Modeling and Simulation of Metabolic Pathways, Gene Regulation and Cell Differentiation

Organizers:

Julio Collado-Vides (Universidad Nacional Autonoma de Mexico)
Ralf Hofestädt (Universität Leipzig, Universität Koblenz-Landau)
Markus Löffler (Universität Leipzig)
Michael Mavrovouniotis (Northwestern University)

October 23-27, 1995

(BioEssays Meeting report)

By isolating and sequencing genes and proteins, by identifying and studying individual enzymes in metabolic pathways, by determining three-dimensional structures of biological macromolecules, and by manipulating the genetic and biochemical composition of cells and observing its consequences, the methods of molecular biology and biochemistry have amassed a large volume of data. The advances of these methods continue to produce an exponential growth in biological data, reflected in the content of databases like GENBANK, SWISSPROT and PIR, used widely by CD ROM or by electronic mail. By contrast, our ability to structure, model, and integrate these streams of biological data has lagged. Computational studies that attempt to capitalise on accumulated biological data were the primary focus of a conference which was part of the year-round series of Dagstuhl-Seminars organised at the Schloss Dagstuhl in Germany. The above four organisers attempt to give, based on the transactions at the conference, their perception of significant issues and emerging themes in Computational and Theoretical Studies of Metabolic Pathways, Gene Regulation, and Cell Differentiation.

Metabolism Enzymes are proteins that catalyse biochemical reactions, by binding to particular substrates and lowering the activation-energy barriers of specific reactions. Sequences of enzyme-catalysed steps form biochemical (or metabolic) pathways, achieving the overall transformation of substrates to a variety of products, to meet the chemical needs of the cell. Metabolic pathways can interact and create complex metabolic networks. The analysis of metabolic networks and eventual construction of novel pathways is the aim of the new research field of metabolic engineering. By its definition, this field requires an integrative view of the

metabolism: In analyzing a pathway, one may discover interactions among pathways with different physiological functions; in synthesizing a new pathway, one can make use of building blocks (enzymes) encoded in the genes of different organisms. The Boehringer Mannheim company has produced a map of metabolic pathways, that provides a colorful visual integrative tool. G. Michal discussed the obstacles, objectives and trade-offs that guided the design and notation of this metabolic wall-chart. It would be clearly desirable to have each metabolite (biochemical compound) appear as a single node in the chart, with all its reactions emanating from it. But given the large number of reactions in which some metabolites participate (like for example, currency metabolites such as NAD or ATP, or, to a lesser extent, common intermediates such as pyruvate), this would create a confusing spaghetti-like appearance. Thus, many metabolites appear as multiple nodes in different parts of the metabolic chart. To allow the user to locate metabolites and enzymes, an index is used, providing coordinates, much like a street index of a city map. Different colors and fonts are used for a categorisation of the metabolites and enzymes. These workarounds are needed because of the inherent weaknesses of maps, textbooks, or other static representations. Many attendees pointed out that a computerised version of the metabolic map would be much easier to use. Indeed, computational metabolic databases with graphical display of pathways are already being developed in various laboratories.

Especially useful for metabolic engineering is the implementation of integrative information systems, that represent genes, enzymes, and metabolic pathways. P. Karp is developing the first integrated metabolic information system for E. coli. The EcoCyc system contains information ranging from structures of metabolites and stoichiometries of reactions, to enzyme cofactors, activators and inhibitors, to protein subunit composition and genetic map localisation of the associated genes. Each object is computationally linked to related objects for easy navigation: A reaction, for example, is linked to its metabolites as well as to particular enzymes that catalyse it. A user could zoom in on a region of the genetic map, click on a gene to obtain detailed information about it, navigate to the enzyme product of the gene, and then to the metabolic pathway containing the enzyme. This system can be accessed at

http://www.ai.sri.com/ecocyc/ecocyc.html on the World Wide Web.

Another integrated metabolic database, which can be accessed at

http://www.mcs.anl.gov/home/compbio/PUMA/Production/puma.html

on the Web, was presented by T. Gaasterland. These information systems will no doubt become an important tool for metabolic engineering. In fact, as presented by Gaasterland, evolutionary metabolic reconstruction exercises are conceivable, taking into account the increasing number of completed genomic sequences available. A more quantitative source for understanding metabolic pathways is that provided by analytical models based on the use of differential equations for simulation of reaction kinetics and metabolite concentrations (H. Heinrich, D.

Kahn, G. Stephanopoulos). Models determine behavior patterns and dominant mechanisms of a biochemical system, and are able to decompose a complex pathway and in some cases determine which pathway steps limit the overall flux towards a desired final product. This analysis is especially fruitful within parts of metabolism which are well understood from a biological point of view. Most mathematical approaches rely on the formulation of the relevant equations for each particular case, which are then solved computationally. The reliance on manually formulated mathematical models does not permit multiple uses of the same information or other types of analysis. In this sense, more flexibility can be obtained by a hybrid approach of equation-oriented methods with integrated metabolic databases, as illustrated by the object-oriented computational encoding of components (metabolites, reactions, cells, medium) of the biological dynamic system, presented by G. Breuel. This interesting approach can support simulation as well as database usage. The concepts of object-oriented programming facilitate the development, maintenance and reuse of mathematical models. Rather than models describing just individual metabolic pathways, this approach aims ultimately to model the whole metabolism as a collection of interacting subsystems. Another approach to interactive simulation of metabolic networks employs discrete models for qualitative modeling. R. Hofestädt, based on the theory of formal languages, automata, and graph theoretical methods, developed a grammatical formalisation of biochemical reactions, that enables the identification of general properties such as metabolic bottlenecks. M. Mavrovouniotis showed the use of thermodynamic arguments that determine the feasibility and direction of biotransformations - and can even place limits on reaction rates. These discrete approaches are able to cope with the usual incompleteness and uncertainty of the available knowledge, which limits the more classical quantitative equation-oriented methods. Classical continuous mathematical methods, on the other hand, are easier to reconcile with experimental measurements, which are usually quantitative.

Deciphering coding regions is now a basic task associated to the computational interpretation of DNA sequences produced by genome projects. The identification of signals in the DNA related to regulatory domains was a topic of discussion at the conference. An example of astute combination of mathematics (more specifically statistics) and biological insights was the discovery by A. Danchin of a specific base-tetramer found at frequencies much lower than would be expected at random. This sequence contains a TGC triplet that is frequently found in binding sites of prokaryotic regulatory proteins, supporting the interpretation that sequences are selected against in certain domains of the genome in order to permit adequate regulatory interactions. This result came out within the framework of statistical analysis of codon usage in E. coli that supported three main biological classes of genes: high expression, low expression, and genes from horizontal transfer. Statistical analyses of larger motifs in DNA sequences are difficult given the presence of overlapping patterns and the limited availability of datasets of well-understood sequences. Other presentations concentrated on methods of pattern recognition

