

The **Physiome Project** (see <http://www.physiome.org/files/>)

The PHYSIOME is the quantitative description of the physiological dynamics or functions of the intact organism. The name comes from "physio-" (life) and "-ome" (as a whole).

The PHYSIOME PROJECT is an integrated multi-centric program to design, develop, implement, test and document, archive and disseminate quantitative information and integrative models of the functional behavior of organelles, cells, tissues, organs, and organisms. The long-range goal is to understand and describe the human organism, its physiology and pathophysiology, and to use this understanding in improving human health. but much or most of what must be learned will come from other species. The project aims toward providing models that summarize information on physiological systems, integrating the observations from many laboratories into quantitative, self-consistent, comprehensive descriptions. The goal is to provide to the community of scientists, physicians, teachers, and to medical health professional and industrial communities, functional descriptions of human biological systems in health and disease. A fundamental and major feature of the program is the databasing of the basic observations for retrieval and evaluation.

A network of Physiome Centers could comprise an adaptable international resource for databasing data on the functional aspects of biological systems covering the genome, molecular form and kinetics, cell biology, up to intact functioning organisms. These many databases would provide the raw information that might be integrated via physiological systems models, and should be structured hierarchically for accessibility and utility. The centers would maintain databases of information and models for retrieval over the Internet. The databases and models will have to accommodate data from many species.

Sample paper: "Metabolic bioinformatics, metabolic reconstructions and mathematical simulation of the cellular metabolism"

by Evgeni Selkov,¹ 2 Evgeni Selkov, Jr.,¹ Yuri Grechkin,³ Igor Goryanin,¹ Milyausha Galimova,¹ Yuri Komarov,¹ Niels Larsen,⁴ Natalia Maltsev,³ Natalia Mikhailova,² Valeri Nenashev,¹ Ross Overbeek,³ Lyudmila Pronevich¹

¹ Laboratory for Metabolic Bioinformatics and Mathematical Simulation, Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142292 Pushchino, Moscow region, Russia, ² Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, 9700 S. Cass Ave., MCS-221, IL 60439-4844, USA, ³ EMP Project team, 142292 Pushchino, Pushchino, Moscow region, Russia, ⁴ Center for Microbial Ecology, Michigan State University, Gilter Hall, East Lansing, MI 4884, USA

We propose to contribute the following components to the Physiome Project:

1. database on Enzymes and Metabolic Pathways, EMP,
2. metabolic and functional reconstructions from genome sequence data, and
3. mathematical simulation of the cellular metabolism.

The Enzymes and Metabolic Pathways Database (EMP) is a general-purpose database [1-3] whose records cover all the main aspects of enzymology and metabolism: enzyme purification and characterization, kinetics, reaction mechanisms, thermodynamics, immunoreactivity, metabolism, regulation, and so forth. It represents all classified enzymes and includes records for some 1,500 organisms. Nearly one-third of the database records specify the enzymology and metabolism of humans and experimental animals. The information contained in EMP complements other widely used data banks on primary, secondary, and tertiary structures of enzymes, non-catalytic proteins and their genes.

The information stored in EMP comes from original articles published in leading international journals; each article is evaluated and translated into an EMP record by an expert annotator. The information is stored primarily as flat ASCII files that can be loaded into different database systems. In 1995, EMP 1.0 was released as a searchable database compatible with the Microsoft Windows platforms and distributed on CD-ROM by Biological Databases, Inc. [1]. In September 1997, EMP contained over 20,000 records compiled from about 14,000 original publications. By the end of 1997, EMP is expected to be accessible via the World Wide Web.

The Metabolic Pathways Database (MPW). Initially a small part of EMP, MPW is the most comprehensive collection of metabolic pathways in computer readable form [4]. The pathway collection, now including some

2,800 diagrams, covers not only primary and secondary metabolisms of the three forms of life, Archaea, Bacteria, and Eukarya, but also membrane transport, signal transduction pathways, intracellular traffic, translation, and transcription. The pathway diagrams were originally encoded as a formatted ASCII text, using pseudographics characters to draw arrows and boxes. A new release of MPW being developed now [5] is due to appear by October this year. In this release, the encoding is based on the logical structure of the pathways and represented by the objects commonly used in design and simulation of electronic circuits. This allows researchers to directly employ the wealth of knowledge and experience accumulated in this area of engineering. In particular, it makes possible automation of the basic simulation operations such as deriving stoichiometric matrices, rate laws, and, ultimately, dynamic models of metabolic pathways.

Metabolic and functional reconstructions from sequenced genomes. With the beginning of complete genome sequencing in 1995, MPW started to play a critical role in the metabolic reconstructions of sequenced genomes. By the term metabolic reconstruction we mean the process of inferring the metabolism and functional organization of an organism from its genetic sequence data supplemented by known biochemical and phenotypic data. Such reconstructions have recently been made for a number of available sequenced genomes. The reconstructions are accessible via the PUMA, WIT, and WIT2 systems, developed by Ross Overbeek and his colleagues [6-8].

Each reconstruction tends to be a consistent conceptual model in which all metabolic pathways are connected to sequence data stored in public databases. Here, consistent means that all intermediates of the metabolism are balanced by sources and sinks.

Initial metabolic reconstructions for a virtual human cell have been attempted under the PUMA [9] and KEGG [10] systems. These reconstructions are based on the known biochemistry [9] and available sequence data [9, 10]. To further extend this effort toward highly differentiated human cells, one needs EST (Expressed Sequence Tag) data to determine which part of the huge human genome is expressed in each specific tissue cell type. We believe that such reconstructions for a limited number of human tissue cells, such as those of smooth and skeletal muscles, liver, kidney, and blood, must be done first to form a solid basis for mathematical simulation of these tissues, organs and then the whole system.

Mathematical simulation of cellular metabolism. Quantitative simulation is the next step after genome sequencing and metabolic reconstruction. It is often argued that any serious quantitative simulation of cellular metabolism is not possible because it requires an enormous amount of quantitative data on kinetic parameters, concentration of intermediates, etc. Moreover, there is a severe limitation of the current reconstruction methodology based on sequence data: its inability to predict regulatory mechanisms and numerical values of all the kinetic parameters required for quantitative simulation and analysis of the metabolism.

We have begun to develop an approach that allows one to bypass this problem. This approach exploits mathematical simulation itself to predict missing control mechanisms and parameter values. The main idea is to optimize the structure and parameters of mathematical models to the known functions of simulated metabolic systems. Such optimization is a generalization of the parameter fitting widely used in simulation studies.

To realize this approach, we have developed a software package, DBsolve [11]. The package includes a large collection of encoded enzymatic reaction mechanisms and metabolic pathways that are used to build up complex metabolic models automatically. Specifically, DBsolve derives ordinary differential equation sets, ODE, from stoichiometric matrices of the enzyme reactions and metabolic pathways stored in EMP. For parameter estimation and model fitting, the software package has an optimization block. It uses a modification of the unconstrained zero-order minimization polygon method of [12]. The bifurcation analysis block is based on a package [13] translated into C++. DBsolve uses a version of Gear's method for integration of a stiff ODE with the controlled stiffness, thereby enabling the solution of hybrid ODE systems with algebraic subsets. An original parameter continuation algorithm is used for computing the algebraic system solution along a parameter range.

The importance of unicellular organisms for the Physiome Project should not be underestimated. Many bacterial metabolic systems are very similar to their mammalian counterparts and can be treated as much simpler experimental models. It has recently been discovered that the cyanobacteria *Synechocystis* and *Synechococcus* have a circadian cell clock that gates cell division [14-16]. We have reconstructed the

metabolism of *Synechocystis* from its sequenced genome [7, 8] and found that the main postulates of a metabolic theory of the clock, developed earlier [17-19], hold true for this cyanobacterium. Thus, this bacterium lends us the shortest way to crack this enigma of cellular physiology and of human physiology in particular.

