Modeling and simulation of metabolic pathways, gene regulation and cell differentiation

October 22-27, 1995. International Conference and Research Center for Computer Science, Schloss Dagstuhl, Saarland, Germany

Ralf Hofestädt, Michael Mavrovouniotis, Julio Collado-Vides and Markus Löffler

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, molecular biologists and biochemists have amassed a huge number of data. The exponential growth in these biological data is reflected in the content of databases like GENBANK, SWISSPROT and PIR. 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 that was part of the year-round series of Dagstuhl-Seminars organised at the Schloss in Germany. The above four organisers attempt to give, based on the transactions at the conference, their perception of the 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 networks. The analysis of metabolic networks and eventual construction of novel pathways are the aims of the new research field of metabolic engineering. By definition, this field requires an integrative view of metabolism: in analysing a pathway, one may discover interactions among pathways with different physiological functions; in synthesising 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, which 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. Ideally, it would be desirable to have each metabolite (biochemical compound) appear as a single node in the chart, with all its reactions emanating from it. Given the large number of reactions, however, in which some metabolites participate (e.g. 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 participants pointed out that a computerised version of the metabolic map would be much easier to use. Indeed, computational metabolic databases with graphical displays 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 the stoichiometries of reactions, to enzyme cofactors, activators and inhibitors, to protein subunit composition and genetic maps 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 onto 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 engineer-

ing. 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 provided by analytical models based on the use of differential equations for simulation of reaction kinetics and metabolite concentrations (**H. Heinrich, D. Kahn** and **G. Stephanopoulos**). Such models predict the behavior patterns and dominant mechanisms of a biochemical system, and are able to decompose a complex pathway and, in some cases, can determine which pathway steps limit the overall flux towards a desired final product. This analysis is especially fruitful in those aspects of metabolism that 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. Reliance on such 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 using equation-oriented methods with integrated metabolic databases. This was illustrated by the object-oriented computational encoding of components (metabolites, reactions, cells and medium) of the biological dynamic system, presented by **G. Beuel**. 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 just describing 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**, using the theory of formal languages, automata and graph theoretical methods, developed a grammatical formalisation of biochemical reactions, which permits 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 predict limits on reaction rates. These 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.

Gene regulation

Deciphering coding regions is now a basic task associated with the computational interpretation of DNA sequences produced by genome projects. In addition, identification of signals in the DNA related to regulatory domains was also discussed. An example of an astute combination of statistics and biological insights was the discovery by **A. Danchin** of a specific base-tetramer found overrepresented at certain distance intervals and underrepresented at others. The overrepresentation corresponds to regions involved in the transition from anaerobic to aerobic growth conditions. This result emerged within the framework of statistical analysis of codon usage in *E. coli* that supported three main biological gene classes: high expression genes, low expression genes, and genes acquired by 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 implication 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 also taking into account the context in which 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 binding affinity 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 based on published experimental information. An interesting observation 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 their own 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 other dealing with a formalisation that incorporates gene activation at a distance, in sigma 54-recognising promoters. This approach is centred 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* networks 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 of the talks came from theory or computer science. An example was the experimental analysis of nitrogen regulation in bacteria, offered by **B. Magasanik**. 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 also briefly discussed 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 absent or inactivated.

Cell differentiation

The coordinated regulation of the expression of genes is primarily responsible for the diversity of differentiated cell phenotypes that unfolds during the development of a higher plant or animal. Differentiation is usually a consequence of the 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, whose activity triggers subsequent regulatory cascades. In these cases, the sequential expression of genes is genetically preprogrammed and the genes cannot usually be turned on out of sequence. Regulatory genes are known to be involved in the control of patterns of differentiation. In some cases regulatory *cis*-acting elements called enhancers and silencers modulate levels of gene expression from nearby promoters; 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** analysed 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 that can explain the spatial and temporal organisation of the intestinal crypt system, relating cellular division, cell differentiation and maturation to the 3-D architecture and formation of cellular clones. The models ranged from stochastic cellular automata, which describe the short term behavior of single intestinal crypts, to a differential equation model of all stages, which disregards the system architecture of crypts. **H. Meinhardt** presented elegant models which are quantitative but simple; they attempt to capture the essential biological behavior of pattern-forming systems.

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. Despite the large volume of available data, however, the information is often incomplete, uncertain and qualitative for any one system of interest. Ways to cope with incomplete information must be further investigated. Informal discussions at the conference revolved around many of the questions that 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, metabolic networks and 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.

Acknowledgement

The organisers thank the VW-Stiftung for its generous financial support for this conference.

Ralf Hofestädt is with the University of Leipzig, Department of Medical Informatics, D-4103 Leipzig, Germany; he is also affiliated with the University Koblenz-Landau, Department of Computer Science; **Michael L. Mavrovouniotis** is with the Department of Chemical Engineering and the Council for Dynamic Systems and Control, Northwestern University, Evanston, Illinois 60208-3120, USA; **Julio Collado-Vides** is with the Centre for Nitrogen Fixation, National Autonomous University of Mexico (UNAM), Cuernavaca, A.P. 565-A, Morelos, Mexico; and **Markus Löffler** is with the University of Leipzig, Department of Medical Informatics, D-4103 Leipzig, Germany.