for binding sites of specific regulatory proteins. The analysis of prokaryotic signals, specifically sigma 70 promoters, presented by G. Hertz, led to the observation that such promoter sequences have a low information content, given the expected number of promoters and the RNA polymerase concentration. One reasonable interpretation is that in such promoters, additional activator sites participate to attract the polymerase to bind and initiate transcription. P. Bucher presented a similar method using weight matrices for binding sites, but taking also into account the context where the site is found. He emphasized the importance of protein-protein interactions in transcription regulation. These computational studies are being complemented by the experimental evaluation of the affinity of binding of proteins to promoter sequences. Improved data would permit the construction of weight matrices for a more precise computational recognition of promoters (M. Ponomarenko). Questions dealing not with individual promoters, but with collections of promoters, regulatory proteins, and the networks produced by their interactions, require integrative reliable databases. The work of E. Wingender has resulted in the TRANSFAC database with information on gene regulation of all eukaryotic organisms where experimental information has been published. An interesting observation he made is that proteins that belong to the same class defined in terms of their DNA-binding domain, i.e. homeo domains, or leucine-zipper-domains, do not interact more frequently with members of the same class than with proteins of a different class. Another database, christened OperonDB and presented by J. Collado-Vides, contains the available regulatory information for (approximately 150) sigma 70 E. coli promoters. Collado-Vides described a grammatical model that generates the sigma 70 collection as well as many new potential regulatory sequences. He presented two new directions of this approach, one dealing with a syntactic computational implementation to predict regulatory domains in genome sequences, and the second one dealing with a formalisation that incorporates gene activation at a distance, in sigma 54 promoters. This approach is centered on the anatomy of cis-regulatory domains. A complementary formal approach using boolean algebra, presented by D. Thieffry, predicts that networks formed by a small number of regulatory genes are more robust than large networks. These predictions have some empirical confirmation in the known E. coli network of transcriptional regulation. A third theoretical approach, presented by M. Savageau, was the demand theory of gene expression. Assuming that regulatory systems are under selection pressure, Savageau predicts that genes subject to a low demand are negatively regulated, whereas genes under high demand are positively regulated. In addition to a good number of prokaryotic genes that follow this rule, switching of regulation (positive to negative or vice versa) in eukaryotic cells is also in agreement with this theory. Essentially, cell-specific genes will be positively regulated in their corresponding cells, but negatively regulated in other cells.

Not all talks were coming from theory or computer science, as illustrated by a lucid description of the experimentally analyzed nitrogen regulation in bacteria, offered by B. Magasa-

nik. This complex yet logical system of regulation involves a cascade of several regulatory proteins connecting the signal of nitrogen availability to the active/inactive form of glutamine synthetase, as well as autoregulation of transcription of the glnA operon. Magasanik was able also to briefly discuss what seems to be a puzzle in the evolution of gene regulation: the plausible fallback role of a less refined regulation of the glnA operon in E. coli, which functions when some regulatory proteins of the primary system are not present or are mutated.

The coordinated regulation of the expression of genes is primarily responsible for the diversity of cell phenotypes that unfolds during the development of a higher plant or animal. During the development of a higher eukaryote, a single cell gives rise, by mitotic cell divisions, to a vast array of cell types, which are often highly specialised, each carrying out only a few metabolic functions. The mechanism by which these cells differentiate from one another during the growth and development of an organism involves control of the expression of these genes at the levels of transcription and transcript processing. Differentiation is usually a consequence of regulation of gene expression (and rather infrequently changes in genome composition, such as in the immune system). Most developmental processes in higher eukaryotes seem to be controlled by preprogrammed circuits of gene expression, where some event triggers the expression of a particular set of genes. The product of these genes functions by turning off the transcription of the first set of genes and/or turning on the transcription of a second set of genes. In turn, one or more of the products of the second set acts by turning on a third set, and so on. In these cases, the sequential expression of genes is genetically preprogrammed and the genes cannot usually be turned on out of sequence. In eukaryotes, hormones can trigger the sequential expression of sets of genes. Regulatory genes are known to be involved in the control of patterns of differentiation. In some cases regulatory elements called enhancers and silencers modulate levels of gene expressions from nearby promoters. However, the question of how these enhancers and silencers work in controlling gene expression remains a challenge. The spectrum of continuous/quantitative to discrete/qualitative techniques presented on this topic parallels the diversity of modeling approaches discussed above for metabolic pathways and gene regulation. J. Reinitz analyzed the process of segment determination in Drosophila by numerically inverting a chemical kinetic equation that describes the regulatory circuitry and accounts for the synthesis rate, diffusion and decay of gene products. C. Potten and M. Löffler described models which can explain the spatial and temporal organisation of the system relating cellular division, cell differentiation and maturation to the 3D-architecture and formation of cellular clones. The models ranged from stochastic cellular automata to describe the short term behavior of single intestinal crypts to a differential-equation model of all stages disregarding the system architecture of crypts. H. Meinhardt presented elegant models which are quantitative but as simple as possible; they attempt to capture the biological behavior not in its full detail but rather in its essential qualitative features.

Concluding Remarks

This successful meeting demonstrated the diversity of theoretical and computational approaches and the diversity of biological systems to which they can be applied. It also highlighted the utility of databases that accumulate and organise biological data and enable the study of complex biological systems on a global scale. However, despite the large volume of available data, the information is likely to be incomplete, uncertain, and qualitative on any one system of interest. Ways to cope with incomplete information must be investigated, in order to exploit this type of data and still be able to draw at least partial conclusions and predictions, as was illustrated many talks at this conference. Informal discussions in the afternoons of the conference revolved around many of the questions remain to be answered in the near future.

How informative are DNA sequences for predicting complex regulatory mechanisms ?

How far will databases bring us into a more integrated understanding of gene regulation and metabolism of the whole cell ?

How much of the complexity of regulatory networks, of metabolic networks and of pattern formation can be explained by standard evolutionary ideas ?

What is a computationally useful definition of a metabolic pathway ?

How much of the biological information available can be understood and reconstructed with integrative theoretical approaches ?

These are some of the intriguing questions of a more global molecular biology in which the computational and theoretical efforts will play a central role.

During the workshop 31 lectures which covered various topics have been presented by participants from different countries. All participants appreciated the stimulating and cordial atmosphere at Schloss Dagstuhl. The always engaged support of the Dagstuhl team was an essential contribution to the success of this seminar.

The organizers wish to thank the VW-Stiftung for its generous financial support and all those who helped to make the workshop an interesting and fruitful research experience.

December 1995

Julio Collado-Vides
Ralf Hofestädt
Markus Löffler
Michael Mavrovouniotis

Monday, October 23

Gene Regulation

M. Ponomarenko (Novosibirsk)
The Structure of Eukariotic Promoters and Computer Methods of its Recognition

P. Bucher (Lausanne)
Mathematical Methods to Characterize Complex Transcriptional Control Regions

J. Collado-Vides (UNAM Mexiko)
Syntactic Recognition of Regulatory Regions in E.coli

D. Thieffry (UNAM Mexiko)
Theoretical Analysis of E. coli Genetic Regulatory Network

M. Savageau (University Michigan)
Evolution of Gene Regulation

A. Danchin (Pasteur Paris)
Some Global "in silico" Analyses from Bacterial Genome Projects

E. Wingender (GBF, Braunschweig)
The Language of the Genome to Control Transcription

B. Magasanik (MIT)
Nitrogen Regulatory Network of Enten Bacteria

Tuesday, October 24

Cell Differentiation

J. Reinitz (New York)
The Dynamics of Segmentation in the Fruit Fly Drosophila melanogaster

M. Löffler (University Leipzig), C. Potten (University Manchester)
Modeling Spatial and Temporal Organisation of Epithelia

A. Dress (University Bielefeld)
A Mathematical Model for Sequential Cell Differentiation Processes

V. Calenbuhr (ISPRA Italien)
Natural Tolerance as a Function of Network Connectivity

S. Bonhöffer (Oxford)
HIV and HBV Dynamics in Vivo

U. Behn (University Leipzig)
Memory in the Immune System

H. Meinhardt (MPI Tübingen)
Models of Biological Pattern-Formation

Wednesday, October 25

Metabolic Pathways

P. Bork (EMBL Heidelberg)
From Genomes via Protein Function to Pathways

G. Michal (Boehringer Mannheim)
Graphic Representation of Biochemical Pathways

R. Hofestädt (University Leipzig, University Koblenz)
Grammatical Formalization of Metabolic Pathways

P. Karp (SRI USA)
Encyclopedia of E. coli Genes and Metabolism

Thursday, October 26

Metabolic Pathways

U. Mischke (Children's Hospital Reutlingen)
Causal Probabilistic Networks as a Tool for Representation of the Knowledge of Inborn Errors of Metabolism

C. Sensen (Halifax)
Sequencing and Analysis of the Sulfolobus solfataricus P2 Genome

M. Mavrovouniotis (Northwestern University)
Modeling Metabolic Pathways with Incomplete Information

T. Gaasterland (University Chicago)
Sequence Interpretation and Metabolic Reconstruction in an Ongoing Genome Sequencing Project

G. Breuel (University Stuttgart)
An Object-Oriented Approach to the Modeling of Bacterial Metabolism

R. Heinrich (HU Berlin)
The Structural Design of Glycolysis, Kinetic and Thermodynamic Constraints

D. Kahn (INRA Frankreich)
Introduction to Metabolic Control Theory

S. Schuster (HU Berlin)
Modern Developments in Metabolic Control Theory

G. Stephanopoulos (MIT)
Metabolic Network Dynamics Analysis as Tool of Metabolic Engineering

Friday, October 27

Metabolic Pathways

B. Pohl (University Würzburg)
Dynamic Mathematical Models in Knowledge-Based Systems

G. Hertz (University Collorado)
Alignment Matrices for Modeling Complex Interactions between DNA and Regulatory Proteins

J. Wertheimer (MIT)
Reasoning from Experiments to Causal Models in Molecular Cell Biology

The Structure of Eukariotic Promoters and Computer Methods of its Recognition

M.P.Ponomarenko, L.K.Savinkova, A.E.Kel, O.V.Kel,
A.N.Kolchanova, Y.V.Kondrakhin, F.Kolpakov, A.G.Romaschenko and A.N.Kolchanov.
Institute of Cytology and Genetics, Novosibirsk

In the present work, eukariotic promoters have been analyzed.
The Transcription Regulatory Regions Databank (TRRD) was created. The TRRD current release contains data on regulatory regions of 270 genes. The ssDNA/TBP-binding efficiency have been measured. On the base of these magnitudes the computer method for TBP-binding efficiency calculation by an arbitrary ssDNA nucleotide sequence was developed. By this method the optimal nucleotide sequences for the TATA box were simulated. It was shown, the simulated and real TATA box sequences were reliably similar in both nucleotide frequencies and the Bucher's matrix method recognitions.
To recognize eukariotic promoters, the computer method of the transcription regularity potential calculation was made up. The 2nd type error (over-recognizing) following application of this method was at list of 2 time lower than following the method of proximal promoters recognizing by its consensuses.
This research was supported by the Fundamental Research Foundation, Russia, grant 04.12469-a.

Mathematical Methods to Characterize Complex Transcriptional Control Regions

Philipp Bucher
Institut Suissede Recherches Expérimentales sur le Cancer
Epalinges sur Lausanne

Eukaryotic gene expression is controlled by complex DNA regions consisting of a multitude of cis-acting regulatory elements. The elements are target sites of transcription regulatory proteins. The activation of a complex control region involves several steps, starting with chromatin reorganisation and ending with the assembly of a functional transcription initiation complex.
Subsets of individual elements act at different stages of this process. In order to predict the regulatory properties of a given control region one has to be able to locate the individual elements and to quantitatively predict their binding strengths to cognate transcription factors. At a second stage of analysis, one has to be able to assess cooperative effects between elements. Both problems can be approached by computer-aided sequence analysis. The method presented in my talk concerns the first one.
The most common eukaryotic promoter elements show a strong over-representation at characteristic distances upstream of experimentally determined transcription initiation sites. My method exploits this property for the purpose of deriving a quantitative description of such an element providing strength estimates for individual instances. The element is formally represented by a position-specific weight matrix plus a region of preferential occurrence. The parameters of an element are estimated by means of a heuristic optimisation procedure using a quantitative measure of local over-representation as an optimisation criterion. Weight matrices characterizing four important eukaryotic promoter elements, TATA-, CCAAT-, GC-box and "Initiator", have been derived this way. Each of these elements has its maximum of local over-representation at a different position relative to the transcription initiation site and corresponds to a distinct local optimum in the search space defined by a general promoter sequence set.

Syntactic Recognition of Regulatory Regions in E. coli

Julio Collado-Vides
CIFN
Universidad Nacional Autonoma de Mexico, Cuernavaca

We have collected and analyzed an exhaustive database of s70 and s54 promoters in E. coli and Salmonella where there is sufficient knowledge on the regulation of these genes. There are around 130 s70 promoters where regulation at the level of initiation of transcription has been characterized, as well as around 20 s54 promoters (including also those in Klebsiella). As a result of the analysis of these two collections, some biological principles have been proposed that set the frame for the variation of alternative mechanisms of regulation in the s70 type of promoters.

Three important general principles came out from our analysis of the s70 collection. Any promoter requires a site for the binding of the regulator, either positive or negative, close to the promoter in order to enable direct contact between the regulator and the polymerase. These sites, called proximal sites, differ from remote sites that occur at positions far away, either downstream or more commonly upstream, like the enhancer-like sites from where s54 promoters are activated.

The second observation is that although repressor proteins can occur within a wide range of positions, around 100 base pairs long over the promoter sequence, the positions for activators are restricted to positions upstream from -30. This distribution makes sense in terms of activating mechanisms. Finally, we observed that particular proteins have preference within promoters for either single sites or for multiple sites. It is clear that binding sites do not occur as unrelated elements. For instance, one can observe in the dataset that groups of sites occur together in different promoters. A system of regulation is defined as the collection of sites in a regulatory domain whose bound regulatory proteins participate in a single mechanism of regulation. There are three types of phrases: positive, negative, and heterologous phrases.

These biological principles have been formalized into a grammatical model describing regulatory mechanisms in s70 promoters. This model generates all the promoters of the s70 collection, as well as many more which are predicted to fit within these general principles.

This grammatical model has been computationally implemented in prolog. The prolog version of the grammar generates all the representations of the 131 regulatory arrays, plus many more (in the order of 4000) arrays, or regulatory sentences that represent potential regulatory domains. The linguistic representation describes these regulatory regions by symbols that correspond to the binding sites of regulatory proteins and the promoter. Coupling these terminal symbols with algorithms or sensors for the identification of protein-specific binding sites (consensus matrices) will permit to make predictions in unannotated DNA sequences located upstream of identified open reading frames. The modeling, its implementation, and its use for predicting new regulatory domains will be discussed.

Regulation in s54 promoters shows quite different properties. Bacterial s54 promoters have much more in common with regulation in higher organisms than s70 promoters. They share with eukaryotic promoters the ability to be activated from protein binding at remote sites, that is to say, from enhancer-like binding sites; the ability of RNA polymerase to bind in a stable conformation to the promoter, where it can wait to be activated, as well as the requirement for an ATPase activity. In other words, s54 promoters are, at least conceptually, an intermediate step between the classic s70 bacterial promoters and promoters from higher organisms. We have initiated the construction of a grammatical model that includes the mechanisms of gene regulation associated with s54 promoters. This will be an important intermediate step to the future search for grammatical modeling of gene regulation in higher organisms.

Gene regulation in eukaryotic promoters involves a much larger number of molecules, making its description and formalization far more difficult than that of the bacterial promoters here presented. Nonetheless, a similar procedure can in principle be followed for the analysis of regulatory mechanisms and promoters in eukaryotic systems.

Theoretical Analysis of E. coli Genetic Regulatory Network

Denis Thieffry
DEM - CIFN - UNAM
Universidad Nacional Autonoma de Mexico, Cuernavaca

This talk focus on the structural and dynamical analysis of genetic regulatory networks. First, the dynamical roles of positive and negative feedback loops are discussed. Then, a generalized logical formalism is presented, allowing the dissociation of complex regulatory networks into well defined sets of simple feedback loops, yet keeping a complete control on the ways in which these loops are interconnected.

To exemplify this approach, we review a series of application to the analysis of models of gene regulatory network, including a model of the genetic regulation of the expression of bacteriophage lambda, and a model for the genetic regulation of arginine metabolism in E. coli.

The idea of the use of the logical structure as a criterion for the classification of genetic regulatory networks is then discussed. A first classification scheme is based on the presence of positive, or of negative, or of both types of feedback loops in the regulatory structure, leading to specific types of dynamical behavior. This scheme can be further refined, allowing for theoretical and experimental comparisons of genetic regulatory networks which may be otherwise unrelated.

Finally, theoretical and experimental considerations are used to obtain a first characterization of the genetic regulatory network of E.coli as a whole, including an estimate of its connectivity, a quantification of the numbers and types of feedback loops, etc. The preliminary results obtained are then discussed within both physiological and evolutionary perspectives.

Evolution of Gene Regulation

Michael A. Savageau
Department of Microbiology and Immunology
University of Michigan

The great diversity of patterns of gene regulation raises questions about its meaning. Because there are many examples of systems that appear to accomplish the same results by different means, some biologists have argued that this diversity is the result of historical accidents that have become fixed in the population. Others have argued that this diversity has been selected for the perfomance of specific functions. The first alternative implies the dominance of stochastic processes, whereas the second implies the dominance of an optimization process for which one must identify an appropriate representation of the variant systems, functional criteria for the selection and physiological constraints within which selection operates. In this paper I examine two well-characterized aspects of gene regulation and show that these can be understood on the basis of designs that have been optimized by natural selection. The first aspect is concerned with the molecular mode of gene control -- positive or negative. The second aspect is concerned with the coupling of regulator and effector gene expression -- directly coupled, uncoupled, or inversely coupled. The results can be summarized by two simply stated rules. (1) The positive mode of gene control occurs when there is a high demand for expression of the effector gene in the organism's natural environment; the negative mode occurs when there is a low demand for expression in the natural environment. (2) Direct coupling occurs with positively-controlled effector genes whose induction characteristic exhibits a high logarithmic gain in expression, or with negatively-controlled effector genes whose induction characteristic exhibits a low logarithmic gain. Inverse coupling occurs with positively-controlled systems having low logarithmic gain, or with negatively-controlled systems having high logarithmic gain. Uncoupling occurs with both positively- and negatively-controlled systems having intermediate logarithmic gain. These rules are in reasonable agreement with experimental data, which consists of more than 100 well-characterized systems in the case of rule 1 and 30 well-characterized systems in the case of rule 2.

Some Global "In Silico" Analyses from Bacterial Genome Projects

Antoine Danchin
Unité de Régulation de l' Expression Génétique
Institut Pasteur, Paris

Availability of large segments of genome sequences allows one to analyze some features of their biological meaning. Starting from the simplest point of view it is possible to investigate the dinucleotide frequency in windows of fixed length (e.g. 500nt). As already published (Nussinov, 1981) there is a general bias in all genomes as compared to a random poisson distribution (e.g.

AT is more frequent than TA). But some windows display extreme uneven distribution. This defined ordered sequence DNA (dos DNA) is distributed in S. crevisiae, E. coli and B. subtilis as well inside genes and outside genes, with a periodicity of about 80 kb. A tentative explanation is that DNA polymerase encounters some physical barrier (due to DNA folding) at regular positions. Another way to study sequences globally is to study the codon usage in genes. In E. coli (and now in B. subtilis) there are three classes of gene differing by their biological function. If one places in a same set all genes having exact counterparts from two organisms (E. coli and B. subtilis, or B. subtilis and S. cerevisiae) one finds that the cloud of gene points in the code usage space (61-dimensions) separates into two well defined entities, indicating that the genomes have different "styles". Furthermore the points are spread in each cloud following a similar pattern, suggesting that a same selecting pressure is operating in each organism, resulting in a same type of codon usage deviation from the average that is specific to the organism, as a function of the biological function considered. Several examples of comparisons between genomes permitting to propose new metabolic functions have also been discussed (for example the hypothesis that the function of polynucleotide phosphorylase is to make CDP as a precuson for DNA synthesis).

Finally, the distribution of GATC sites in the E. coli chromosome has been studied. It has been shown that following conceptual trends initiated by information theories (algorithmic complexity and Bennett's logical depth), it was possible to find a base line to study the real chromosome. This permitted us to show that GATC sites are overrepresented at a period of ca. 1100 bp (this corresponds to the long patch mismatch repair system). They are underrepresented when parts of CAP binding sites (TGTGATCNNGATCACA). Finally they are overrepresented when separated by 10, 19, and 70 bp. This generally corresponds to region having clusters of GATC motifs. These clusters are mostly located in genes involved in the transition from anaerobic growth conditions to aerobic growth conditions.

These global analysis of long sequences allow one to give hints about new regulatory pathways in bacterial metabolism.

As a motto for this week I proposed a sentence from Democritus' master, Leucippus:

Nothing comes to being by itself, but all because of a reason, and under the constraint of necessity.

Chance and necessity is not a Greek thought, and is not in Democritus.

This should remind us to read and follow our precursors.

The Language of the Genome to Control Transcription

Edgar Wingender

Department of Genome Analysis
GBF Braunschweig

The words of the genomic language that give the information when and where genes have to be expressed (transcribed) usually are sequence elements between 5 and 25 base pairs. Several of these words constitute the whole sentence of the regulatory regions of a gene, comprising promoter and/or enhancer structures as principal and subordinate clauses, respectively and presumably following a specific syntax. As a 'full stop' of these sentences, S/MARs (scaffold/ matrix attached regions) appear to define the borders for the activity of enhancers and other types of regulatory elements. To decode this language, we started to characterize the constituting words by collecting the available experimental data in a database of transcription regulating factors and their binding sites (TRANSFAC) (Wingender, 1994). Among other sources, this database is available through the WWW (http://transfac.gbf-braunschweig.de). As a part of the database, a collection of nucleotide distribution matrices is offered to the user which can be applied for sequence scanning purposes to identify putative regulatory elements. Moreover, based on the sequence data of TRANSFAC, weighted matrices and consensus descriptions have been constructed using two programs developed by T. Werner and coworkers (Frech et al., 1993; Quandt et al., submitted). Appropriate sequence analysis tools using these matrices and consensus descriptions will be available through the WWW in near future as well.

A language can be used for information transfer only if the recipient is able to understand it. The biological recipient of regulatory genomic signals is the set of transcription factors available in each specific cell type. To assess what a cell can understand, the FACTORS table of the TRANSFAC database does not only hold information about the names and synonyms of transcription factors and their physicochemical properties, but comprises also data about their presence in individual cell types and tissues. Moreover, most factors act as homo- and heterodimers (and, occasionally, as higher oligomers) which opens up a considerable combinatorial variability for the integration and interference of distinct signalling pathways aiming at different factors. Modelling the underlying regulatory mechanisms will help us to understand how the one-dimensional information carrier DNA is able to organize a complex multicellular organism.

Nitrogen Regulatory Network of Enten Bacteria

Boris Magasanik
Department of Biology
MIT

The external stimulus for the activation of nitrogen regulated genes is a drop in ammonia. This

stimulus affects the enzyme glutamine synthetase (GS), resulting in the intracellular signal, a drop in glutamine. The signal is transduced by the signal transducers PII and Utase to the modulator, NRII, which in turn phosphorylates the response regulator NRI. NRI-phophate activates transcription at the glnAp2, sigma54-dependent promoter of the glnALG operon, coding respectively for GS, NRII and NRI. When ammonia in the medium is high, PII in combination with NRII dephosphorylates NRI-phosphate and thus blocks initiation at glnAp2. A decline in intracellular glutamine, resulting from a low ammonia concentration stimulates Utase to uridylylate PII, and thus to liberate NRII to act as kinase for NRI.
The increase of the levels of GS allows the cell to utilize ammonia in low concentration. The increase in NRI-phosphate, resulting from increased transcription of NRI raises the level of this activator sufficiently to activate transcription at sigma54-promoters for the activators Nif A, which activates the nif (nitrogen fixation genes) or NAC, which activates the genes for degradation of histidine, proline, and urea, capable of supplying the cell with ammonia.

The Dynamics of Segmentation in the Fruit Fly Drosophila melanogaster

John Reinitz
Brookdale Center, Mt. Sinai Medical School, New York

We analyze the process of segment determination in Drosophila using the following approach. We find the circuitry by numerically inverting a chemical kinetic equation that contains parameters that describe the regulatory circuitry together with synthesis rate, diffusion and decay. The inversion is accomplished by means of fitting to observed expression patterns by simulated annealing. Using this method, we analyze the control of pair-rule gene expression by the gap genes. Today we present two results, which show:

1) Each of the 8 borders of even-skipped stripes 2 - 5 is under the control of a certain gap domain.
2) The gap genes specify exactly one set of pair-rule stripes.

Modeling Spatial and Temporal Organisation of Epithelia

Markus Löffler[1], Christopher Potten[2]
[1]Department of Medical Informatics, University of Leipzig
[2]CRC Dep. of Epitelial Biology, Christie Hospital (NHS) Trust Paterson Institute, Manchester

Epithelia are highly structured tissues which are often characterized by macroscopic stability despite a high daily regeneration and accociated microscopic dynamic. The objective of the talk was to describe models which can explain the spatial and temporal organisation of the systems relating cellular division, cell differentiation and maturation to the 3D-architecture. This implies assumptions on cell migration and formation of cellular clones.
Three models were discussed:

1) A stochastic cellular automaton to describe the short term behavior of single intestinal crypts. This model provided support for the cellular pedigree concept of stem cells => transitamplifying => mature cells.
2) A stochastic threshold dependent Galton-Watson process to describe the long term dynamics of populations of crypts on the basis of a hidden Markov process occuring at the stem cell level.
3) An ODE model of all cell stages disregarding the system architecture of crypts. This model was used to examine short term data on the recovery after a severe perturbation.

Requirements for a comprehensive unified dynamic network model were discussed.

A Mathematical Model for Sequential Cell Differentiation Processes

Andreas Dress
Department of Mathematics
University of Bielefeld

Cell differentiation processes are often modeled in terms of compartments comprising cells of a given well-defined type which are “fed” by cells from compartments comprising cells in an earlier stage of development, get enlarged by cell division, and loose cells either by cell death or by further differentiation into later stages of development.
To check these models empirically, it is imperative that the cell-differentiation dynamics implied by the combination of these processes are worked out explicitely, so that relevant system parameters can be evaluated and numerical relations (“identities”) between these parameters which are implied by the model can be derived analytically and then examined experimentally. In the lecture, a first attempt in this direction was presented, based on the most simple dynamics imaginable (that is, using constant rates of change all over), and some consequences were

derived which, in principle, are amenable to experimental examination.

Natural Tolerance as a Function of Network Connectivity

V. Calenbuhr[1,3] , H. Bersini[2] , F. J. Varela[1]
[1]CREA, Ecole Polytechnique, Paris
[2]IRIDIA, Université Libre de Bruxelles
[3]please send correspondence to this author

This article investigates the following basic question: in the relatively stable molecular environment of a vertebrate body, can a dynamic idiotypic immune network develop a natural tolerance to endogenous components? Our approach is based on stability analysis and computer simulation using a model that takes into account the dynamics of two agents of the immune system, namely, B-lymphocytes and antibodies. We investigate the behavior of simple immune networks in interaction with an Ag whose concentration is being held constant as a function of the connectivity matrix of the network. The latter is characterized by the total number of clones, N, and the number of clones, C, with which each clone interacts. The idiotypic network models typically become unstable in the presence of this type of Ag. We show that idiotypic networks that can be found in particular connected regions of NC-space show tolerance towards auto-Ag without the need for ad hoc mechanisms that prevent an immune response. These tolerant network structures provide dynamical regimes in which the clone which interacts with the auto-Ag is suppressed instead of being excited such that an unbounded immune response does not occur. Possible implications for the future treatment of auto-immune disease such as IvIg-treatment are discussed in the light of these results. Moreover, we propose an experimental approach to verify the results of the present theoretical study.

HIV and HBV Dynamics in Vivo

Sebastin Bonhoeffer
Wellcome Center for the Epidemiology of Infections Disease
Department of Zoology
University of Oxford

Recently new anti-retroviral drugs became available that are very potent inhibitons of viral replication. Administration of these drugs to infected patients is followed by a rapid decline of the free virus population over several orders of magnitude. Crucial kinetic parameters for the

dynamics of viral infections in vivo can be estimated by fitting mathematical models to data for the decline of free virus in drug-treated patients. Data from 22 HIV infected individuals showed a half life of $t_{1/2}$ = 2.0 +,- 0.9 days in the virus producing cell population and a half life of $t_{1/2}$ < 1.5 +,- 0.5 days in the free virus population. These data indicate that HIV infection is a highly dynamic process with high rates of viral reproduction and clearance. Comparison with HBV infection shows that the free virus populations may turnover at similar rates in both infections. The half life of the virus producing cell population in HBV, however, is between 1 and 2 orders of magnitude larger than in HIV. These results may help to explain some of the crucial properties of HIV infection, the enormous genetic diversity and the rapid emergence of drug resistant strains.

Memory in the Immune System

Ulrich Behn
Department of Theoretical Physics
University of Leipzig

In the talk, first the nonlinear dynamics of ad hoc assumed small subsystems of the idiotypic network is discussed. Especially we consider idiotype-antiidiotype cycles coupling to an antigen. The attractors of the dynamics describe, depending on the parameters, the virgin state, a healthy immunized state, and a state of chronic infection; in other words; immunological memory. Several models of increasing complexity preserve this property. A therapy of chronic infection by specific stimulation through injection of antigen is shortly discussed.
In the second part the architecture of the idiotypic network is investigated analyzing a bit chain representation of receptors which leads to the problem of random cluster formation on a hypercube. It is possible to choose the parameters such that a randomly chosen antigen is recognized almost surely by antibodies which are members of small clusters. In this care the idiotypic network consists of many small clusters (the peripherical system) and one large throughly connected cluster (the central immune system). This justifies to consider small clusters separately which presure memory.
In the third part, an internal structure of the central immune system is assumed idealizing experimental results of Kearney et al. We consider a fully connected core and coupled pairs of idiotype and antiidiotype which couple weakly to the core. Self-antigen is supposed to couple to both the core and the idiotype. The attractors of this system describe then - among others - a tolerant state towards self-antigen and an autoimmune state. Infection by an antigen coupling either to the idiotype or an antiidiotype may induce transitions fro and back between the tolerant

and the autoimmune state which may be exploited for designing a therapy.

Models of Biological Pattern-Formation

Hans Meinhardt
MPI für Entwicklungsbiologie, Tübingen

Models of biological pattern formation and their coupling will be discussed. It will be proposed that the following processes play a key role.
(i) Primary pattern formation is accomplished by autocatalysis and long ranging inhibition. Gradients, periodic distributions and stripe-like pattern can be generated in this way.
(ii) Cells obtain a stable state of differentiation by direct or indirect autoregulation of genes accompanied by a mutual competition among alternative genes. In this way, only one of several alternative genes can remain active within a particular cell. Which of the genes becomes activated can be under the control of a gradient generated by the mechanism mentioned above. After the correct gene activation has been achieved, the gene activity is independent of the evoking signal.
(iii) By mutual long range stabilization of cell states, a controlled neighbourhood of structures can be achieved. Segmentation such as seen in insects is proposed to result by a cyclic mutual activation of such locally self-stabilizing cell states. For the segmentation of insects, the repetition of at least three cell states are necessary to generate this periodic structure that has an internal polarity of the repetitive units.
(iv) Boundaries between regions generated by these mechanisms can obtain organizing properties for the finer subdivision of an organism. Substructures such as eyes, legs or wings are proposed to be initiated around the intersection of two borders. This mechanism accounts for the pair-wise initiation of these structures at the correct position and with the correct handedness. Classical experiments and recent molecular-genetic observation with insect and vertebrate limbs will be discussed in the view of this model.
Many such elementary steps are required for development of higher organisms. To allow the generation of complex patterns in a reproducible way, it is assumed that these elementary steps are coupled to each other in a hierarchical way. The patterns of one level exert a strong influence on the subsequent pattern. Therefore, each subsequent pattern has a precise spatial relationship to pattern of the hierarchically higher level. With Hydra as an example it will be shown how two organizing regions can organize a morphogenetic field from two opposite poles (head and foot) and how two structures can emerge in a predictable arrangement. The same mechanisms can lead also to complex space-time patterns. Pigmentation patterns on the shell of tropical snail are

given as examples.
The models are given as coupled non-linear differential equations. By computer simulations it is shown that their dynamic properties correspond to the experimental observations.

From Genomes via Protein Function to Pathways

Peer Bork
Max-Delbrück-Center for Molecular Medicine, Berlin-Buch and
EMBL, Heidelberg

More than 80% of all known genes have at least one identifyable homologue in current databases, for the majority of them functional predictions are possible. However, they have to be done very carefully as both overprediction and missed functional indications hamper a more complex pathway analysis considerably.
With the progress of the genome sequencing projects, we are becoming able to compare the protein composition and to analyse the functional composition in different model organisms. Based on data from Mycoplasma capricolum, E.coli, H.influenza, yeast and that from higher eukaryotes, a picture of the drift of protein function from metabolism to regulation and communication can be observed. More "modern" eukaryotic proteins, involved in communication and regulation, tend to have a modular architecture, the complete elucidation of which require sophisticated analysis tools.
Comparative analysis also allows to trace the evolution of metabolic pathways by adding information about homology to each enzyme of a pathway. Evolution of pathways via gene duplications seems to be a frequent theme; however, there are enzyme families that seem to be inserted into different pathways. With the constraint of knowing the complete gene pool of an organism, comparative sequence analysis also can explain nutrient requirements and guides the prediction of novel pathways. As for most of the modular regulatory proteins pathways have not yet been identified, the focus is here a classification of the modules that leads, together with experimental work, to the identification of interaction partners as a first step in pathway identification.

Graphic Representation of Biochemical Pathways

Gerhard Michal

Basic problems are dicussed, which arise during designing of metabolic charts (for example the 'Biochemical Pathways'chart, which is distributed by Boehringer Mannheim).
Verbal descriptions easily allow the emphasizing of certain aspects and putting aside others. Graphic representations result in higher compression of data and thus allow quick recognition of the essentials of a statement, but also of portions left off or dealt with in less detail.
The following aspects have to be considered:

- 3 dimensions of space
- the progress in time
- the variations among the kingdoms of biology, orders, families etc.
- the variations within an individual: between organs, at various stages of life, nutritional changes etc.
- regulation of biological activity, especially
    * by changing the amount of enzymes (during transcription, translation or by degradation)
    * by changing the activity of enzymes (by activation or by inhibition via different mechanisms or by degradation).

In case of printed material, this has to be shown in only 2 dimensions. Some help can be obtained by graphic means (colors, line shapes etc.). Computers are less restricted.
The 'language' (style of representation) depends on the kind of readers addressed. Chemical formulae show a high degree of abstraction. More recent developments require additionally indicating biological structures. Usually, full detail of these structures distracts from the reactions, but some resemblance to them is helpful. There is also the task of distinguishing between the various types of regulation. Unless one can devote a special drawing to it, one has to resort to graphic means (color, dashed vs. solid lines etc.). Finally, didactic aspects should be considered carefully in order to improve legebility.
All these aspects have to be adapted to the general scope of the task (overall or detailed representation, available space etc.).
The rules differ by the means employed (wallchart, books, computer representation). The first two ones are more useful for an overall survey, the latter one for going into more detail.

Grammatical Formalization of Metabolic Pathways

Ralf Hofestädt
Department of Medical Informatics, University of Leipzig
Department of Computer Science, University of Koblenz-Landau

Biotechnological methods allow the analyzing of biochemical processes. Enzymes are biosynthetic products of specific structure genes, which catalyze biochemical processes. Metabolic pathways are cascades of biochemical reactions, which can interact and create complex metabolic networks. Analyzing and synthezising of metabolic networks is the main aim of the new research field of metabolic engineering. A fundamental element for the realization of metabolic engineering is the implementation of integrative information systems, which represent genes, enzymes, and metabolic pathways. Moreover, modeling of metabolic processes in combination with such information systems will be the basic tool for metabolic engineering. Therefore, dynamic models are important, which allow the implementation of useful interactive simulation programmes.
In the research area of modeling metabolic processes different models are discussed. Abstract models are based on binary automata or logical approches which allow the qualitative discussion on an abstract level. Analytical models are based on the usage of differential equations which allow the exact simulation of concentration rates. However, the simulation of kinetic effects is possible. The disadvantage of this approach is the effort of the computational complexity and the fact that concentration rates are not available nowadays. Discrete models allow the qualitative modeling of metabolic networks and are based on the theory of formal languages, automata, graph theoretical approaches, and methods of artificial intelligence.
The aim of our work is to develop a new concept for the modeling and simulation of metabolic processes. Therefore, we defined a grammatical formalization. In this talk we present this new method which represents the first method for the interactive modeling and simulation of complex metabolic processes. The approach is important in the research field of metabolic engineering, because analysis of metabolic pathways is becoming more and more importance in biomedicine and biotechnology. The reason is that genetic defects cause diseases and must be identified. An important application is the detection of metabolic bottlenecks because such configurations signal specific concentration rates, which are based on genetic defects.

Encyclopedia of E. coli Genes and Metabolism

Peter Karp
SRI International, AI Center, Menlo Park

EcoCyc is a knowledge base of E. coli genes and metabolism that runs on Unix Workstations, and through the WWW (see http://www.ai.sri.com/ecocyc/ecocyc.html). Its graphical user-interface creates drawings of metabolic pathways, of individual reactions, and of the E. coli genomic map. Users can call up objects through a variety of queries (such as retrieving an

enzyme by a substring search), and then navigate to related entities shown in the resulting display window. For example, a user could zoom in on a region of the genetic map, click on a gene to obtain detailed information about it, and then navigate to the enzyme product of the gene, and then to the metabolic pathway containing the enzyme.
Metabolic pathway drawings are produced automatically, and can be drawn in several styles, such as with compound structures present or absent. The EcoCyc knowledge base currently contains information about 100 metabolic pathways, 300 enzymes, 580 enzymatic reactions, 1200 metabolic compounds, and 2500 E. coli genes. It will eventually contain information about all pathways, enzymes, and reactions of E. coli metabolism. EcoCyc contains extensive information about each enzyme, including its cofactors, activators and inhibitors (qualified by type), subunit composition, substrate specificity, and molecular weight. Individual values in the knowledge base are extensively annotated with citations to the literature, as are comment fields.

Causal Probabilistic Networks as a Tool for Representation of the Knowledge of Inborn Errors of Metabolism

U. Mischke, G. Frauendienst-Egger, F.K. Trefz
Children's Hospital Reutlingen,
Medical School of the University of Tübingen
(Sponsored by the German Ministry of Science and Technology grant MEDWIS A40)

In our project we have been developing a knowledgebased system for the diagnostic support of inborn errors of metabolism (KBS-DIAMET). Inborn errors of metabolism are heretical diseases. Each single disease is very seldom, but all together they are quite frequent.
Even for a specialist it is difficult to know each disease, but a fast precise diagnosis is necessary, as most clinical signs could be avoided by a special dietary treatment. The knowledge about inborn errors of metabolism is very uncertain, because it is difficult to obtain statistic as the patients are spread world wide. Additionally, the knowledge is very dynamically increasing, the progress in biochemical and genetic diagnostic procedures leads to 10-15 new diseases found every year.
We decided to use "Causal Probabilistic Networks" (CPN) to structure this knowledge. CPN are based on the Formula of Bayes. Qualitatively a directed acyclic graph has been built, with the nodes representing medical entities and the edges describing their dependencies. The quantitative relationships have been scored with "extended linear models"(ELM). For ELM the assumption is made, that each parent-node has a defined impact onto the child node. Also it is required that all variables are either discrete or continuous and that they are normally distributed. The

problem was solved choosing adequate ranges and the resulting error was kept small.
Mainly we chose CPN together with ELM for our system, because this is a tool handling uncertainty which is highly orientated towards the thinking processes of the physicians. Also it remains consistent if new diseases are added into the network.
We developed a network (about 120 nodes) dealing with all diseases resulting in hyperammonemia. This was implemented into the expert system shell "HUGIN". Via dynamic data exchange HUGIN was connected to EXCEL. In close co-operation to the physicians we constructed an user-interface. A first internal validation showed a good functionality of the system, diagnoses are found with a high probability.
We plan to complement the prototype towards hypoglycemia and lactic acidosis in future. For this we want to give extern experts the possibility to include their knowledge. Secondly the facilities for decision support should be included into the system. A database will be constructed for saving and updating the knowledge and last but not least an extern evaluation will be necessary.

Sequencing and Analysis of the Sulfolobus solfataricus P2 Genome

Christoph W. Sensen
(principal investigators: Robert L. Charlebois, Mark A. Ragan, W. Ford Doolittle)
National Research Council
Institute for Marine Biosciences

The Sulfolobus genome project is the only all-Canadian genome sequencing project set up to determine the entire DNA sequence (3.1 Mbp) of an organism. The sequence of the crenarchaeote Sulfolobus solfataricus P2 is obtained on a cosmid by cosmid basis using automated sequencing techniques. The sequencing strategy allows us to analyze clean sequence in contigs at least 40 kbp long. The size of the cosmid inserts allows a good quality control of the sequence assembly. To date, about 25% of the Sulfolobus genome has been sequenced by the Canadian team. The sequence data produced by the genome project are used to develop tools for automated computer analysis of entire genomes in collaboration with computer scientists from various institutions. The most significant collaboration is with Dr. Terry Gaasterland from the Argonne National Laboratory. The goals of her research are presented in a separate abstract in this abstractbook.

Modeling Metabolic Pathways with Incomplete Information

Michael Mavrovouniotis
Chemical Engineering Department and
Council on Dynamic Systems and Control
Northwestern University

A set of techniques for the qualitative analysis of metabolic pathways are based on thermodynamic and kinetic limits.
In the first technique, thermodynamic bottlenecks are defined as those pathway sections that present thermodynamic obstacles to the flux. Distributed bottlenecks create obstacles through the combination of two or more marginally feasible reactions. An algorithm formally identifies all bottlenecks.
A second technique computes the necessary Gibbs energies for the thermodynamic analysis, through group contributions.
A third technique computes limits on enzymatic reaction kinetics, using limits from diffusion; thermodynamic constraints; and partial experimental data. This method can be used to interpolate or extrapolate kinetic measurements.

Sequence Interpretation and Metabolic Reconstruction in an Ongoing Sequencing Project

Terry Gaasterland
Argonne National Laboratory, Department of Computer Science
University of Chicago

This year, 1995, we experienced the first fall genome sequence. Within three years, we anticipate not only Haemophilus influenzae, Mycoplasma genitalium but Solfalobus solfataricus, Escherichia coli, Methanoccocus janaschii, Methanobacteriom thermautotrophicum, Mycobacterium legrae, Richettsia prowazechi, Helicobacter pylori, and Saccharomyces cerevisiae.
Analyzing these sequences with today's tools against today's databases with today's understanding of domain motifs would take 3 people 1 year per megabase. Over the next years, the tools will change, improve and grow in number. The sequence databases continue to expand - for both proteins and DNA.
The curation community pays close attention to serving tools and maintaining data. However, analysis of whole genomes has remained a human (super human!) task.
MAGPIE (Multipurpose Automated Genome Project Investigation Environment) provides sequencing projects with a data collection manager and a logic-programming based automated reasoning environment for maintaining. The emerging picture of a genome. Automated decisions

and organization lay on top of hierarchical navigable evidence for genome sequence features. Reconstructing the metabolism of an organism is an essential module of genome sequence interpretation. MAGPIE provides the basis for feeding the reconstruction - a list of enzymes together with confidence levels and phylogenetic distances. A reconstruction engine (joint with Ross Overbeek and Eugeni Selkov) uses the MAGPIE output together with EMP (Selkov's Enzyme and Metabolic Pathway database) to generate reconstructed primary metabolism as interconnected, possibly disjoint, topolares - ripe for further investigation.

An Object-Oriented Approach to the Modeling of Bacterial Metabolism

G. Breuel, A. Kremling, E.D. Gilles
Institut für Systemdynamik und Regelungstechnik
University of Stuttgart

One of the objectives of mathematical modeling and dynamical simulation of bacterial metabolism is the development of computational and theoretical techniques, which determine behavior patterns and dominant mechanisms of a biochemical system. Different parts of the metabolism are understood very well from a biological point of view.

A smaller number of mathematical models describing certain metabolic pathways or the expression of a specific operon can be found in literature. Considerably less attention has been given to the modeling of the whole metabolism as a collection of interacting subsystems. Some of the main reasons for this are, that the available knowledge is usually sparse, uncertain and often only qualitative.

To overcome these problems metabolic knowledgebases have been constructed by different research groups (e.g. ECOCYC). Besides these knowledge bases, novel modeling and simulation languages have been developed to support the modeling of systems in chemical engineering (e.g. MODEL.LA) and biology (e.g. METASIM).

These systems employ a decomposition and a structuring of the modeling knowledge for the specific application area.

They rely on concepts of object-oriented programming to facilitate the development, maintenance and reuse of mathematical models. Another object-oriented approach for the mathematical modeling of chemical engineering processes was developed by Marquardt, Gilles et. al. In comparison to other approaches, this concept contains a more thorough structuring of the modeling knowledge of the application area. In our approach this concept was modified and extended to cope with the domain specific issues of biological and biochemical systems. The most important step towards the development of a modular structured concept to the modeling

of bacterial metabolism is the structuring of the available biological and genetic knowledge. To model the dynamical behavior of bacterial metabolism, the metabolism is considered as two coupled and interacting networks. In a decomposition of the two network modeling objects are introduced. The introduced modeling objects only imply the structural aspects of metabolic processes. Behavioral modeling objects are introduced to describe the behavior of any structural modeling object (e.g. reaction kinetic).
Consequently, a behavioral modeling object is always directly linked to a structural modeling object. The ideas presented in this contribution can be seen as a first step towards the development of an object-oriented representation of the introduced modeling objects in a data model for the application area of bacterial metabolism.

The Structural Design of Glycolysis, Kinetic and Thermodynamic Constraints

Reinhart Heinrich
Department of Theoretical Biophysics
Humboldt - University Berlin

There exists a rather large number of models dealing with the simulation of the dynamics and the elucidation of control properties of glycolysis. In these investigations the stoichiometric coefficients which define the topology of enzymic networks and the kinetic constants of enzymes are considered as given parameters and it is not attempted to give any explanation for the observed values. In the present study it is theoretically analyzed whether the structural design of contemporary glycolysis may be explained on the basis of optimization principles originating from natural selection during evolution.
Particular attention is paid to the problem how the kinetic and thermodynamic properties of the glycolytic pathway are related to its stoichiometry with respect of the number and location of phosphorylation and dephosphorylation sites. Phosphorylation by inorganic phosphate as well as by ATP is taken into account. The mathematical analysis of an unbranched chain shows that the requirement of high ATP producing rate favours a structural design which includes not only ATP producing reactions (P sites) but also ATP consuming reactions (C sites). It is demonstrated that at fixed overall thermodynamic properties of a chain the ATP production rate may be enhanced by optimizing the location of the coupling sites as well as by kinetic optimization. The ATP production rate is increased if the C sites are concentrated at the beginning and all the P sites at the end of the pathway. A maximum is achieved in dependence on the number of coupling sites. It is analyzed how the optimal ATP production rate depends on the total number of steps of an energy converting pathway. In extended version of the model the effects of

internal feedback regulation, of variable enzyme concentrations and of a splitting of C6 compounds into two C3 compounds, which in glycolyis takes place at the aldolase reaction, are taken into account.
The theoretical results for optimal states which are in general agreement with the structural properties of contemporary glycolysis are compared with those for non optimized states. Of particular theoretical interest are hypothetical pathways owning an "antiglycolytic" design with a reverse location of ATP consuming and ATP producing reactions compared to glycolysis. Several combinatorical problems concerning the toal number of different stoichiometric designs are solved.

Introduction to Metabolic Control Theory

Daniel Kahn
Biologie Moléculaire des Relations Plants-Microorganisms
INRA - UNRS, Castanel-Blozan

Metabolic control theory is a mathematical formalism which allows the computation of systemic sensitivities (“control coefficients”) in terms of network structure and enzyme sensitivities (“elasticity coefficients”).
Complex systems can be approached by decomposing them into smaller modules, provided moiety conservation cycles are confined within the modules. Control of the entire system can be calculated in terms of the control within each module and interactions between the modules.
Modular decomposition of complex systems appear particularly relevant to the study of the regulation of gene expression and of regulatory cascades. However experimental methodology needs considerable improvement before we can reach a reliable quantitative description of regulatory networks.

Modern Developments in Metabolic Control Theory

Stefan Schuster
Department of Biology
Humboldt - University Berlin

The basic quantities of Metabolic Control Theory are presented. In particular, the two distinct definitions of control coefficients (in terms of derivatives with respect to enzyme concentrations

and in terms of derivatives with respect to reaction rates) are compared, and the interrelations to response coefficients are discussed. The recently introduced co-response coefficients are shown to be independent of the enzyme subject to perturbation under certain conditions. The summation theorems of Metabolic Control Theory are valid whenever the control coefficients are independent of the choice of perturbation parameter, which is satisfied under weak conditions. Cases where these conditions are not fulfilled include dynamic metabolite channelling and moiety-conserved cycles together with high enzyme concentrations. The problems occuring in these cases can be resolved by using control coefficients with respect to elemental steps of enzyme catalysis. These coefficients have also turned out useful for showing that there are no completely rate-limiting steps in enzymatic mechanisms. Other modern developments in Metabolic Control Theory are briefly summarized.

Metabolic Network Dynamics Analysis as Tool of Metabolic Engineering

Gregory Stephanopoulos
Department of Chemical Engineering
MIT

Understanding the dynamics of metabolic pathways is an important step of metabolic Engineering, the directed and purposeful modification of metabolic pathways for the overproduction of metabolites. We present two paradigms of pathway dynamics and metabolic engineering:
The first deals with the distribution of metabolic flux between competing pathways and factors affecting it. Using the tetrahydrodipicolineate brunch point of lysine biosynthesis as example, fluxes through the two competing pathways were analyzed in terms of H_4D concentration variations as well as at varying enzymatic activities. It was determined that in-vitro kinetic data of the competing enzymatic reactions can provide valuable insights on the outcome of flux distribution as well as means by which the latter can be affected. Conclusions drawn from this example were applied to the analysis of the aspartyl-semialdehyde, hormoscrine and threonine branchpoints leading to metabolic modifications for the successful overproduction of lysine, threonine and isoleucine, respectively.
The second paradigm addresses the issue of control distribution of complex metabolic networks. Concepts of Metabolic Control Analysis have been extended to groups of reactions and systematically applied to the analysis of the control structure of the aromatic aminoucid pathway. Furthermore, the response of the pathway was studied to the manipulation of one or more of the constituent bioreactions. It was found that, although rather limited results are possible with the

modification of a single reaction, significant application of the overall network flux is possible with the proper modification of a small (2-3) number of reactions. Group MCA permits the implementation of these findings in the laboratory.

Dynamic Mathematical Models in Knowledge-Based Systems

Bernhard Pohl
Department of Hygiene and Microbiology
University of Würzburg

Knowledge-based systems are computer programs which consist of a knowledge base and an (expert) system shell. The knowledge base is represented in a formal language, optimised for the specific problem, e.g. representation of physiological models or construction of text modules for output. The system shell comprehends more general functions as there are database, inference, explanation, and communication to the user. Knowledge base representation languages are often rule- or frame-based and then reflect algebraic equations. In this lecture it is argued that a modelling language for compartment (dynamic) systems should be included which are usually described in differential equations. Until now there is almost no knowledge based system shell which can deal with dynamic models - remember that an explanation function then is also required for dynamic models.
While algebraic equation knowledge uses forward- or backward-chaining inference in most cases, differential equation knowledge is applied for continuous simulation and parameter fitting algorithms. I expect that integration of dynamic models in knowledge based systems will greatly enhance development of knowledge based systems for medical applications.

Alignment Matrices for Modeling Complex Interactions between DNA and Regulatory Proteins

Gerald Z. Hertz, Gary D. Stormo
Department of MCD Biology
University of Colorado

Using log-likelihood statistics to compare sequence alignments, we have been able to determine alignments from multiple, unaligned, functionally related, DNA and protein sequences. The scoring formula we have used previously does not allow for insertions and deletions in the alignments. We have now used large-deviation statistics to extend the scoring formula to allow for insertions and deletions. The insertion-deletion penalty of this scoring scheme depends exclusively on the observed alignment rather than on previous observations or the user's

intuition. We also describe how to incorporate correlations between positions of the alignment and describe the close relationship between our formulas and hidden Markov models. Finally, we present results of applying this new scoring formula to align a set of E. coli promoter DNA sequences to derive complex alignment patterns that can potentially be used to identify promoters in the E. coli genome.

Reasoning from Experiments to Causal Models in Molecular Cell Biology

Jeremy Wertheimer
Artificial Intelligence Laboratory
MIT

I describe a system that represents and reasons about experiments and causal models in the domain of molecular cell biology. Given a description of an experiment, the system computes the ramifications of the experiment on the causal models of the relevant cellular mechanisms. The system can automatically generate causal diagram - of the type found in textbooks and survey articles - from descriptions of experimental data. These models can be used for mechanism-based retrieval of biological research articles. This research contributes a representation for experiments and causal models in molecular cell biology, and a set of inference methods for reasoning about these experiments and models.