Acknowledgment. This work was supported by U.S. Department of Energy, under U.S. Department of Energy, under Contract W-31-109-Eng-38, and award OR00033-97CIS001.

References

1. <http://www.biobase.com/EMP>
2. Selkov E., Basmanova S., Gaasterland T., Goryanin I., Gretchkin Y., Maltsev N., Nenashev V., Overbeek R., Panyushkina E., Pronevitch L., Selkov E., Jr., Yunus I. *Nucleic Acids Res.*, 1996, 24 (1), 26-29
3. <http://www.biobase.com/emphome.html/homepage.html/pags/pathways.html>
4. Selkov E., Galimova M., Goryanin I., Gretchkin Y., Ivanova N., Komarov Y., Maltsev N., Mikhailova N., Nenashev V., Overbeek R., Panyushkina E., Pronevitch L., Selkov E., Jr. *Nucleic Acids Res.*, 1997, 25 (1), 37-38
5. <http://beauty.isdn.mcs.anl.gov/~selkovjr/tmp>
6. http://www.mcs.anl.gov/home/compbio/PUMA/Production/puma_graphics.html
7. <http://www.mcs.anl.gov/home/compbio/WIT/wit.html>
8. <http://www.mcs.anl.gov/home/overbeek/WIT2/CGI/user.cgi>
9. <http://www.mcs.anl.gov/home/compbio/PUMA/Production/ReconstructedMetabolism/reconstruction.html>
10. <http://www.genome.ad.jp/kegg/kegg3.html>
11. <http://sirius.iteb.serpukhov.su/~igor>
12. Swann, W. H. In: *Numerical Methods for Unconstrained Optimization* (W. Murray, ed.), London, Academic Press, 1972, pp. 13-28
13. Khibnik, A., Kuznetsov, Y., Levitin, V., and Nikolaev, E. *Physica D*, 1993, 62, 360-370
14. Kondo T., Strayer C. A., Kulkarni R. D., Taylor W., Ishiura M., Golden S. S., Johnson C. H. *Proc. Natl. Acad. Sci. U.S.A.* 1993, 90(12), 5672-5676
15. Aoki S., Kondo T., Ishiura M. *J. Bacteriol.*, 1995, 177(19), 5606-5611
16. Mori T., Binder B., Johnson C. H. *Proc. Natl. Acad. Sci. U.S.A.*, 1996, 93(19), 10183-10188
17. Sel'kov E. E. *Ber. Bunsenges. Phys. Chem.*, 1980, 84, 399-402
18. Avseenko N. V., Lisnichuk L. Ya., Sel'kov E. E. *Biofizika*, 1987, 32(2), 248-252
19. Selkov E. E., Lisnichuk L. Ya., Avseenko N. V. *Biofizika* 1989, 34(3), 459-456

Chapter 1

Accelerated Biological Simulation Research (ABSR)

Societal Needs for Accelerated Simulation Research in Biology

The pace of extraordinary advances in molecular biology has accelerated in the past decade due in large part to discoveries coming from genome projects on human and model organisms. The advances in the genome project so far, happening well ahead of schedule and under budget, have exceeded any dreams by its protagonists, let alone formal expectations. Biologists expect the next phase of the genome project to be even more startling in terms of dramatic breakthroughs in our understanding of human biology, the biology of health and of disease.

Why can we not wait for a computational science effort until the conclusion of the genome project, and what are the broader implications? Significant personal and economic costs will be borne by society for delays in our exploiting the discoveries of the genome projects on behalf of the Nation. The importance of each individual, each human being on the planet, the individual's productivity and contribution to society, the quality of that individual's life including that of our shared environment, has transcendental value. A major breakthrough in technology, the human genome project, puts us on the brink of truly extraordinary progress in realizing the transcendental goal of human well-being. Only today can biologists begin to envision the necessary experimental, computational and theoretical steps necessary to exploit genome sequence information for its medical impact, its contribution to biotechnology and economic competitiveness, and its ultimate contribution to environmental quality.

Individualized medicine- the recognition of individual differences in drug and treatment response, in disease development and progression, in the appearance and thus diagnosis of disease-justifies a much more aggressive approach to utilizing DNA sequence information and the introduction of simulation capabilities to extract the implicit information contained in the human genome. Each year thousands die and a hundred times more people suffer adverse reactions of various extents to drugs that are applied in perfectly correct usage, dose and disease specificity. Cancer in particular stands out as a disease of the genes. Two patients with what appear to be identical cancers (based on cellular pathology) at the same stage of malignancy, growth and dissemination will have very different responses to the same therapy regime; for example, one could rapidly fail to respond to treatment, become terminal in weeks, and the other could respond immediately and ultimately recover fully.

Re-engineering microbes for bioremediation will depend directly on the understanding derived from the simulation studies proposed here. The design of new macromolecules and the redesign of microbe metabolism based on simulation studies, truly grand challenges for applied biology, will contribute to a wide range of environmental missions for the Department, including the environmental bioremediation of the Nation's most contaminated sites, often mixed waste for which no economically feasible cleanup technology currently exists. Many aspects of sustainable development, changing the nature of industrial processes to use environmentally friendly processes, depend on the same kinds of advances.

Similarly, in the longer term the understanding of living systems will also contribute to environmental research on carbon management.

To exploit the inherent genome information derived from knowing the DNA sequences, computational advances, along with related experimental biotechnology, are essential. Knowing the sequence of the DNA does not tell us about the function of the genes, specifically the actions of their protein products - where, when, why, how the proteins act is the essence of biological knowledge required. Encoded in the DNA sequence is a protein's three dimensional topography, which in turn determines function; uncovering this sequence-structure-function relationship is the core goal of modern structural biology today. The goal of the accelerated computational biology initiative is to link sequence, structure and function, and to move from analysis of individual macromolecules to macromolecular assemblies and complex oligomeric interactions that make up the complex processes within the cell.

Elucidating the Path From Protein Structure, to Function, to Disease

A particular sequence derived from the genome encodes in a character string the three-dimensional structure of a protein that performs a specific function in the cell. The primary outcome of the Argonne Workshop on Structural Genomics emphasized the goal of exploiting the relationship between sequence and structure at a level unattainable before the coordinated effort to map and sequence the genomes of many organisms. It is in fact a logical extension of the genome effort to systematically elaborate DNA sequences into full three dimensional structures and functional analysis; this new effort will require the same level of cooperation and

collaboration among scientists as was necessary in the original genome project.

The Structural Genomics Initiative is gaining momentum and will primarily focus initially on the infrastructure and support necessary to realize high throughput experimental structures by x-ray crystallography and high field NMR. Success of this effort is in some sense guaranteed, and the computational challenges that are posed when a structural genomics initiative of this scale is contemplated is already clear cut. It is the outgrowth from this direction, a computational biotechnology initiative-that will capture everything from sequence, structure and functional genomics to genetic networks and metabolic engineering to forward folding of macromolecules and kinetics to cellular level interactions. This approach makes all the modeling and simulation of macromolecules fair game and moves toward the complex systems approach, the complicated level of real living systems rather than individual macromolecules.

Elucidating the path from structure, to function, to epidemiological consequences for proteins of selected pathogens directly relates structural analysis to national health. As sequence information continues to accumulate at a rapid pace from the Genome Program, and as structure information becomes increasingly available from the Structural Genomics Initiative, the challenge will be to integrate biological information from the molecular level of sequence and structure, to the macroscopic level of function. Such an integrated approach lies at the heart of our ability to understand and combat disease. Since the early days of the Genome Project DOE has been a pioneer in establishing databases of molecular information and in developing algorithms and computational tools to analyze this information. In the

past, this effort has concentrated on the human genome, with the goal of understanding basic molecular mechanisms central to biological function. Establishing such an understanding is critical to developing a molecular interpretation of disease, and augmenting these database with structural information from the Structural Genomics Initiative will prove immensely valuable.

An Accelerated Biological Simulation Research Initiative

The goal of the ABSR effort within the time horizon of the Strategic Simulation Initiative is to:

- *Characterize the link between protein sequence and fold topology.* The emphasis on the experimental determination of a complete set of representative tertiary folds is based on the success of comparative modeling. One can often deduce the fold topology for a new sequence by simply finding another similar sequence with a known structure, even when sequence identity is very low. The ~100,000 protein coding genes expected in the human genome is too large to handle experimentally, but protein modelers estimate that ~10,000 structures for protein domains would be sufficient to use these algorithms to determine the fold topology of the remaining 90% of gene products. Algorithms of scale have been developed that attempt to find this sequence-fold match, but more sophisticated versions that exhibit more severe scaling are needed to be genuinely successful in this regard.

- *Quantitative determination of protein structure from folding or conformational searches.* Although the gross backbone configuration (or tertiary fold) of proteins remains invariant under sequence mutation, quantitative

differences in structure between wild type and mutant can have important macroscopic effects on protein function. The quantitative prediction of structure and folding behavior is critical for the development of successful pharmaceutical drug targets, protein redesign of enzymes for bioremediation, and prediction of disease manifestation of new pathogens. The next step beyond predicting fold topology is the quantitative determination of protein structure starting first from the tertiary fold prediction, and ultimately directly from sequence. This is the heart of computational complexity in molecular biology at present. The simulation methodologies have commonality with the areas of materials and combustion chemistry, and share the same severe scaling issues due to large length scales, long time scales, and system size scaling that define the need for high-end computing in quantitative modeling.

- *Simulate the biochemical function of individual gene products.* A robust and predictive approach to protein structure ties directly into our ability to model and understand protein function. Simulations will be important for predicting detailed structural changes and the fluctuations that drive enzymatic reactions, how protein structures recognize each other, and associate to form the multi-protein complexes, and prediction of the mode and energies with which molecules bind to proteins in metabolic reactions and molecular signaling processes. The strong biological connection between structure and function means that the same modeling issues, primarily a good energy surface description and the means to explore it, will be important for simulating biochemical function as well.

From “Beyond Genome 2003” – a conference held in San Diego this June

Advances in genomics and proteomics have increased our understanding of biological systems at the molecular level. Systems Biology modeling can elucidate how individual system components interact, integrate, and function to form a complex organism. Examples of computational modeling, in conjunction with empirical research, are providing a greater understanding of disease states and target prioritization. This combination may increase the probability of successful drug development. This conference will bring together experts from life and computational sciences to discuss the state of the field. The conference was intended for those researchers interested in Systems Biology tools, methods, and concepts to develop or improve drug development strategies.

Novel Clone Resources and Assay Technologies for Functional Studies; John Carrino, Ph.D., Vice President, Research and Development, Invitrogen Corporation. The presentation will focus on the development and characterization of a flexible clone resource for use in various applications directed toward determination of protein structure and function. The varied, fluorescence-based assay technologies in combination with the clone resource provide a powerful platform for the development of biochemical and cell-based assays for functional analysis to support both basic science and HTP drug discovery efforts.

An Information Perspective, Dr. Michael Hucka, Staff Scientist, California Institute of Technology: The Systems Biology Mark-up Language (SBML) Level 2. SBML is becoming a de facto standard in the modeling community. It provides a common, application-neutral XML format for representing models of biological function described in terms of networks of biochemical reactions. SBML Level 2 is the latest edition of the SBML specification.

Tools for Large-Scale Distributed Simulations in Systems Biology: Dr. John Ambrosiano, Research Scientist, Computational Biosystems, Computer and Computational Sciences Division, Advanced Computing Laboratory, Los Alamos National Laboratory: Systems biology views biological processes as complex dynamical systems. It offers the promise that by integrating experiment, theory, and simulation we can one day disentangle the daunting complexities of molecular and cellular biology to produce quantitative predictive models, which would allow, for example, new drug therapies to be designed rather than to be discovered by trial and error. Complex systems analysis often relies heavily on simulation as a virtual laboratory in which to explore the implications of the many abstract systems models that are continually offered by theorists to explain the processes of life. Our collaborations with theoretical and experimental biologists to build simulation tools in support of systems biology have given us some insights into the special nature of these problems. They have also helped us to appreciate the challenges simulation scientists face solving them. This talk will describe some of these challenges and the strategies we have undertaken to overcome them, within the context of an open-source software development project we call BioReactor.

Informatics for Systems Biology Knowledge Integration: Dr. Bret Peterson, Project Officer, National Center for Research Resources, National Institutes of Health. Translational research that moves discoveries from the lab bench to the clinic will increasingly draw on information collected across a wide range of disciplines. Efforts in biomedical informatics (e.g., the BIRN, <http://www.nbirn.net>) are aimed at creating technologies that allow for integration of information across disciplines to form coherent knowledge bases. Such knowledge bases will support queries that will provide the basis for building systems biology models that encapsulate current understanding and inspire new clinical practices. These efforts at integration benefit from work towards data standards and shared vocabularies, but they also assume that adoption of these practices will be imperfect, particularly across disciplines. Consequently, the informatics approaches also encompass computer science research aimed at merging heterogeneous data sources by mapping semantics across ontologies.

Systems Biology from an Information Technology Perspective: Linking Diverse Biological Data Sources: Dr. Ronald Shymko, Head of Scientific Computing, Novo Nordisk A/S. A pharmaceutical research organization can gain a competitive edge through its ability to manage and integrate multiple, large data sources. Technical difficulties need to be overcome, for example, linking different types of information content and different data formats. A greater challenge is to create a common conceptual framework within which to develop IT tools that can be applied across the data sources. Finally, it is important to provide practical guidelines and examples to formulate queries that give real benefit to the drug development process. Our experiences in these areas will be discussed, along with practical solutions to some of the problems.

Control Systems Analysis of Signaling Pathways: Dr. Herbert Sauro, Assistant Professor of Computational Biology, Keck Graduate Institute of Applied Life Sciences. This talk will discuss network motifs in signaling pathways in terms of classical and metabolic control theory. A review of the versatility of the humble cascade cycle, from digital integral to

operational amplifier control systems, will be given. An understanding of what the signaling networks are actually doing computationally will serve to guide the drug targeting community in much more rational ways.

Biological Complexity and Robustness: Dr. John Doyle, Professor of Control and Dynamical Systems, BioEngineering & Electrical Engineering, California Institute of Technology. Biological organisms are highly constrained in that they have not just evolved but have necessarily evolved in ways that are robust to uncertainties in their environment and their component parts. These are extremely severe constraints, not present in other sciences but essential in engineering, that emerge primarily at the network level in both biology and engineering. This talk will outline a promising new theoretical and computational infrastructure that explicitly exploits the highly structured, evolved, and "robust yet fragile" nature of biological systems and argue that this is necessary to avoid being overwhelmed by their sheer complexity.

Playing Practical Games with Bacteria and Viruses: Dr. Adam Arkin, Assistant Professor, Bioengineering, University of California, Berkeley; Assistant Investigator, Howard Hughes Medical Institute; Member, Biophysical Sciences Division, Lawrence Berkeley National Laboratory. How do pathogenic bacteria sense their environment to deploy different survival strategies? Why do some viruses, like HIV, allow their host to live for long periods whereas others like Ebola do not? How precisely are these strategies encoded in the organism's biochemistry and genetics and how closely do they need to be followed to guarantee its survival? What are the optimal strategies for defeating these organisms or forcing them to do our bidding for industrial or medical benefit? Here I will demonstrate, using examples from our research on *Bacillus subtilis* stress response and the design of HIV gene therapeutic strategies, how molecular biology combined with methods from statistical physics, nonlinear dynamics, and game theory can be used to pose and partially answer these questions as well as illustrate some of the profound challenges in doing so.

Exploiting Vertebrate Sequence for Insights into Human Biology: Dr. Len Pennacchio, Staff Scientist, Life Sciences Division, Lawrence Berkeley National Laboratory. The recent explosion of genomic sequence availability from higher eukaryotes has sparked new strategies to uncover functional regions of the human genome. Specifically, the wealth of sequence data has provided the means for comparing the genomic sequence from numerous vertebrate species. This modern strategy is based on the search for evolutionary conserved sequences and the hypothesis that highly conserved sequences are functionally important. The new paradigm of using genomic sequence data to select regions of the human genome for functional studies has provided a vast data set ripe for targeted analysis. We will describe our successful use of comparative genomic approaches to reveal functional regions of the human genome.

Statistical Analysis of T-Cell Gene Expression Time Courses: Dr. Adrienne James, Postdoctoral Research Scientist, Computational Biology and Chemistry Group, Novartis Research Institute. DNA microarrays have received a great deal of attention in recent years. However, such experiments commonly have small numbers of replicates making data analysis hard to interpret. Compounding these difficulties, issues such as probe set sequence and annotation quality are often not addressed. An exploratory analysis of several microarray time series experiments using Affymetrix GeneChip® technology is undertaken. A qualitative feature extraction pattern string representing these time series is generated by identifying statistically significant changes in signal levels. A clustering of identical and related pattern strings is performed. This analysis is used to identify groups of genes of possible immunological importance. The reproducibility and relevance of these results to pharma research is discussed.

Functional Implications of Alternative Designs for Genetic Switches and Signal Transduction Modules: Dr. Michael Savageau, Professor and Chair, Department of Microbiology and Immunology, University of Michigan Medical School. Biological systems with alternative designs can realize functions that appear to be similar. However, careful analysis often reveals functional differences and design principles that distinguish between alternatives. The integrated behavior of alternative designs can be elucidated by the use of concepts and methods from biochemical systems theory. The method of mathematically controlled comparison reveals system design principles by eliminating extraneous differences between the alternatives being compared. Two examples will be considered here: genetic switches that exhibit either continuous graded or discontinuous all-or-none dynamics and two-component signal transduction modules that either integrate cross-talk from a physiologically relevant signal or suppress cross-talk from irrelevant noise.

Genetic and Functional Approaches to Delineate Signal Transduction Pathways, Especially in the PI3K and NFkB Pathways: Dr. Tak Mak, Professor, Department of Medical Biophysics, University of Toronto. The main interest in my laboratory is to understand genetic and biochemical pathways that affect cell survival and apoptosis. To this end, we use multiple technologies to dissect these pathways. We concentrate on the two signaling cascades that lead to cell survival, namely, those leading to the activation of the transcription factors NFkB and those affected by the tumor suppressor gene PTEN. To advance our knowledge, we use microarray gene profiling, genetic modifying screens, and expression cloning to identify new genes and document relationships between known genes in these pathways. Biochemical studies are also employed to establish these genetic links. Finally, we generate mutant mouse models to study the physiological roles of

these genes. A recent example of our findings in the NF κ B pathways includes the identification of the role of BCL10 as a gene involved in antigen receptor signaling for the activation of these transcription factors. For the PTEN pathways, we demonstrated that DJ-1 may impact on the function of this tumor suppressor gene by regulating the stability of the RNA of this phosphatase. We hope that our findings will enrich the knowledge of these physiologically important pathways and allow the identification of mechanistically based drug targets for the treatment of autoimmune disorders and malignancies.

Determining Efficacy of Targets and Lead Compounds via a Data-Driven Computer Simulation of Cancer: Mr. Colin Hill, Chief Executive Officer, President, Chairman, and Founder, Gene Network Sciences, Inc. Using the Diagrammatic Cell Language™, GNS has created the largest known network of interconnected signal transduction pathways and gene expression networks controlling human cell proliferation and apoptosis. Time-course experiments measuring mRNA abundance and protein activity are conducted on Caco-2 and HCT116 colon cell lines to constrain unknown regulatory interactions and kinetic parameters via BioMetrics™ sensitivity analysis and the Digital Cell™ parameter optimization methods. Using the cell simulation, GNS tested efficacy and toxicity of various drug targets and their corresponding lead compounds, many of which are in early to late clinical trials. While many of the targeted therapies led to increased cancer cell apoptosis, GNS cell simulation revealed that a secondary agent was often necessary to induce a strong apoptotic effect. The simulation is also used to predict efficacy of various therapeutic combinations on cancer cells with specific mutational profiles. Thus once a promising therapy is found, the simulation can be used to determine the types of cancers it would impact maximally and under what conditions. The simulation becomes a powerful tool that can incorporate patient-specific data on the DNA, RNA, and protein levels for assessing efficacy of cancer therapeutics in specific patient populations and can greatly impact success of a given therapeutic strategy.

Tools for Generating, Managing and Mining Research Data: Systems Biology in Drug Discovery, Dr. Bernhard O. Palsson, Co-Founder and Chairman, Genomatica; and Professor of Bioengineering and Adjunct Professor of Medicine, University of California, San Diego

The Silicon Cell and Network-Based Drug Design: Prof. J. L. Snoep, Department of Biochemistry, University of Stellenbosch. One can now make computer replicas of important pathways in living cells and use these silicon cells to find new drug targets, to discover unknown gene functions, and to direct metabolic engineering.

Discovery of Disease Markers Using a Systems Biology Platform: Dr. Stephen Naylor, Chief Technology Officer, Beyond Genomics, Inc. Integration of gene expression, proteomics, and metabolomics technologies is central to the understanding of often disjointed data in complex multifactorial diseases. An example of such an integrative "systems biology" platform using progressive analytical and in silico approaches will be presented. Specifically, we focus on approaches for marker discovery and contextualization of key components of pathophysiologic processes. Data flow through our platform will be described, starting with experiment design approaches and analytical technologies that allow us to search for prognostic, diagnostic, and progression markers. We will introduce statistical and pattern recognition techniques that are used in combination with bioinformatics tools for optimal marker selection. A similar approach is also used for identification of key components and mechanisms of the disease. Our platform will be illustrated using two recent studies on atherosclerosis and Alzheimer's Disease transgenic mouse models.

In Silico Microbe Models and the Drug Discovery Process: Dr. Bernhard O. Palsson. Genome-scale in silico models have been built for several microbial cells. These models can predict a range of phenotypic functions. We explore their uses for guiding the drug discovery and development process.

A Systems Biology Approach to Understanding Cellular Functions: Dr. Hiroaki Kitano, Project Director, ERATO Kitano Symbiotic Systems Project, Japan Science and Technology Corporation; and Senior Researcher, Sony Computer Science Laboratories, Inc. A systems-level understanding of biological systems requires a set of principles and methodologies that links the behaviors of molecules to system characteristics and functions. Ultimately, cells, organisms, and human beings will be described and understood at the systems level grounded on a consistent framework of knowledge that is underpinned by the basic principles of physics.

Biology to Improve Decision Making in Drug Discovery: Dr. Didier Scherrer, Senior Scientist, Entelos, Inc. Using today's technologies, target validation and compound selection continue to be major challenges in the discovery process. Will eBiology really offer a novel paradigm in drug discovery? It is about mathematical modeling, statistics and decision theory applied in a systematic way to improve knowledge management and prediction of efficacy in the clinic. Integration of existing knowledge into a single contextual framework of human physiology, creates an environment for rapidly evaluating a target and focusing laboratory experiments to validate the target and screen drug candidates efficiently.

Genomics as the Foundation for the Discovery of New Therapeutic Products: From Theory to Practice: Dr. Vivian Albert, Vice President, Preclinical R & D, Human Genome Sciences, Inc. Ten novel therapeutic agents have entered clinical trials based on genomic discoveries. Nine of these have been based on discoveries made by Human Genome Sciences. Lessons learned from translation of theory to practice in the discovery of new therapeutic products based on genomics will be presented.

Applying the Entelos Metabolism PhysioLab to Pharmaceutical R&D: Dr. Xiaobin Zhang, Biosystems Engineer, Medical Informatics, Johnson & Johnson Pharmaceutical Research & Development, LLC. JJPRD is using the Metabolism PhysioLab from Entelos to evaluate drug targets and compounds from discovery through to development. This systems model provides a comprehensive and up-to-date knowledge base on the physiology of human metabolism and allows us to simulate experiments and trials of novel interventions and combination therapies. Having such a model facilitates discussions with experts from diverse backgrounds and has helped uncover ideas that contravene accepted dogmas. Examples and insights from our work will be presented.

A Systems Biology Approach for Ab Initio Hepatotoxicity Based on Global Transcription Profiles: Dr. Xuding Dai, Senior Data Analyst, Merck & Co., Inc. We will demonstrate the combination power of global transcriptional profiling and machine learning with real examples of drug toxicity prediction. Different from what has been reported in the field of toxicogenomics, it is the first time that the accuracy and generality of a hepatotoxicity predicting model are validated. The method we will demonstrate will not only have direct applications in drug toxicity prediction but also may provide a stepping stone for the modeling of complicated biological systems.

Objective Construction of Biological Pathways: Integrating Prior Knowledge and Bayesian Analysis of Gene Expression Data: Dr. Dean Bottino, Physiome Sciences, Inc. "Standard" or "defined" biological pathways, as might be found in any biochemistry or cell physiology textbook, or from various online resources, are an important source of information, representing currently understood relationships (e.g., reactions and interactions) between biological factors. However, gene or protein expression data often suggest previously unknown relationships between (sometimes unknown) biological entities. This can make it difficult to analyze results if it leads to either the development of complex visualization tools that lose the biological pathway context of the system, or to the combination of standard pathways into unwieldy "superpathways." I will describe a system that combines known biological relationships with relationships inferred by Bayesian analysis of gene expression data to construct a biological pathway that objectively includes the key elements as determined by the data.

The Use of Well-Defined in Vitro Models and Gene Expression Profiling to Identify Novel Therapeutic Targets: Dr. Mary E. Gerritsen, Senior Director, Vascular Biology, Millennium Pharmaceuticals, Inc.

In Silico Approaches to Predicting Drug Metabolism, Toxicology, and Beyond: Dr. Sean Ekins, Associate Director, Concurrent Pharmaceuticals. Discovery and optimization of new drug candidates are becoming increasingly reliant upon the combination of experimental and computational approaches related to drug metabolism, toxicology, and general biopharmaceutical properties. With the considerable output of high-throughput assays for cytochrome P450-mediated drug-drug interactions, metabolic stability, and assays for toxicology, we have orders of magnitude more data that will facilitate model building.

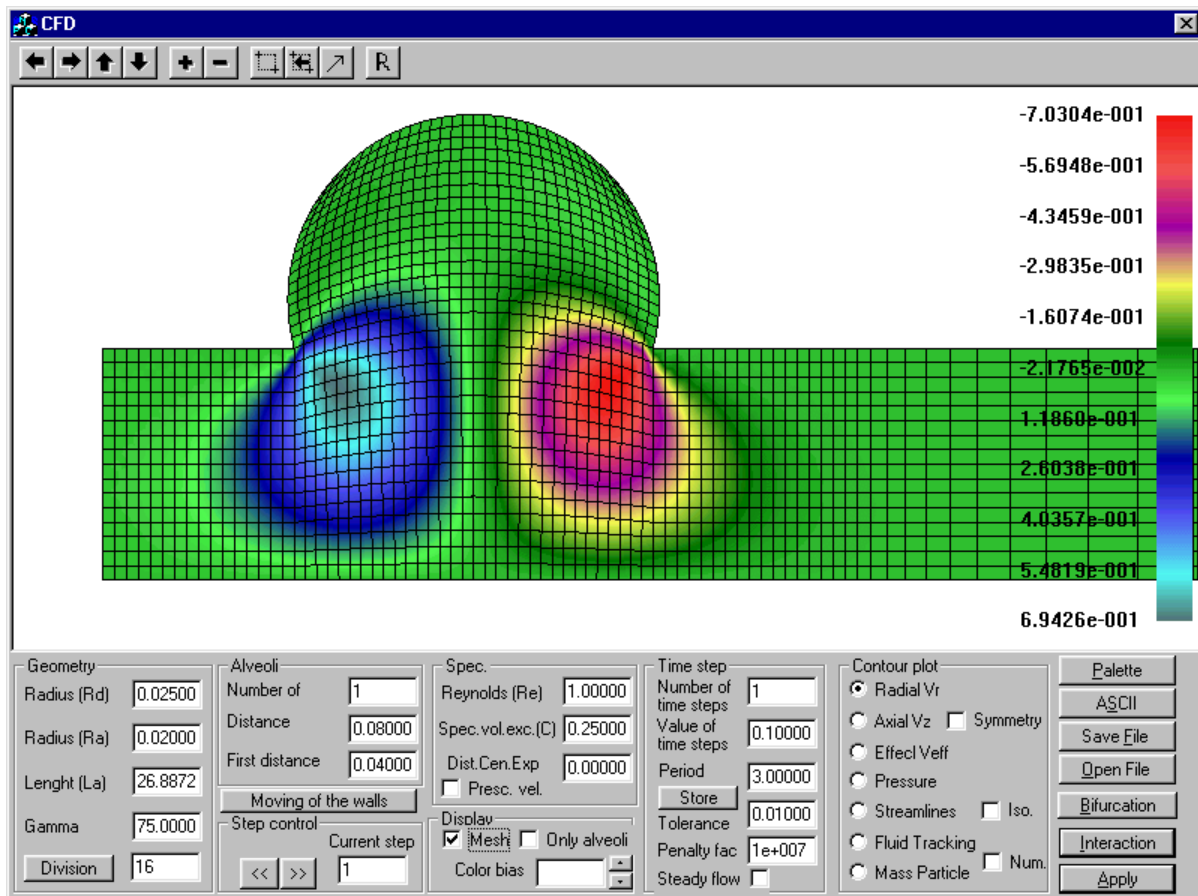
From Gene Networks to Drug Targets: Dr. Jian Zhu, Senior Scientist, Novartis Institute of Biomedical Research, Novartis AG. We have developed a system for high-throughput cell-based assays. The system was used to investigate genes' function and their roles in signaling pathways. Combining gene function information with results from gene expression analysis, regulatory sequence analysis, and literature mining, we were able to identify gene networks consisting of potential drug targets. A database and a web interface have been developed to enable efficient data mining. The system has become an important tool for target identification and prioritization.

Modeling air flow through the alveola

CFDAL is special software to model the air flow within the alveola, including particle tracking, and was developed in collaboration with the Harvard School of Public Health. The software also provides information about the velocity and pressure fields, particle tracking and streamlines (with detection of the stagnation point.) CFDAL represents an innovative software environment for the alveolated duct flow in rhythmically expanding pulmonary acinus. The complete object-oriented method is implemented and the program runs on the Windows platform.

Computational fluid dynamics (CFD) is essential for simulating the human physiological flows. Non-engineering professionals, especially medical doctors or biomedical researchers, will find general-purpose software for CFD analysis a complicated affair. However, CFDAL provides automation, eliminates user-dependent mistakes, drastically reduces time to generate the computer model, so that there is no need for the user to know the details of pre-processing procedures for CFD computer analysis.

Availability of the software for transfer data in other Windows applications is supported automatically. The OpenGL inside C++ language is used for building the graphical user interface. CFDAL is developed for the simulation of tidal breathing by an automatic generation of a realistic geometric model of an alveolated duct that expands and contracts. Acinar fluid mechanics plays an important role in explaining aerosol deposition, chaotic mixing and other phenomena in the lung periphery. CFDAL can be used as a tool to optimize the design of devices in the aerosol industry



Recreating Biopathway Databases towards Modeling and Simulation

Masao Nagasaki ⁽¹⁾, Atsushi Doi ⁽²⁾, Hiroshi Matsuno ⁽²⁾, and Satoru Miyano ⁽³⁾

⁽¹⁾ Graduate School of Information Science, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
masao@ims.u-tokyo.ac.jp

⁽²⁾ Graduate School of Science and Engineering, Yamaguchi University, 1677-1 Yoshida, Yamaguchi 753-8512, Japan
atsushi@ib.sci.yamaguchi-u.ac.jp, matsuno@sci.yamaguchi-u.ac.jp

⁽³⁾ Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo, 108-8639, Japan
miyano@ims.u-tokyo.ac.jp

keywords: Simulation, Modeling, Biopathway databases, KEGG, BioCyc, Petri net

Introduction

Genomic Object Net (GON) (<http://www.genomicobject.net/>) is a biopathway modeling and simulation platform that employs the notion of hybrid functional Petri Net with extension (HFPNe) that extends the hybrid functional Petri Net (HFPN) [3, 4], and is developed with JAVA. With this platform, we have succeeded in modeling and simulating glycolytic pathway of *E. coli*, boundary formation by notch signaling in *Drosophila*, and apoptosis induced by *Fas* ligand, etc. For the modeling and simulation of a biopathway, suitable information selection from public biopathway databases, such as *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [1] (<http://www.genome.ad.jp/kegg/>) and BioCyc [2] (<http://www.biocyc.org/>), would be useful. Although the first aim for these pathway databases is to reorganize biochemical information for usage on computers and is not for modeling and simulation of biopathways. Thus, we have developed a way to transform these pathway databases so that the converted biopathways can run on GON. The transformation of the static biopathway models in KEGG and BioCyc leads users (biologists) to modify and refine the resulting dynamic biopathway models for their own interests on GON. This will be a first step for recreating biopathway databases towards simulation.

Converting Metabolic Pathways in KEGG and BioCyc

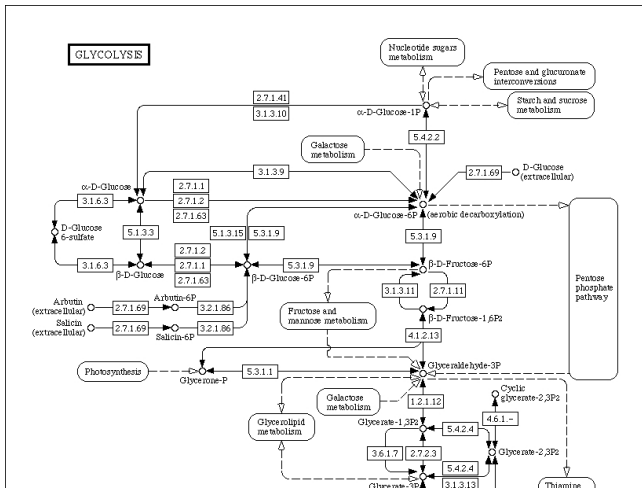
KEGG Pathway conversion and simulation

For each KEGG map, all enzyme reactions in the map are converted to HFPN. Then we obtain a simulatable reaction map. In KEGG map some compounds that appear in a reaction are omitted and some reactions in which the same enzyme is involved with more than two reactions are omitted for understandability. In our method, those omitted compounds and reactions are included but made invisible on Genomic Object Net.

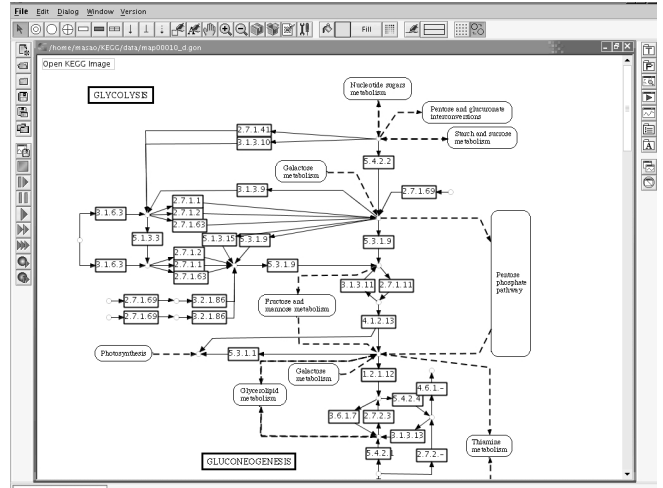
After the above process, as shown in Fig. 1, Glycolysis/Gluconeogenesis KEGG metabolic pathways in Fig. 1(a) is converted to Fig. 1(b). These two figures look almost the same, although the Fig. 1(b) can be simulated on Genomic Object Net. On the platform, as in Fig. 1(c) users can acquire information which enzymes work on each step and as in Fig. 1(d) users can acquire information from time-series values of enzymes and other compounds that are plotted on 2D graphs.

BioCyc pathway map and its conversion

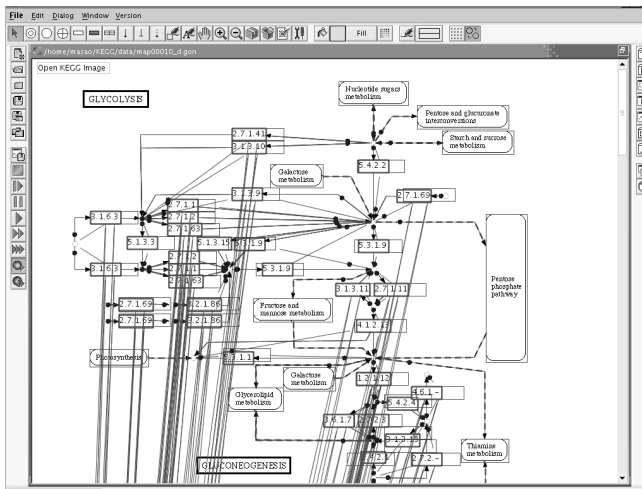
BioCyc is another collection of Pathway/Genome Databases [2]. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism, with the exception of the MetaCyc database, which is a reference source on metabolic pathways from many organisms [2]. As a conversion target, we selected EcoCyc that is a Literature-derived Pathway/Genome Databases for *Escherichia coli*. We can also convert pathways on EcoCyc to our platform as was done for KEGG.



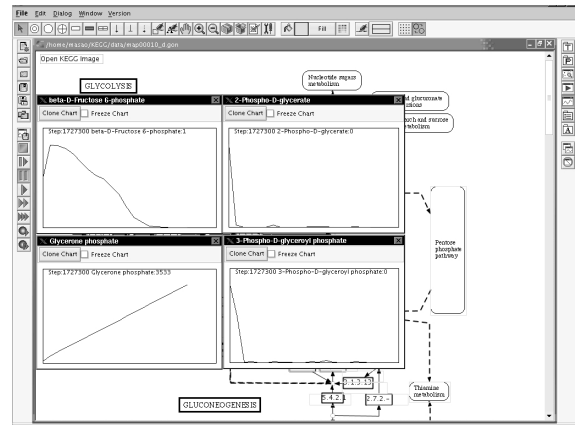
(a) KEGG pathwaymap



(b) Conversion result on GON



(c) Animated simulation on GON



(d) Time-series 2D graphs on GON

Figure 1. A conversion result for a KEGG map (glycolysis/gluconeogenesis map). All metabolic pathway maps of KEGG are also converted.

References

- [1] M. Kanehisa and S. Goto, KEGG: Kyoto encyclopedia of genes and genomics. *Nucleic Acids Res.* 28(1):27–30, 2002
- [2] P.D. Karp, M. Riley, S. Paley and A. Pellegrini-Toole, The MetaCyc Database, *Nucleic Acids Research* 30(1):59–61, 2002
- [3] H. Matsuno, A. Doi, M. Nagasaki and S. Miyano, Hybrid Petri net representation of gene regulatory network. *Proc. Pacific Symposium on Biocomputing* 5:338–349, 2000.
- [4] H. Matsuno, A. Doi, H. Hirata and S. Miyano, XML documentation of biopathways and their simulations in Genomic Object Net, *Genome Informatics* 12:54–62, 2001.

Description of the Physiome Project: the PHYSIOME is the quantitative description of the physiological dynamics or functions of the intact organism. The name comes from "physio-" (life) and "-ome" (as a whole).

The PHYSIOME PROJECT is an integrated multi-centric program to design, develop, implement, test and document, archive and disseminate quantitative information and integrative models of the functional behavior of organelles, cells, tissues, organs, and organisms. The long-range goal is to understand and describe the human organism, its physiology and pathophysiology, and to use this understanding in improving human health. but much or most of what must be learned will come from other species. The project aims toward providing models that summarize information on physiological systems, integrating the observations from many laboratories into quantitative, self-consistent, comprehensive descriptions. The goal is to provide to the community of scientists, physicians, teachers, and to medical health professional and industrial communities, functional descriptions of human biological systems in health and disease. A fundamental and major feature of the program is the databasing of the basic observations for retrieval and evaluation.

A network of Physiome Centers could comprise an adaptable international resource for databasing data on the functional aspects of biological systems covering the genome, molecular form and kinetics, cell biology, up to intact functioning organisms. These many databases would provide the raw information that might be integrated via physiological systems models, and should be structured hierarchically for accessibility and utility. The centers would maintain databases of information and models for retrieval over the Internet. The databases and models will have to accommodate data from many species.

Metabolic bioinformatics, metabolic reconstructions and mathematical simulation of the cellular metabolism. By Evgeni Selkov,^{1, 2} Evgeni Selkov, Jr.,¹ Yuri Grechkin,³ Igor Goryanin,¹ Milyausha Galimova,¹ Yuri Komarov,¹ Niels Larsen,⁴ Natalia Maltsev,³ Natalia Mikhailova,² Valeri Nenashev,¹ Ross Overbeek,³ Lyudmila Pronevich¹. From (1) Laboratory for Metabolic Bioinformatics and Mathematical Simulation, Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142292 Pushchino, Moscow region, Russia, (2) Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, 9700 S. Cass Ave., MCS-221, IL 60439-4844, USA, (3) EMP Project team, 142292 Pushchino, Pushchino, Moscow region, Russia, and (4) Center for Microbial Ecology, Michigan State University, Gilter Hall, East Lansing, MI 4884, USA .

We propose to contribute the following components to the Physiome Project:

1. database on Enzymes and Metabolic Pathways, EMP,
2. metabolic and functional reconstructions from genome sequence data, and
3. mathematical simulation of the cellular metabolism.

The Enzymes and Metabolic Pathways Database (EMP) is a general-purpose database [1-3] whose records cover all the main aspects of enzymology and metabolism: enzyme purification and characterization, kinetics, reaction mechanisms, thermodynamics, immunoreactivity, metabolism, regulation, and so forth. It represents all classified enzymes and includes records for some 1,500 organisms. Nearly one-third of the database records specify the enzymology and metabolism of humans and experimental animals. The information contained in EMP complements other widely used data banks on primary, secondary, and tertiary structures of enzymes, non-catalytic proteins and their genes.

The information stored in EMP comes from original articles published in leading international journals; each article is evaluated and translated into an EMP record by an expert annotator. The information is stored primarily as flat ASCII files that can be loaded into different database systems. In 1995, EMP 1.0 was released as a searchable database compatible with the Microsoft Windows platforms and distributed on CD-ROM by Biological Databases, Inc. [1]. In September 1997, EMP contained over 20,000 records compiled from about 14,000 original publications. By the end of 1997, EMP is expected to be accessible via the World Wide Web.

The Metabolic Pathways Database (MPW). Initially a small part of EMP, MPW is the most comprehensive collection of metabolic pathways in computer readable form [4]. The pathway collection, now including some 2,800 diagrams, covers not only primary and secondary metabolisms of the three forms of life, Archaea, Bacteria, and Eukarya, but also membrane transport, signal transduction pathways, intracellular traffic, translation, and transcription. The pathway diagrams were originally encoded as a formatted ASCII text, using pseudographics characters to draw arrows and boxes. A new release of MPW being developed now [5] is due to appear by October this year. In this release, the encoding is based on the logical structure of the pathways and represented by the objects commonly used in design and simulation of electronic circuits. This allows researchers to directly employ the wealth of knowledge and experience accumulated in this area of engineering. In particular, it makes possible automation of the basic simulation operations such as deriving stoichiometric matrices, rate laws, and, ultimately, dynamic models of metabolic pathways.

Metabolic and functional reconstructions from sequenced genomes. With the beginning of complete genome sequencing in 1995, MPW started to play a critical role in the metabolic reconstructions of sequenced genomes. By the term metabolic reconstruction we mean the process of inferring the metabolism and functional organization of an organism from its genetic sequence data supplemented by known biochemical and phenotypic data. Such reconstructions have recently been made for a number of available sequenced genomes. The reconstructions are accessible via the PUMA, WIT, and WIT2 systems, developed by Ross Overbeek and his colleagues [6-8].

Each reconstruction tends to be a consistent conceptual model in which all metabolic pathways are connected to sequence data stored in public databases. Here, consistent means that all intermediates of the metabolism are balanced by sources and sinks.

Initial metabolic reconstructions for a virtual human cell have been attempted under the PUMA [9] and KEGG [10] systems. These reconstructions are based on the known biochemistry [9] and available sequence data [9, 10]. To further extend this effort toward highly differentiated human cells, one needs EST (Expressed Sequence Tag) data to determine which part of the huge human genome is expressed in each specific tissue cell type. We believe that such reconstructions for a limited number of human tissue cells, such as those of smooth and skeletal muscles, liver, kidney, and blood, must be done first to form a solid basis for mathematical simulation of these tissues, organs and then the whole system.

Mathematical simulation of cellular metabolism. Quantitative simulation is the next step after genome sequencing and metabolic reconstruction. It is often argued that any serious quantitative simulation of cellular metabolism is not possible because it requires an enormous amount of quantitative data on kinetic parameters, concentration of intermediates, etc. Moreover, there is a severe limitation of the current reconstruction methodology based on sequence data: its inability to predict regulatory mechanisms and numerical values of all the kinetic parameters required for quantitative simulation and analysis of the metabolism.

We have begun to develop an approach that allows one to bypass this problem. This approach exploits mathematical simulation itself to predict missing control mechanisms and parameter values. The main idea is to optimize the structure and parameters of mathematical models to the known functions of simulated metabolic systems. Such optimization is a generalization of the parameter fitting widely used in simulation studies.

To realize this approach, we have developed a software package, DBsolve [11]. The package includes a large collection of encoded enzymatic reaction mechanisms and metabolic pathways that are used to build up complex metabolic models automatically. Specifically, DBsolve derives ordinary differential equation sets, ODE, from stoichiometric matrices of the enzyme reactions and metabolic pathways stored in EMP. For parameter estimation and model fitting, the software package has an optimization block. It uses a modification of the unconstrained zero-order minimization polygon method of [12]. The bifurcation analysis block is based on a package [13] translated into C++. DBsolve uses a version of Gear's method for integration of a stiff ODE with the controlled stiffness, thereby enabling the solution of hybrid ODE systems with algebraic subsets. An original parameter continuation algorithm is used for computing the algebraic system solution along a parameter range.

The importance of unicellular organisms for the Physiome Project should not be underestimated. Many bacterial metabolic systems are very similar to their mammalian counterparts and can be treated as much simpler experimental models. It has recently been discovered that the cyanobacteria *Synechocystis* and *Synechococcus* have a circadian cell clock that gates cell division [14-16]. We have reconstructed the metabolism of *Synechocystis* from its sequenced genome [7, 8] and found that the main postulates of a metabolic theory of the clock, developed earlier [17-19], hold true for this cyanobacterium. Thus, this bacterium lends us the shortest way to crack this enigma of cellular physiology and of human physiology in particular.

References

1. <http://www.biobase.com/EMP>
2. Selkov E., Basmanova S., Gaasterland T., Goryanin I., Gretchkin Y., Maltsev N., Nenashev V., Overbeek R., Panyushkina E., Pronevitch L., Selkov E., Jr., Yunus I. *Nucleic Acids Res.*, 1996, 24 (1), 26-29
3. <http://www.biobase.com/emphome.html/homepage.html/pags/pathways.html>
4. Selkov E., Galimova M., Goryanin I., Gretchkin Y., Ivanova N., Komarov Y., Maltsev N., Mikhailova N., Nenashev V., Overbeek R., Panyushkina E., Pronevitch L., Selkov E., Jr. *Nucleic Acids Res.*, 1997, 25 (1), 37-38
5. <http://beauty.isdn.mcs.anl.gov/~selkovjt/tmp>
6. http://www.mcs.anl.gov/home/compbio/PUMA/Production/puma_graphics.html
7. <http://www.mcs.anl.gov/home/compbio/WIT/wit.html>
8. <http://www.mcs.anl.gov/home/overbeek/WIT2/CGI/user.cgi>
9. <http://www.mcs.anl.gov/home/compbio/PUMA/Production/ReconstructedMetabolism/reconstruction.html>
10. <http://www.genome.ad.jp/kegg/kegg3.html>
11. <http://sirius.iteb.serpukhov.su/~igor>
12. Swann, W. H. In: *Numerical Methods for Unconstrained Optimization* (W. Murray, ed.), London, Academic Press, 1972, pp. 13-28
13. Khibnik, A., Kuznetsov, Y., Levitin, V., and Nikolaev, E. *Physica D*, 1993, 62, 360-370
14. Kondo T, Strayer C. A., Kulkarni R. D., Taylor W., Ishiura M., Golden S. S., Johnson C. H. *Proc. Natl. Acad. Sci. U.S.A.* 1993, 90(12), 5672-5676
15. Aoki S., Kondo T., Ishiura M. *J. Bacteriol*, 1995, 177(19), 5606-5611
16. Mori T., Binder B., Johnson C. H. *Proc. Natl. Acad. Sci. U.S.A.*, 1996, 93(19), 10183-10188
17. Sel'kov E. E. *Ber. Bunsenges. Phys. Chem.*, 1980, 84, 399-402
18. Avseenko N. V., Lisnichuk L. Ya., Sel'kov E. E. *Biofizika*, 1987, 32(2), 248-252
19. Selkov E. E., Lisnichuk L. Ya., Avseenko N. V. *Biofizika* 1989, 34(3), 459-456

Some Modeling Projects utilizing E-Cell

E2coli

An energy metabolism model of *Escherichia coli* was constructed using E-CELL System. The metabolism is represented as a hybrid model which includes quantitative kinetics based approach and qualitative stoichiometric approach. Central pathways of energy metabolism are represented as a kinetics based model. Metabolite concentrations are measured using Capillary Electrophoresis and Mass Spectrometer (CE/MS). An automated survey system was developed to get practical kinetic data and to uniform survey quality. Some pathways are represented as stoichiometric model because of unavailability of kinetic data. The stoichiometric model and quantitative kinetic model can be simulated simultaneously using a method we developed for hybrid simulation. A model database was also developed to manage data of the hybrid model. We plan to include gene expression model to express catabolite repression into the energy metabolism model. The simulation result will be compared with experimental data using isotope labeled glucose. Acknowledgement: This work was supported by New Energy and Industrial Technology Development Organization. (Kenta Hashimoto)

e-Rice

The complete genomic sequence of rice has been elucidated and its metabolism thoroughly researched, making rice an ideal prospect for modeling and simulation. As a basis for the in silico rice modeling project, e-Rice, the preliminary goal is to develop a model for simulating plant cell primary metabolism using E-CELL Simulation Environment. (Nobuyuki Ishii)

Erythrocyte

We have constructed the computer model of the human erythrocyte using E-CELL system. The model has three major metabolic pathways including glycolysis, the pentose phosphate pathway, nucleotide metabolism, and ion transport systems. And it has reached to a steady state, which is very closed to that of the real erythrocyte. Using this model, we carried out simulation experiments of the glucose-6-phosphate dehydrogenase (G6PD) deficiency. These simulation experiments suggested that glutathione metabolism, including the pathways producing reduced glutathione (GSH), the export system of oxidative glutathione (GSSG), and the glyoxalase system, activated in the condition, and highly influence the symptom of G6PD deficiency. After the addition of GSH metabolism to the conventional model, the lifetime, predicted from ATP level, of the cell was longer and the ratio of GSH/GSSG was higher. These results suggested that pathways that are insignificant in normal erythrocyte function may become essential; when studying abnormalities, surrounding metabolic pathways should be included in the simulation model. So, we are restructuring the whole human erythrocyte model on the E-CELL simulation environment. On accomplishing this task, the hybrid of stoichiometric and kinetic modeling methods (Yugi. K. and Nakayama. Y. et al.) was applied. We are challenging the expansion of erythrocyte model regarding not only the metabolic pathways but also its robustness and tolerance, adding the following functions: pH dependence of enzymes, osmotic balance, electroneutrality, and oxygen and carbon dioxide transportation by hemoglobin. And the automatic transformation tool of kinetic equations into S-system and GMA forms, which are consistent with a number of specific features of biochemical systems, have been developed for calculation the steady-state concentration of intermediates. (Yoichi Nakayama)

Circadian rhythm

In this work, we have estimated a gene regulatory network and signaling pathway of the circadian rhythm in *Synechococcus* sp. PCC 7942 using the E-CELL System. It is widely known that there is a problem in exploring a broad parameter space of biosystems, and the problem is difficult to solve

simply using the genetic algorithm (GA). In this work, we introduced the activation energy for each protein-protein interaction for eliminating of the searching space. We use the Smoluchowsky equation for representing the activation energy in our computer model, and explored parameters which give oscillating conditions. (Fumihiko Miyoshi)

E-Neuron

E-Neuron project aims the reproduction of nerve behaviors using kinetic equations of several reactions at molecular level in a neuron. In order to construct our project theoretically, we chose the concept of "systems biology" as our base theory. We call this theory "systems molecular neuroscience". It is that the equation system itself and the range of each parameter mainly cause a behavior rather than the value of each parameter. Therefore, we have to elucidate "parameter space for a behavior" of our model. Our recent works are simulations of signal transduction of nerve growth cone and synaptic plasticity. (Shinichi Kikuchi)

Mathematical Analysis

Our E-CELL project needs kinetic data for all reactions, however it is generally difficult to obtain them only from the literature. One solution is to measure the value directly. Actually, we started this "wet-approach" in IAB, Keio Univ. We think that another solution, "dry-approach", is also required together that estimates the parameters computationally from the limited data. Mathematical analysis group mainly aims to develop novel parameter estimation system. Moreover, our group are developing metabolic control analysis, metabolic flux analysis and flux balance analysis. Recently, we try to apply control theory to cell simulation. (Shinichi Kikuchi)

Cell Signaling

My current research has two main threads: (i) implementing a stochastic module in the E-Cell Simulation Environment Version 3, and (ii) modelling the bacterial chemotaxis signalling pathway. In (i), I am implementing the StochSim (1) algorithm within the E-Cell environment. This algorithm is similar but not equivalent to another stochastic algorithm for simulating chemical kinetics developed by Gillespie, and has properties that are particularly desirable for simulating cell signalling pathways. In (ii), I am extending my thesis work on modelling the bacterial chemotaxis receptor complex (3), in which I investigated the properties of a hypothetical spatial interaction between membrane receptors that could explain the high gain in the chemotaxis pathway (4). Because of its simplicity, this pathway is an ideal system for developing both modelling and experimental methods that can be applied to a wide variety of cell signalling systems. (Tom Shimizu)

References:

- (1) Morton-Firth, C. J., & Bray, D. (1998) "Predicting temporal fluctuations in an intracellular signalling pathway" *Journal of Theoretical Biology* 192:117-128.
- (2) