limited the number of differences of predicted vs known sites. The algorithm to predict operons was based on two observations: The fact that the distances in-between genes that belong to an operon follow a distribution with a peak of around 70bp, as opposed to the distribution of distances in-between genes of different operons that is much flat including larger distances. The second observation is that genes within an operon tend to belong to the same physiological class, as classified by Monica Riley. Finally, we compared the relative consistency of these independent predictions. 16% of operon regions contain a binding site for a regulatory protein as opposed to only 10% of non-coding regions internal to operons. This low number of predicted regulation is due to the limited number of binding sites available for different proteins. We expect to find between 250 to 350 regulatory proteins in the complete genome, whereas we currently only have site information for around 50 regulatory proteins, which corresponds to 1/7, a number consistent with the 16% regulated operons found. We are aware that these different predictions can be improved, and therefore emphasize that the predictions should be taken with caution.

Integration of gene expression with metabolism in kinetic simulations

Pedro Mendes, Institute of Biological Sciences, University of Wales, United Kingdom

Computer simulation of kinetic models has an important role in the biochemical sciences. It serves to check the consistency of our theories with observed behaviour, it allows one to ask "what-if" questions that can reveal non-intuitive properties of the system, it can be used to find estimates for kinetic parameters and it is an educational tool. Although biochemical kinetic simulations have been

performed since the early days of analog computers, these have focused mainly on metabolism. Only a small number of these simulations have focused on the kinetics of gene expression and even a smaller number integrate gene expression with metabolism. The purpose of this paper is to discuss the issues involved with the integration of gene expression and metabolism in kinetic simulations. This integration is becoming more important as the complete genome sequencing projects are coming to an end and experimentally the focus will change to the analysis of the kinetics of these systems. Kinetic modeling of biochemical systems is based on a quantitative description of the rates of the various processes (enzymatic reactions but also other steps like transport). This description is based on a kinetic function for each step. Each kinetic function is characterised by a number of parameters whose values need to be determined in order to solve the equations - the simulation stage. One problem with this type of modeling is that the more detailed we want to make it the larger it will be the number of parameters that will have to determine. Therefore there is a need to create high level representations of the systems such as to minimise the number of parameters. In particular for the metabolic part this passes by describing each enzyme catalysed reaction as a unit or even to lump several of these into one single unit (step). For gene expression I argue that the ideal representation would be of the transcription of each gene to be represented as one single step and also the translation of each mRNA as one single step. This means that we must derive appropriate kinetic functions for each of these steps. So far, most kinetic models of gene expression use kinetic functions for transcription and translation that are not satisfactory as they do not represent the effects of saturation and consider the supply of building blocks as unlimited. Additionally to the problem of the kinetic functions, modeling of gene expression and metabolism requires that we estimate the parameters of those functions. This will require time-resolved measurements of the metabolic and genetic components after appropriate perturbations. Traditionally such experiments are very difficult to carry out, especially in vivo. But recent technological developments are making such experiments possible. Nucleotide array chips are able to measure the concentration of a large number of mRNA molecules and liquid- chromatography mass-spectrometry techniques can do the same for proteins. Measurement of all (or a large number of) small molecular weight metabolites are rather more difficult but developments in spectroscopy and chemometrics are also making this task easier. The existence of these technologies per se is not enough, we have to use them in a way in which they provide the right kind of data from which parameters can be estimated. This means following the time course after a certain perturbation has been applied to the system. Software for simulating such systems is readily available (for example my own software Gepasi) so in principle the limitation is in the data aquisition phase. In conclusion, the combination of gene expression and metabolism in kinetic models is essential as the two aspects of cellular dynamics are intimately coupled and only in special circumstances is it valid to ignore one or the other. This would be cases

in which, for example, gene expression is significantly slower than metabolism so that one can study how the system evolves in the short term without considering enzyme synthesis and breakdown. Some surprises are expected to be revealed when such combined modeling of gene expression and metabolism is carried out. For example, there is evidence from simple models, that the dynamic stability is inversely related with the stability of the mRNA molecules. This means that the faster the turnover of mRNA is the more stable is the system as a whole (i.e. it stabilises quickly). On the other hand, if mRNA is very stable the system tends to oscillate after perturbations, sometimes never achieving a steady state, or taking very long to do so. Only by doing quantitative kinetic simulations can we begin to understand the interplay of gene expression and metabolism and how control is distributed in these systems.
http://www.enzyme.demon.co.uk/pedro.html
http://gepasi.dbs.aber.ac.uk/softw/gepasi.html
http://gepasi.dbs.aber.ac.uk/softw/gepasi.html

The regulatory network that controls flowering in Arabidopsis

Luis Mendoza, Lab. Genetica Molecular y Evolucion, Instituto de Ecologia, UNAM, Mexico

In this seminar a genetic regulatory network that controls flower morphogenesis in Arabidopsis is presented. The model takes into account the transcriptional regulatory relationships of eleven genes that intervene in different aspects of flower development. Extracting data from the literature regarding mRNA expression, it was possible not only to establish the topology of the network but also to the relative strengths of their interactions. With the use of such data, a dynamic system made of difference equations was constructed. Since there is not quantitative data for the expression of those genes, the model uses only binary elements. Additionally, it was necessary to implement a biologically-based updating methodology christened semi-synchronic. The NET model reaches six attractors; four of them corresponding to experimentally observed patterns of gene expression found in the floral organs of Arabidopsis (sepals, petals, stamens and carpels). The fifth state corresponds to a non-flowering stage, and the sixth attractor found in the model never occurs in the wild type plants. Also, it was presented a preliminary analysis of the model using the loop efficiency methodology as developed by the group of R. Thomas. Those results shows that it is possible to obtain five steady states in the system: i) a saddle point between the states of flowering and non-flowering state, ii) a saddle point between the "A" state present in sepals

and petals, and the "C" state present in stamens and carpels, iii) a saddle point between the "B" state present in petals and stamens, and a non-"B" which is found in sepals and carpels, iv) a saddle point between high and low levels of "A" activity, and finally v) a focus in a state intermediary between a flowering and a non-flowering state. Taken all together, the results indicate that the topology of the NET model is sufficient to explain the actual architecture of the Arabidopsis flowers.

Cell Signaling Networks Database

Takako Takai-Igarashi, Division of Chem-Bio Informatics, National Institute of Health Sciences, Tokyo, Japan

We develop a database for cell signaling networks in human cells. The final goal of this project is to make a computational model for biological phenomena such as development, differentiation, carcinogenesis, and aging. Cell Signaling Network Database (CSNDB) bases on an object-oriented database management system, ACEDB. We represent signaling networks as diagrams that are produced automatically by the system. We prepared pre-filtering system for diagram production; a required set of signaling networks is selected according to a user's requests. We use a rule-based production system, CLIPS, for the filtering system. In living cells, as cascades have cross-talk and feedback interactions, the whole networks are highly complex and flexible. CSNDB is a new tool to stock diverse and complex cell signaling data and to provide them as dynamically constructed cascades through the user-friendly interface, using combination of object-oriented and rule-based production techniques. We consider that CSNDB will be useful for estimating biological effects caused by variousextracellular stimuli. Igarashi, T. and Kaminuma, T. Development of Cell Signaling Networks Database, Pacific Symposium on Biocomputing '97, pp.187-197, (1997), World Scientific.

Exploring biochemical models with Gepasi, a kinetics simulator

Pedro Mendes, Institute of Biological Sciences, University of Wales, Aberystwyth, UK

Kinetic models of biochemical pathways and gene expression systems are becoming very important (Mendes 1998, these proceedings) now that full genome

sequences are becoming available at a fast pace. Kinetic models of biochemical pathways are in sufficiently complex and there is a strict requirement of software for their simulation as in general these models do not have a known analytical solution. Here I demonstrate the biochemical kinetics simulator Gepasi (Mendes, 1997, Trends Biochem. Sci. 22, 361-363), a software package for the simulation, optimisation and analysis of biochemical kinetics. This program is freely available on the Internet at
http://gepasi.dbs.aber.ac.uk/softw/Gepasi.html
http://gepasi.dbs.aber.ac.uk/softw/Gepasi.html
Gepasi is a Microsoft Windows program, with a user-friendly front-end based around a control-panel paradigm. This program was written with the explicitly intention of being easy to use but at the same time to be powerful and above all to follow the most correct numerical algorithms. With Gepasi a non-specialist in computing, such as the average biochemist, can define and simulate a biochemical model. Pathways are entered by typing the component chemical reactions in the usual chemical syntax, then kinetic types are selected for each reaction from a list of predefined ones. If needed, further kinetic types can be added by the user in the form of a rate equation. After setting the values of all the kinetic parameters and initial concentrations, the program is ready to simulate the pathway: available are a steady-state solution and a time course of reaction progress. The pathway is also analysed in terms of its structural properties (mass conservation relations and elementary modes, see Schuster 1998, these proceedings) and steady state solutions are further characterised in terms of metabolic control analysis, and stability analysis. All this is available automatically for each simulation. Gepasi is tightly coupled with the free plotting program gnuplot to display results in publication quality and has a help file containing introductions to the various aspects of metabolic kinetics and simulation, including full bibliographic references to the relevant literature. This makes the program useful for both research and education. The power of computer simulation is however not limited to running simple simulations. Computers excel in repetitive tasks and Gepasi takes advantage of this. The program can be instructed to carry out a series of simulations, where some parameters are changed from simulation to simulation ("scans"). This effectively allows one to map the behaviour of the model with respect to the parameters varied. Gepasi is not limited in the number of parameters selected for scans but obviously this is limited by the combinatorial nature of this procedure. For example if one wanted to map the behaviour of a pathway for 10 parameters and one would 5 values for each one, this would effectively require 5^10 simulations. To get around this problem the program is also capable of carrying out optimisation. The user selects any variable of the model (or perhaps some complex function of variables) to be either maximised or minimised and then Gepasi can use a series of numerical nonlinear optimisation methods to solve the problem. If the user is able to state the problem as an optimisation problem, then the software will be able to search for a solution. Gepasi

has available many optimisation methods, from steepest descent to genetic algorithms and simulated annealing, such that several can be used to attempt to solve the optimisation problem. This is important because it is well known in numerical analysis that no single optimisation method is the best for all problems. The optimisation routines can also be used for parameter estimation from experimental data, the program then attempts to minimise the sum of squares of residuals between experimental and simulated points. This is extremely useful for model building.

Molecular Database Integration: Analysis of Metabolic Network Control

Andreas Freier, Michael Höding, Ralf Hofestädt, Matthias Lange, Uwe Scholz, Otto-von-Guericke-University Magdeburg, Institute for Technical and Business Information Systems, Bioinformatics and Medical Informatics, Magdeburg, Germany

The development of the Magdeburger Molecular Information System (MMIS) is the goal of our project. The architecture of our prototype allows the access onto two different molecular database systems which allow the analysis of metabolic pathways. The access to the molecular knowledge (genes, proteins, and pathways) is realized by using the information system KEGG which allows the access to every known metabolic pathway including the related gene and proteins. Information about the gene regulation is available via the TRANSFAC database system. Our WWW-Server connects both molecular database systems. This integration tool represents the kernel of our MMIS. Furthermore, our system offers the simulation tool Metabolika for the analysis of metabolic pathways. Metabolika allows the interactive simulation of biochemical networks. Therefore, molecular knowledge can be transfered into analytical metabolic rules - the language of Metabolika. Based on that information transfer, the simulation of complex metabolic networks is available. The configuration of Metabolika is represented by the actual metabolite concentrations of the virtual biochemical reaction space. Metabolika allows the calculation of (all) possible configurations (derivation tree) based on the selected metabolic knowledge (biochemical scenario) and the start configuration. The visualization tool and the Graphical User Interface (GUI) realize the interactive analysis of the corresponding derivation tree. The idea of our MMIS is to present a virtual laboratory for the analysis of molecular processes. Therefore, we integrated different database systems which represent molecular and medical knowledge. The graphical user interface gives the user access to a compact local information system. The access to the molecular knowledge

will be realized by the direct access to the heterogeneous database systems. In case of modeling and simulation of metabolic processes the specific biochemical knowledge will be identified by using these database systems. In the next step this knowledge will be transfered automatically into the language of analytical metabolic rules, the language of Metabolika. The simulation of this biochemical reactions will be produced by Metabolika. For the visualization and statistical analysis of the derivation tree tools are available.

Indexing and Retrieval of Complex Data Sets or Is it a good idea to store metabolic data in an ooDB?

Thomas Mück, Institut für Angewandte Informatik und Informationssysteme, Universität Wien, Austria

Efficient storage and processing of metabolic information (e.g., pathways) relies to a large extent on the application of graph theoretical concepts, data structures and algorithms. Therefore object-oriented and object-relational database management systems could provide a technically sound storage and retrieval layer for metabolic information systems, simulation packages and expert systems. In particular the advanced modeling features of such systems like object identity, type hierarchies, association paths and the general principle of data encapsulation yield significant enhancements with respect to expressive power. Thus complex objects like hypergraph representations of metabolic pathways can be handled in a convenient and, above all, efficient way. Additionally the object-oriented database paradigm provides excellent means for state-of-the-art semantic database integration in the context of so called federated databases. This could be helpful for solving several integration problems caused by the technical as well as semantic heterogeneity of the different metabolic information repositories currently existing.

Databases integration and automatic knowledge acquisition on regulatory regions of eukaryotic genome

N. A. Kolchanov, M. P. Ponomarenko, A. E. Kel, Yu. V. Kondrakhin, A. S. Frolov, F. A. Kolpakov, O. V. Kel, E. A. Ananko,

E.V. Ignatieva, O. A. Podkolodnaya, I. L. Stepanenko, T. I. Merkulova, V. N. Babenko, D.G Vorobiev, S.V. Lavryushev, Yu. V. Ponomarenko, A. V. Kochetov, G.B Kolesov Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia; L. Milanesi, Istituto Di Tecnologie Biomediche Avanzate, Consiglio Nazionale Della Ricerche, Milano, Italy; V. V. Solovyev, The Sanger Centre Hinxton, Cambridge, UK N. L. Podkolodny, Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia; E. Wingender, T.Heinemeyer, Gesellschaft für Biotechnologische Forschung mbH, Braunschweig, Germany

Various experimental data on eukaryotic gene expression are being rapidly accumulated. The number of databases on gene expression and a variety of software for the analysis of these data grow fast. But these resources are dispersed and weakly integrated. That's why among the main problems is the creation of the unified Internet-accessible media to provide the maximal integration of biocomputing resources on gene expression regulation and to admit the effective user navigation in the integrated resources. The system GeneExpress is developed toobtain specific knowledge on the DNA regulatory regions stored in the databases. It comprises 5 basic units in the current version. (1) Transcription Regulation unit contains the TRRD - database on transcription regulatory regions of eukaryotic genes; (2) Transcription Factor Binding Site Recognition unit contains programs for sites analysis and recognition; (3) ACTIVITY is the module for sites activity prediction by their nucleotide sequences; (4) mRNA Translation unit is designed for analysis of translational properties of mRNAs; (5) GeneNet is the database on gene networks and signal transduction pathways. The database integration into the GeneExpress is based on the Network Browser of Sequence Retrieval System (SRS). GeneExpress is available via Internet
http://wwwmgs.bionet.nsc.ru/Systems/GenExpress

Integration of protein data : ProDom, XDOM and genome projects

Daniel Kahn, Jérome Gouzy, Laboratoire de Biologie Moléculaire des Relations Plantes-Microoorganismes, I.N.R.A./C.N.R.S., Castanet-Tolosan Cedex, France Florence Corpet, Laboratoire de Génétique Cellulaire, I.N.R.A., Castanet-Tolosan Cedex, France

The combinatorial nature of proteins makes it necessary to decompose every protein sequence into domains before clustering on the basis of homology. This is done in the ProDom database of domain families, which provides also a number of tools for protein domain analysis : (1) graphical representation of protein domain arrangements ; (2) domain homology search utility; (3) links to primary databases, SWISS-PROT,PROSITE and PDB ; (4) multiple alignment utility ; (5) links to 3-D modeling with SWISS-MODEL, where appropriate.
http://www.toulouse.inra.fr/prodom.html
We propose to systematically organise our knowledge on proteins (sequences, structures, biochemical data) around the concept of protein domain families, because domains appear to be the fundamental building blocks of proteins. We have applied domain analysis to all known or predicted proteins from 13 completed microbial genomes. 112 domain families were found in all 13 species, and appear therefore as 'universal'. Many more domain families were found conserved in all archaea and in yeast, than between all bacteria and archaea, or between all bacteria and yeast. From this perspective archaea and yeast would appear more closely related to each other than to bacteria. Finally, statictics of multi-domain proteins indicate 2 classes of proteins in B. subtilis, with one class of highly multi-domain proteins. These include families of polyketide synthase homologues and of peptide synthetase homologues.

GeneNet: a Database for Gene Networks and its Automated Visualization through the Internet

Fedor A. Kolpakov, Elena A. Ananko, Grigory B. Kolesov and Nikolay A. Kolchanov Laboratory of Theoretical Molecular Genetics, Institute of Cytology and Genetics, Novosibirsk, Russia

The gene network concept. The physiological functions of organisms are accomplished through the coordinated regulation of the expression of a large number of genes. Hence, there exist complex networks: the gene ensembles functioning in a coordinated manner to provide vital functions, the fine regulation of physiological processes, and the responses to external stimuli. The functional elements of a gene network are: (1) a gene ensemble interacting when certain biological functions are performed; (2) proteins encoded by these genes; (3) signal transduction pathways providing gene activation in response to an external stimulus; (4) a set of positive and negative feedbacks stabilizing the parameters of the gene network (autoregulation) or providing a transition to a new functional state; and (5) external signals, hormones, and metabolites that trigger the gene network or correct its operation in response to the changes in physiological parameters.

Databases on gene networks. Experimental data on the features of gene function have been rapidly accumulated during the last ten years resulting in development of several specialized databases. The major of them are
(1) KEGG, the Kyoto Encyclopedia of Genes and Genomes
http://www.genome.ad.jp/kegg/kegg.html
(2) BRITE, the Biomolecular Reaction Pathways for Information Transfer and Expression
http://www.genome.ad.jp/brite/brite.html
(3) CSNDB, the Cell Signaling Networks Database
http://geo.nihs.go.jp/csndb.html
(4) SPAD, the Signaling Pathway Database
http://kintaro.grt.kyushu-u.ac.jp/eny-doc/spad.html
(5) GeNet, the Gene Networks Data Base
http://www.iephb.ru/ spirov/genet00.html
All the above databases contain manually drawn interactive diagrams of the signal transduction pathways and gene networks described. Automated construction of diagrams from the formalized information appears to be a promising direction. EcoCyc was the first convincing demonstration of the efficiency of automated diagram generation for metabolic pathways; however, the gene network databases available are not provided with such tools. The GeneNet database. We have developed an object-oriented database GeneNet, compiling the information on the gene networks of antiviral response and erythropoiesis regulation. The information contained in the databases IIG-TRRD and ESG-TRRD, respectively, was used for their formalized description. A chemical formalism was employed as a basis for describing the events occurring in the gene network. Thus, any event is described as follows:

$$A_1 + \ldots + A_N \xrightarrow{B_1+\ldots+B_M} C_1 + \ldots + C_K$$

where A is the entities entering into reaction; B, the entities affecting the course of reaction; and C, the products of reaction. Basing on this model, we consider two types of relationships between entities: (1) reaction (indicated by double arrow), that is, formation of a new entity or acquisition of a new property by the entity, and (2) regulatory event (single arrow), that is, the effect of an entity on certain reaction. The following entities participating in the events are considered: (1) cells (tissues, organs); (2) genes; (3) proteins and protein complexes; and (4) nonprotein regulatory substances and metabolic products. In the GeneNet database, each type of gene network components is described in a separate table: (1) CELL, containing the information on cells, tissues, and organs; (2) PROTEIN, on proteins and protein complexes; (3) GENE, on genes and their regulation patterns; (4) SUBSTANCE, on nonprotein regulatory substances and other metabolic products; (5) STATE, on physiological functions and the state of the organism; (6) RELATION, on relationships between the gene network components; (7) SCHEME contains the formalized description of the gene network

graphs that including: the list of gene network components and the relationships between them, and instructions for optimal their optimal arrangement in the diagram; and (8) LITER, containing the references to the original papers. The GeneNet has references to the databases EMBL, SWISS-PROT, TRRD, TRANSFAC, EPD, and MEDLINE. Automated generation of gene network diagrams. The GeneNet database is designed to allow the automated construction of the gene network diagrams basing on their formalized description. A specialized Java program was created for this aim. It is accessible via the Internet at
http://wwwmgs.bionet.nsc.ru/systems/MGL/GeneNet/
A diagram of the gene network is a graph with nodes corresponding to entities and arrows representing relations between the gene network components. Each component of the gene network has its own image reflecting the peculiarity of the object. Information about the structure of the gene network graph is contained in the SCHEME table. A compartmentalization is characteristic of all biochemical reactions in the organism. Hence, the gene network is described at three hierarchical levels: (1) organism level; (2) single cell level; and (3) single gene level. This allows us both to take into account that the components of a gene network are scattered through different organs, tissues, cells, and cellular compartments and to describe different regulation levels of the gene network. The system of filters. The table SCHEME contains a consolidated description of the gene network based on experimental data obtained in different species, cell types, and under different conditions. The default diagram is built basing on the entire table. However, the system of filters allows the user to select for graphical representation only those objects and their interrelationships that have been described experimentally in a specified species, cell type, and/or in response to a certain stimulus. The work was supported by the Russian Foundation for Basic Research (grants Nos. 96-04-50006, 97-07-090309, 97-04-49740, and 98-04-49479), Russian National Program on Human Genome, and Russian Committee on Science and Technology. The authors are grateful to O.A. Podkolodnaya for kindly providing the information on regulation of erythropoiesis.

Integrated Analysis of Metabolic Pathways, Sequence Evolution and Genome Organization

Hiroyuki Ogata, Wataru Fujibuchi, Susumu Goto, and Minoru Kanehisa, Institute for Chemical Research, Kyoto University, Kyoto, Japan

We introduce and discuss a new computational method for automatic extraction of functional units by making use of genomic data and biochemical pathway data

in KEGG
http://www.genome.ad.jp/kegg/
In order to obtain functional clues of a gene in a complete genome, it is customary to perform similarity search of the gene against database sequences. However, the search alone leaves, at least, one third to one half of genes in a genome ashypothetical. To overcome this situation, we have been focusing on functional units, sets of genes or gene products, which make basal building blocks of cellular functions
http://www.genome.ad.jp/dbget-bin/get_htext?Ortholog
Although the homology search against the functional units represented in the ortholog tables and the following examinations of completeness of the units are obviously useful for gene function prediction, collection and compilation of the units are time consuming and to be automated. To this end we have recently developed a method to automatically extract the functional units. The method is based on a concept of graph, where a node is a gene or a gene product and an edge is a link or a relationship between genes or gene products. By comparing two biological networks represented as undirected graphs, it detects local clusters of corresponding nodes that represent links of genes and/or gene products. Different kinds of links make different networks or graphs. For example, a genome is seen as a set of genes that are one-dimensionally linked, so it is represented by a linear graph. A set of interacting gene products in a biochemical pathway is another type of graph. The utility of the method is demonstrated in the following two comparisons . If the method is used for a comparison of a genome with a set of known biochemical pathways, it extracts gene clusters that play their roles at close positions on the biochemical pathways. An analysis on metabolic pathways showed that most of the gene clusters in E. coli thus detected corresponded to enzyme operons. By comparing each known genome against metabolic pathways we observed many common gene clusters that were conserved among multiple organisms, as well as many organism-specific gene clusters. This type of analysis would be useful for reconstructing and characterizing functional units of biochemical pathways. If the method is used for a comparison of a genome versus another genome with correspondence information of sequence similarity, it extracts pairs of orthologous gene clusters. While it is well known that global arrangement of orthologous gene clusters on the genome can be highly shuffled between two distantly-related bacterial lineage, we also observe gene shuffling events by translocations, inversions, insertions and deletions within some of the orthologous gene clusters. Practically, the extraction of orthologous gene clusters gives important clues for identification of orthologs that have been missed by simple homology searches. In both examples, most of the genes in each cluster appear to have functional relation with each other. Extracted clusters are merged and represented into ortholog tables, which would be useful for function prediction of genes. We believe comparative analysis of networks of biological entities at this level of abstraction would be fruitful for development of practical tools

such as for automatic annotation of gene functions.

The Metabolic Diseases Database

Manuela Prüß, Institute for Technical and Business Information Systems, University Magdeburg, Germany

We present a database which combines medical and molecular knowledge of a special type of metabolic diseases for the computer supported detection of inborn errors. In this database, called Metabolic Diseases Database, or MDDB, we collect both the medical and the biochemical and genetic data on hyperammonemias. Hyperammonemias are diseases which are characterized by disturbances in the synthesis of urea, amino acids or other organic acids. This are inborn errors, basing on different gene defects, which lead to deficient enzymes, so that special biochemical reactions can not be catalyzed and the regarded metabolic pathways are blocked. We want to present not only the medical data, like general information on the disease, symptoms, laboratory findings and therapy, but also the molecular data. This include information on genes, gene variants and their description, and gene regulation elements, as far as there are data available. The data on enzymes include also general information like EC number, synonyms and structure of the enzyme and information on the catalyzed biochemical reaction, with structural formulas of the substances. The pathways regarded in case of given diseases are also shown by a diagram. The database is a relational one. It was built according to the entity relationship schema, and it runs on a windows PC.

The GeNet Database

John Reinitz, Mt. Sinai Medical School, Brookdale Center for Molecular Biology, New York, USA
The GeNet Team: Dr. Maria G. Samsonova - Group Leader Dr. Alexander V. Spirov - Data Base Curator Vasiliy N. Serov - Data Base Administrator, Programmer Katherina G. Savostyanova - Researcher Svetlana Yu. Surkova - Researcher Olga V Kirillova - Researcher Institute for High Performance Computing and Data Bases, St.Petersburg, Russia

GeNet is a hypertext database. The concept of genetic networks forms a basis for information structuring in this database-each of the thus far characterized genes is

considered as a node of a genetic network, while the links between nodes represent the interactions of genes or their products. The are two parts in GeNet. The EmbryoNet part holds information on genetic networks in sea urchin, Drosophila and vertebrates and contains 6 types of data: genetic network maps, gene entries, gene sequence entries, regulatory region entries, bibliographies and regulatory interactions. The NetModel part of GeNet holds the models of genetic networks. Two entry points allow the user to browse and to search the database. The Java applet NetWork enables a user to construct interactively the genetic network of interest, as well as to visualize and to evaluate the genetic network dynamics in framework of Boolean network model. With NetWork it is possible to model the effects of the mutations in the network, as well as to reveal gene interactions compensating for these mutations.

Examples on post-transkriptional reguation and metabolic pathways

Thomas Dandekar, European Molecular Biology Laboratory, Heidelberg, Germany

Metabolic pathways are not only interconnected to other pathways in the cell, additional perspectives to analyze their regulation are metabolic pathway alignment and regulatory steps mediated by RNA. Metabolic pathway alignment reveals differences in substrate flux, conversion and regulation in different species (example shown: glycolysis). After an introduction to regulatory elements in mRNA we next discuss different examples for post-transcriptional regulation in the citric acid cycle by iron-responsive elements. Combining these and other approaches (e.g. differential genome analysis) will allow us to assemble a more complete picture of regulatory and metabolic networks.

Simulation Experiments on the Role of Spatial Arrangement in Enzymatic Networks

Klaus-Peter Zauner and Michael Conrad, Department of Computer Science, Wayne State University, Detroit, USA

Biochemical networks depend on an intricate interplay of conformation, kinetic, structural, and (molecular) dynamic factors. We have developed a simulation system that abstracts this interplay. The simulator provides a theoretical labora-

tory for investigating the role of dynamically changing structures in biomolecular information processing and control. Macromolecules and various small molecules (and ions) are represented in a 3D-simulation space. The macromolecules can act catalytically on the small molecules in their local environment. They may also interact through attractive or repulsive forces with other macromolecules in their vicinity. The force (or dynamic) interactions and the catalytic properties of each of the macromolecules is dependent on its conformational state. Macromolecules are represented by dodecahedra, each of whose twelve faces is a finite automaton. The states of this compound finite automaton represent the conformational states of the macromolecule. The state transitions depend on the local milieu and on the state of neighboring macromolecules. The whole system forms a loop: conformation controls binding and catalytic interactions that influence supramolecular structure and chemical milieu. Structure and chemical milieu in turn influence conformation. The simulator was used to study a network composed of five types of enzymes with two competitive paths. The enzymes were represented by 3300 localized dodecahedra placed in a $1\mu m^3$ simulation space. The kinetic parameters and the number of enzymes of each type were chosen to be symmetrical for both of the competitive paths. The flux through these paths can be modulated by changing the relative spatial distribution of the enzymes in the simulation space. The results, in the case under consideration, demonstrate that the steady state concentrations of the milieu components are significantly affected by the arrangement of the macromolecules.

E-CELL vs StochSim: System-wide and molecularly-detailed approaches to simulation of cellular processes

Tom Shimizu, Trinity College, Cambridge, UK

E-CELL and StochSim are both generic simulators of cellular processes, and use quantitative models to simulate the biochemical reactions. They also share the common aim of reproducing and making predictions about observable cell behaviour under a given set of conditions. However, the approaches that the two systems employ strongly contrast from eachother, and can be viewed as complementary. E-CELL attempts to study system-wide properties of the cellular system from a macroscopic perspective by simulating all of the reactions which occur within the cell, simultaneously. A novel object-oriented classification hierarchy is provided for fascilitating the modeling of large biochemical systems with diverse molecular components. E-CELL's present limitations, much of which we expect to overcome through ongoing work, include the limitied accuracy of nu-

merical integration, limitations imposed by deterministic, mass-action kinetics assumptions, as well as the problems of handling stiffness. Simultaneously to development of the software, an extensive effort is being made within our group to construct diverse models of various scales (ranging from single metabolic pathways, signalling pathways, gene regulation networks, to entire cells) for simulation in E-CELL. StochSim's molecularly-detailed approach aims to overcome one of the most significant limitations of E-CELL and most other biochemical simulators which assume that reactions take place in even-mixture solutions. In reality, the cytosol of a living cell is packed with proteins and other large molecules. By modeling the stochastic behaviour of individual molecules, rather than the deterministic concentration dynamics of molecular species, StochSim simulations can provide realistic results even in very small volume compartments which arise in cellular environments due to macromolecular crowding. The major limitation of the StochSim approach is that the computational load of simulating at such a level of detail imposes limits to the size/complexity of the system it can be applied to. StochSim has already proven to be useful in a detailed analysis of the signalling pathway for bacterial chemotaxis, and can easily be applied to the analysis of many other systems. I plan to integrate these approaches in future work in order to overcome the limitations of each simulator. Developed by Takahashi K., Shimizu T., Hashimoto K. and Tomita M. et al. at the Laboratory for Bioinformatics, Keio University, Japan. More information is available at:
http://bio.mag.keio.ac.jp
Developed by Morton-Firth, C. at the Department of Zoology, University of Cambridge, UK. More information is available at:
http://www.zoo.cam.ac.uk/zoostaff/morton
E-CELL: Software Environment for Whole Cell Simulation. Tomita M., Hashimoto K., Takahashi T., Shimizu T.S., Matsuzaki Y., Miyoshi F., Saito K., Tanida S., Yugi K., Venter J.C. and Hutchison III C.A. 1998. Bioinformatics (accepted).
Predicting Temporal Fluctuations in an Intracellular Signalling Pathway. Morton-Firth C.J., Bray D. 1998. Journal of Theoretical Biology 192(1):117-128.

Towards a Virtual-Lab-System for Metabolic Engineering: Development of Biochemical Engineering System Analyzing Tool-Kit (BEST-KIT)

Masahiro Okamoto, Department of Biochemical Engineering and Science, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Fukuoka, Japan

True understanding of complexity in bioprocess reaction network requires new

approaches to both mathematical modeling and system analysis by user-friendly computer simulator. Since most biochemical and bioprocess phenomena are the result of synergistic interactions among the components of reaction networks, any viable approach must be based on a nonlinear formalism whose structure permits efficient evaluation, even if the number of components and reaction process is relatively large. On the other hand, a rapid increase in CPU capability of recent prevailing computer enables us to analyze the dynamic property of a large scaled network system. Furthermore, a design of efficient graphical user interface (GUI) and familiar web browser environment can make our interaction with computers much easier and more productive. This study aims to the implementation of an efficient, user-friendly and web-based "biosimulator" named BEST-KIT (Biochemical Engineering System analyzing Tool-KIT) for analyzing a large scaled nonlinear reaction network such as "Metabolic Pathways". The BEST-KIT mainly consists of the following four modules: 2) mass-action module, 3) power-law module, 4) enzyme-kinetics modules, 5) metabolic-map module. In the module of mass-action, there are several remarkable properties such as (i) using the "mouse", user can easily design and update an arbitrary reaction scheme (nonlinear system) in the editing window (working area) through an efficient GUI even if the number of reaction components is relatively large. (ii) after editing the scheme, cumbersome simultaneous nonlinear differential equations base on generalized mass-action law can be automatically produced without writing troublesome equations. (iii) By using "server-client system", numerical calculation and nonlinear optimization of the constructed scheme are carried out in the server (virtual CPU-server) through internet and the results are sent back to the client (user's PC or workstation) and are visualized there. For this purpose, source-code of this module is programmed in C and JAVA-applet. The second module (power-law module) was designed for the case where the details of the process that govern expressions and interactions between system components are well not known. Given the several time-courses (value-change with time) of observable system components, all parameters involved in the nonlinear system formalized in "S-system or BST" are estimated in order to be fitted to the given time-courses. The genetic algorithm (GA) and structure skeletalizing was adopted as the method for nonlinear numerical optimization of huge number of parameters to be estimated. The third module (enzyme-kinetics module) was designed for simulating dynamic properties of enzymatic reaction under the assumption of steady-state. In this module, user can easily design the enzymatic reaction system by connecting substrates, products, inhibitors and activators. Since over the 10 kinds of velocity functions under the steady-state such as Michaelis-Menten type, Competitive inhibition type, Mixed nonessential activation type..., are prepared, user can design the system by selecting those reaction type and relevant reaction species. The numerical calculation of differential equations is carried out in the server as I mentioned above. In the fourth module (metabolic-map module), since graphical pictures (clickable map) of metabolic pathway (KEGG style) are saved,

user can clip out the sub-system to be analyzed within a map. The simulataneous differential equations of assigned sub-system are automatically derivated. The BEST-KIT is one of the general-purpose user-friendly simulator of metabolic and nonlinear networks, and even those who are unfamiliar with computer technology and with computer programming can easily use. Our final goal is the development of virtual-lab-system for metabolic engineering. For the purpose of this, we are planning to put BEST-KIT on our internet WWW server for open usage. You will access to BEST-KIT and carry out computer simulations from "any" platform through web-browser in the very near future.

Modeling signal pathways

Simone Bentolila, IGM, University Marne la Vallée, Noisy le Grand Cedex, France

The integration of the various types of cellular activities in a multicellular organism is mainly performed by the nervous system, the endocrine system and the immune system. In fact although DNA is indispensable for the life of the cell, outside of the context of the living cell and the intercellular communications network, DNA is basically inert. Memory exists at 2 levels: the memory of the species which consists of the unchangeable DNA on one hand, and the active memory or short-term memory of each cell, which is its metabolic state at each instant. Enzymes constitute the short-term memory of the cell, its identity, and the network of indicators of what is going to happen. It is the same for the intercellular communications network. In fact DNA is indispensable to the life of the cell: during apoptosis, or programmed cell death, in which endonuclease enzymes which digest the cell's own DNA are activated, thus rendering the DNA unusable. DNA is necessary for life, it is a matrix which is read and interpreted by the cell according to its own identity, its own biochemical context and its environment. The identity of a differentiated cell is maintained by its metabolism; a cell which loses control of its regulation dedifferentiates and loses its identity. Enzymes activate or inhibit the metabolic and differentiation pathways in which the cell may engage: some enzymes regulate the expression of genes and the synthesis of proteins from the DNA template (both structural proteins and the enzymes themselves); while others enzymes catalyse metabolic reactions, anabolic, in which products required by the cell are synthesized, and catabolic in which substrates are degraded to elementary subunits with production of energy. Intercellular communications are mainly conducted by secreted proteins (ex: hormone, growth factors, cytokines, antibodies) and exogenous ligands (ex: antigen). These circulating substances transmit a signal to competent cells using trans-receptor proteins as intermediaries; the signal is then relayed to

the interior of the cell by transduction by an enzymes cascade. The signal may be transmitted to the nuclear by transcription factors which provoke the expression or repression of a gene, or to the cytoplasm where a metabolic pathway may be activated or inactivated. Without the enzymes to promote a given pathway, reactions would occur, but would proced so slowly that the products of of a given reaction might be degraded before they could serve as substrates for the next reaction of the pathway. The progression of enzymes that serve as catalysts for a metabolic pathway form a code which switch on or off, these enzymes form the code for the metabolic pathway or word of the language. Molecules interact by contact and chemical interactions, a binding between two molecules may produce activation or inhibition of the catalytic site of one of the two molecules. This usually involves allosteric alteration or covalent modification by phosphorilation. An enzyme can be described by these 2 sites: the catalytic site and the allosteric site. An inhibitor may bind to the catalytic site and block it (isosteric modification, inhibition of the catalytic site by an analog of the substrate) or to the allosteric site and cause a change in conformation which will activate the catalytic site. An enzyme which catalyse covalent binding (interconversion), usually by phosphorilation or dephosphorilation, of another enzyme may cause activation or inactivation of the latter. In a previous paper (Bentolila, 96) we described a context sensitive grammar which models the 4 main types of genes regulation. The proposed model considers two types of objects: transcriptional units on DNA and regulatory or structural proteins which are synthesized, and which are, in the case of regulatory proteins, themselves destined to activate or repress other transcriptional units in a later phase. A transcriptional unit is described by the list of its active sites (operator, promoter, binding sites for transcription factors). A regulatory protein is described by the list of its active sites (binding domain, activation domain, binding domain for ligand). The DNA sites and the protein domains are the terminal symbols of the proposed grammar. The interaction of these proteins with the DNA, and in certain cases preliminary interactions between proteins, leads to one of two antagonistic actions: expression or repression of the transcriptional unit. These protein-protein and protein-DNA interactions are grouped into syntactic categories (induction, inhibition, initiation complex, repressor complex, activation complex) which are called biological binding operators. The expression/repression actions are described by grammar rules which provide the chain of execution by biological binding operators for the four activable/repressible regulatory systems modulated by positive/negative co-factors. If we suppose that the semantics of biological binding operators is already implemented (using a database), it is sufficient to write a context-free grammar which describes the order of application of biological binding operators, similar to the context-free grammar of arithmetic operators, for example. The grammar that we have developped describes the series of operations that leads to either the activation or inactivation of an anabolic / catabolic pathway (ex: glycogensis / glycogenolysis) or the expression / repression of a protein (ex: an-

tibody, ccytokines). We have applied this model to 2 examples: the key enzymes involved in sugar metabolism regulation in the liver which is under hormonal control; and a simplified model of the immune response.
Bentolila S., (1996) A grammar describing "biological binding operators" to model gene regulation. Biochimie 78, 335-350.

Mapping Metabolic Networks and Gene Expression Data via Protein Structure Prediction

Ralf Zimmer, GMD-SCAI, Schloss Birlinghoven, Sankt Augustin, Germany

Differences of gene expression in normal and diseased cells can lead to valuable hints for identifying potential drug targets - a problem of direct relevance to the pharmaceutical industry. - Due to the genome projects and the corresponding sequencing technology a wealth of sequence data and even complete genome sequences are available. - Charted interaction networks compile (partial) information of metabolic relationships and signal transduction mechanisms. - DNA chip technology allows to measure gene expression (mRNA level) on a genomic scale (in the case of several organisms on whole genomes). The problem is to map the sequences and the expression data onto disease connected metabolic/signalling pathways. We discuss the application of protein structure prediction methods such as 123D to a set of sequences overexpressed in yeast during the diauxic shift, i.e. the reprogramming of yeasts metabolism from anaerobic growth (fermentation) to aerobic respiration due to the exhaustion of glucose. With DNA chip techniques changes of expresssion levels of any yeast protein have been measured over time. Interestingly, several hundred proteins of still unknown function seem to be over- or under-expressed during the well-studied and fairly well understood metabolic relations of the diauxic shift. With our prediction methods it is possible to assign probable functions and testable structure models with reasonable prediction significance to a large fraction of these unknown proteins intractable with standard sequence analysis methods. Thus, the method allows to map the genomic sequence data onto evolving metabolic network knowledge and vice versa: the prediction procedures can be significantly improved using additional biological, i.e. metabolic, information and the partial metabolic networks can be extended by previously not yet identified components.

Computer tools for the analysis of yeast regulatory sequences

Jacques van Helden and Julio Collado-Vides, Centro de Investigación sobre Fijación de Nitrógeno, Universidad Nacional Autónoma de México, Cuernavaca, Mexico; Bruno André, Laboratoire de Physiologie Cellulaire et de Génétique des Levures, Université Libre de Bruxelles, Bruxelles, Belgium.

A series of computer tools were developed for the analysis of the non-coding sequences mediating transcriptional regulation, with a special focus on upstream regions from the yeast Saccharomyces cerevisiae. These tools are publicly available on the web
(http://copan.cifn.unam.mx/Computational_Biology/yeast-tools).
Basically, three classical problems can be addressed: 1) Search for known regulatory elements in known upstream regions: The first tool (upstream-region) allows to directly extract the upstream region of any known gene or predicted ORF from the genome. These sequences can then be searched (with the program dna-pattern) for specific patterns provided by the user. The program feature-map automatically draws a graph showing the location of these patterns within the upstream regions. Several genes can be represented in parallel on the same map, allowing visual detection of positional specificity. 2) Search for unknown regulatory elements in a set of known upstream regions: Oligo-analysis allows the detection of unknown regulatory elements shared by a set of co-regulated genes. The program counts the number of occurrences of each oligonucleotide among the set of input sequences, and calculates the statistical significance of each of them in order to detect over-represented patterns. The specificity of this implementation is that it uses a frequency table to calculate a distinct expected number of occurrences for each oligonucleotide. The frequency table was built by measuring oligonucleotide frequencies in the set of all non-coding genomic sequences, for each oligonucleotide of size 1 to 8. Despite its simplicity, this tool has been shown efficient in 8 out of 10 known families of co-regulated genes, and could become particularly useful for extracting unknown regulatory elements from the numerous functional families discovered by large-scale gene expression measurement (e.g. DNA microarrays). 3) Search for unknown upstream regions regulated by a known element: The complete genome can be searched for a given pattern (genome-search), and the closest ORFs from each match can be inferred (neighbour-orfs). Alternatively, pattern search can be directly performed on a subset of genomic sequences restricted to the upstream regions from the 6200 predicted ORFs (all-upstream-search). The site also includes a series of utility programs, which perform generation of random sequence (random-seq), automatic drawing of XY graphs from numeric data in columns (XYgraph),

inter-conversions between various sequence formats (convert-seq). Several tools are linked together in order to allow their sequential utilisation (piping), but each one can also be used independently by filling the web form with external data. For instance, any DNA sequence can be submitted to pattern searching (dna-pattern) or oligonucleotide analysis. The feature-map accepts several types of data for input: Swissprot, Transfac, RegulonDB, MatInspector, Signal Scan, Patser, Gibbs sampler, dssp. This widens the scope of the site to the analysis of non-regulatory and/or non-yeast sequences.

Evolutionary Optimization and Metabolic Control Analysis

Edda Klipp, Humboldt University Berlin, Institute of Biology, Theoretical Biophysics, Germany

Modeling of cellular metabolism is often done as simulation of the behaviour of the system assuming a known structure of the system, the pattern of influences and the quantitative strength of these influences. Another approach is the prediction of features of metabolic systems (e.g. the number of reactions of a pathway, enzyme concentrations, values of kinetic constants) as a conclusion from optimality principles [1,2]. These optimality principles can be regarded as an expression of the observation that biological systems are under evolutionary pressure. Here, the principle is used that the total amount of enzyme in a cell can be considered as limited or even minimized [3]. This can be explained by the fact that cellular metabolism is characterized by the necessity of parsimony in the use of resources. The synthesis of enzyme in the cell costs energy and material. The capacity of the cell to store enzyme (i.e. protein) is limited due to osmotical reasons. Steady states of metabolic systems can be described in terms of metabolic control analysis, where flux control coefficients are defined as

$$C_{\nu_k}^{J_j} = \frac{\nu_k}{J_j}\frac{\delta J_j}{\delta \nu_k} \tag{1}$$

describing the normalized change of a steady state flux (J_j) caused by a change of a certain reaction rate (ν_k). Optimized states are characterized by a special distribution of flux control in the system. In unbranched chains of enzymic reactions in states of minimal total enzyme concentration at given flux the distribution of the flux control coefficients is identical to the distribution of the relative concentrations of the individual enzymes [1]. Reactions catalysed by those enzymes that must be present in a high concentration compared to the total amount of enzyme to maintain the given flux also exert a strong flux control and vice versa.

That differs remarkably from non-optimized situations. Consider metabolic networks with a linearity between reactions rates and enzyme concentrations. At fixed steady state fluxes J the application of the principle of minimized total enzyme concentration yields the following relations between optimized enzyme concentrations and flux control coefficients:

$$C^T E = E \tag{2}$$

where C represents the matrix of normalized flux control coefficients, T indicates the transpose and E is the vector of optimized enzyme concentrations. For an unbranched chain formula (2) is equivalent to the statement of proportionality between flux control coefficients and optimized enzyme concentrations. For branched systems a matching between the control coefficients and the optimized enzyme concentration can be concluded, too. Enzymes of reactions which excert more control over the fluxes in the system than others are present in a higher concentration in states of minimized total enzyme. Eqn. (3) offers a set of relations between the control coefficients in addition to the well-known summation and connectivity theorems. For the matrix of scaled elasticities (expressing in normalized form the change of a reaction rate caused be the change of a metabolite concentration) it follows from eqn. (2) under consideration of the connectivity theorem for systems without conservation relations ($C\epsilon = 0$) that

$$\epsilon^T E = 0 \tag{3}$$

For practical purposes the use of equations (2) or (3) may allow the calculation of control coefficients if one knows fewer elasticities than would be necessary using only the summation and connectivity theorems since they supply as many new relations as there are independent internal metabolites included in the system. Hitherto, optimization principles have been applied assuming that during evolution the cellular systems have had enough time to adopt to given conditions and that changes have been fixed by mutation and selection. However, it is known that organisms are able to change gene expression patterns remarkably with changing environmental conditions (e.g., in yeast cells one can observe the change of the m-RNA level of roughly 2000 genes during the diauxic shift switching from enzymes necessary for the degradation of glucose to ethanol to enzymes of the lower part of glycolysis and of gluconeogenesis). The distribution and the total amount of enzyme present in a cell are well regulated and quickly adapted to the environmental conditions. Hence, the enzyme distribution can be understood only in the context of genetic regulation.

Heinrich, R., Klipp, E., Control analysis of unbranched enzymatic chains in states of maximal activity, J. theor. Biol. 182 (1996) 243-252.

Heinrich, R., Montero, F., Klipp, E., Waddell, T.G., Meléndez-Hevia, E., Theoretical approaches to the evolutionary optimization of glycolysis. Thermodynamic and kinetic constraints, Eur. J. Biochem. 243 (1997) 191-201.

Brown, G.C., Total cell protein concentration as an evolutionary constraint on the metabolic control distribution in cells, J. theor. Biol. 153 (1991) 195-203.

Why and how to build a conceptual bridge between mechanistic regulatory biology and quantitative genetics

Stig Omholt, Agricultural University of Norway, Department of Animal Science, Aas, Norway

The concepts of additive, dominance and additive by additive (epistatic) genetic variance of a metric character keep a central position within theoretical machinery of quantitative genetics used in such fields as plant and animal breeding, evolutionary biology, medicine and psychology. The estimation of these variance components is normally based upon performance covariances between relatives within family. The theoretical foundation for this was developed by Fisher (1918) by the use of a single locus model with two alleles, one dominant and one recessive, in a random mating population. Quantitative genetic theory of today has not in principl moved beyond this Fisherian basis. In fact, in the most sophisticated linear mixed animal models the coefficients of relationship between relatives, which are highly instrumental for the estimation of variance components, come from Fishers one locus model. The use of quantitative genetic theory has been a highly succesful entreprise within animal and plant breeding, even if its conceptual foundation is highly dubious. The theory is built upon the premise that to a large degree the contribution from each locus to the genotypic values of a metric character can be added (inter-locus additivity). Why and how genomic regulatory networks behave in such a way that the "bean bag model" of Fisherian genetics has such a predictive power when implemented within a mathematical-statistical phenomenological methodological apparatus remains to be explained. Part of the explanation may be found if we are able to establish a conceptual bridge between mechanistic regulatory biology in a wide sense and the generic phenomena of quantitative genetics. That is, if we are able to construct regulatory models catching the essential features of regulatory networks behind metric characters that produces the generic phenomena dominance, additivity and epistasis, we may be able to understand under which regulatory conditions the various phenomena are realised. We have recently addressed this question by providing simple regulatory networks displaying the phenomena of additivity, dominance, overdominance and epistasis as generic features. We show by analytical and numerical means that dominance may be an intra- as well as inter-locus interaction phenomenon, that overdominance is likely to be an inter-locus regu-

latory phenomenon, and that genetic dominance is the rule, not the exception, in the case of intra-locus interaction. In the interlocus case we find that there will be considerable room for intra-locus additive gene effects. We think our approach is instrumental for developing the theoretical foundation of quantitative genetics as well as population genetics so that we will be capable of making use of the coming huge amount of molecular biological information at the individual level also on population level phenomena.

Computer-aided Structural Analysis of Biochemical Reaction Systems

Stefan Schuster, Max-Delbrück Center for Molecular Medicine, Dept. of Bioinformatics, Berlin, Germany

The dynamic modeling of biochemical systems is often hampered by the fact that the kinetic parameters of enzymes, such as Michaelis constants and maximal activities, are imperfectly known. In the light of the recent advances in genome research, there is renewed interest in metabolic modeling. This is, however, another type of modeling than the kinetic approach. Now, structural approaches are needed, where we do not care about kinetic parameters, but rather about the topological structure of metabolism including signal transduction pathways. Our first studies on structural properties of metabolic networks concerned conservation relations among metabolite concentrations. A necessary condition for a conservation relation to represent conservation of chemical units is that all coefficients be non-negative. Adapting methods from convex analysis, we have developed an algorithm for calculating a complete set of non-negative conservation relations for systems of any complexity. This algorithm was implemented as computer programs in Turbo-Pascal and C. Another focus of our research has been on pathway analysis. It is not always straightforward to detect the precise pathway that leads from any one starting point to some product(s). Therefore, we have developed the concept of elementary flux mode. This term denotes any minimal set of enzymes that can operate at steady state with all irreversible reactions proceeding in the appropriate direction. The enzymes are weighted by the relative flux they carry. Any real flux distribution can be represented as a superposition of elementary modes. We have developed an algorithm for detecting all elementary modes for systems of any complexity. This algorithm is based on the Gauss-Jordan method and includes special conditions for meeting the irreversibility constraint and for excluding non-elementary modes. The algorithm has been implemented by several colleagues in three different programming languages. For illustration, we analyse a reaction scheme comprising the tricarboxylic acid cy-

cle, glyoxylate shunt and some adjacent reactions of amino acid metabolism in E. coli. This scheme gives rise to 16 elementary modes all of which are interpretable in terms of biochemical function. For example, two modes represent futile cycles. A mode via aspartate corresponds to a main pathway which had been proposed for Haemophilus influenzae. There are biochemical findings indicating that several non-glycolytic mycoplasma species such as M. hominis lack not only phosphofructokinase and aldolase, but also glucose-6-phosphate dehydrogenase. We found that there is no elementary mode bypassing the abovementioned enzymes. It can be concluded that there is some missing link in the metabolism of M. hominis. These examples show that the analysis presented can be used as a guideline in the reconstruction of bacterial metabolism.The approach based on elementary modes appears to be a helpful tool for understanding the complex topology of metabolic networks by a rational approach. It might also be helpful in teaching biochemistry.

Parallel Knowledge Discovery System for Amino Acid Sequences - BONSAI Garden

Takayoshi Shoudai, Universität Lübeck, Institut für Theoretische Informatik, Lübeck, Germany

We have developed a machine discovery system BONSAI which receives positive and negative examples as inputs and produces as a hypothesis a pair of a decision tree over regular patterns and an alphabet indexing
http://bonsai.ims.u-tokyo.ac.jp/bonsai
This system has succeeded in discovering reasonable knowledge on transmembrane domain sequences and signal peptide sequences by computer experiments. However, when several kinds of sequences are mixed in the data, it does not seem reasonable for a single BONSAI system to find a hypothesis of a reasonably small size with high accuracy. For this purpose, we have designed a system BONSAI Garden, in which several BONSAI's and a program called Gardener run over a network in parallel, to partition the data into some number of classes together with hypotheses explaining these classes accurately.

Application of Conceptual Clustering to the Recognition of the Hierarchical Structure of Transcriptional Control Domains

Patrizio Arrigo, C.N.R. Istituto Circuiti Elettronici, Genova, Italy

One of the most relevant task in functional genomics is the discovery of the syntactical rules that drive the gene expression. Many tools based on matemathical and biophysical approaches was applied, these methods are able to detect the binding sites of DNA and transcriptional factors. More difficult is the discovery of functional correaltions between these features. Recently some authors consider the genome like a linguistic text an they applied methods derived from computational linguistic to the analysis of this kind of text. The main difference between linguistic and biolinguistic is the availability of dictionaries and grammatical rules in latter field, insted this knowledge is relatively scarce in biolinguistic. The first step for more complex analysis is the capability to recognize potential functional word along the linear genomic sequence, in other word we need to reduce the sequence redundancy. In this work a new combined methodology is applied to process a subset of g-protein coupled receptors in order to evaluate the possibility to detect nucleotide domains and test their relations with structural or functional region of the corresponding protein. The CDS can be considered like a 'noisless' text then is more easy to evaluate the correlations between features on genomic sequence and proteins. The method combine the potentiality of an unsupervided neural clustering and informational and statistical parameters in order to extract and select domains on nucleotide sequence, their translation in the corresponding peptide and their positioning along the protein sequence. The results obtained on this dataset evidence the a good correlation between the features selected on CDS and functional regions on g-protein coupled membrane receptor. The preprint of the paper is available at the following address: http://www.biocomp.unibo.it/piero/arrigo/title.html

Hydrogen Bonds in Biopolymer Structures - Variations on an Old Theme

Jürgen Sühnel, Biocomputing, Institut für Molekulare Biotechnologie, Jena, Gemany

Hydrogen bonds belong to the most important interactions in biopolymer structures. Currently, we know the three-dimensional structures of almost 8000 biopo-

lymers and have a growth rate of about 4 new structures per day. There is almost no structure report which does not contain any information on hydrogen bonds. This means, that there is an incredible amount of information on H-bonds in biological macromolecules. Nevertheless, there is a variety of open questions and controversial issues. In this contribution about three of them is reported: the role of C-H...O and C-H...N interactions in biopolymers, the free energy contribution of H-bonds to protein stability, and unusual base pairs in nucleic acids. Recently, it has been claimed that C-H...O and C-H...N interactions may be a long- neglected stabilizing force in biopolymers. We report on a systematic geometric analysis of these interactions in experimental RNA structures. It is shown that in the RNA backbone the C2=92(H)...O4=92 and C5=92(H)...O2=92 interaction are the most promising candidates from the hydrogen bonding perspective. Taken together these interactions connect two segments of the sugar phosphate-backbone to a seven-membered ring. As DNA lacks the 2'-OH group, this structural motif is specific for RNA. Hence, it is tempting to speculate that it may contribute to the different structural features of DNA and RNA. This conjecture requires further studies. So far, there is no consensus about the free energy contribution of hydrogen bonding to protein folding. Recently, it has been proposed to convert statistical information from databases of three-dimensional experimental structures into Helmholtz free energies. For peptide hydrogen bonds between amino acids with a sequence distance larger than 8 the result was that the free energy difference between the short-distance minimum and large distances is close to zero. Hence, it was concluded that peptide H-bonds do not contribute to protein stability in terms of free energy. There is, however, a barrier for the disruption of H-bonds, which can be assumed to be responsible for maintaining protein structure when formed. These results were obtained for the N...O interaction without taking into account hydrogen atoms. We have calculated the H-atom positions, which can be done in a reliable manner for the peptide N-H groups. The very same analysis with the very same structure dataset then results in a free energy gain of slightly more than 1 kT. This value is comparable to typical hydrophobic contacts. Therefore, the claim that peptide hydrogen bonds do not contribute to protein stability has to be revised.The standard view of nucleic acids is that they consist of Watson-Crick base pairs. Especially, from the increasing number of RNA structures we know now, that non-Watson-Crick or non-canonical base pairs often occur. They have usually two standard hydrogen bonds. Recently, unusual base pairs with only one or no direct hydrogen bond and water-mediated base pairs have been detected. Little is known about their geometric properties and interaction energies. We have performed high-level quantum chemical calculations on these complexes and report results on a water-mediated UC base pair.

The Cell as an Expert System

Jaime Lagunez-Otero, Instituto de Quimica, Ciudad Universitaria, Mexico

In order to propose treatments for human ailments including those of geriatric origin, it is extremely useful to understand the underlying molecular biology. Fortunately, due to the important advances in the technology used in the field of genomics and to the results of the various genome projects, a vast amount of information is becoming available to researchers interested in determining the ethiology of pathological processes. Furthermore, important advances have also come about in the design of methods for the control of gene expression and the modification of genetic material. Genia is a project directed towards the organization and exploitation of genomic knowledge and technology. We are using paradigms adopted from the field of artificial intelligence in order to code protein interactions in the form of logical circuits as well as bio-structural data. With this framework, we believe we can take particular systems and propose comprehension schemes as well as potential therapeutic strategies. The knowledge bases include an in-house selectd set of 80 elements involved in signal transduction and rules describing their interactions. However, information will be taken from monitoring whole sets of genes using array and microarray technologies. The software is an expert system shell, presently Nexpert-Object. The technologies we are following are triplex forming oligonucleotides and other molecules able to recognize specific sequences in DNA. It should be observed then, that the signal transduction processes in the cell allow for the analysis of both external and internal parameters and for the taking of decisions in the transcription and protein synthesis system. This processes can be well emulatad with an expert system such as the one we are constructing.

Visualization of Biochemical Pathways

Franz J. Brandenburg, Bernhard Gruber, Michael Himsolt, Falk Schreiber, Institut für Theoretische Informatik, Universität Passau, Germany

This project deals with the visual representation of biochemical information. Important aspects are the representation of arbitrary parts of pathways, the comparison of biochemical reactions in different organisms, the view of regulative processes, and the consideration of compartments. The data will be managed by a database. The result of a query will mostly be a pathway. Pathways are

visualized by drawing algorithms. These treat pathways as annotated graphs and draw them in a nice way. Because of the many views to the database and the pathways, automatic drawing algorithms are necessary here. The pathways can be annotated with further information, such as structural formulas, regulative processes or compartmentation. This is neccessary to accommodate to different users, such as researchers, students or doctors. Arranging parts of pathways into single objects enhance the readability of the obtained diagrams. The underlying information is based on the Boehringer poster "Biochemical Pathways" edited by G. Michal.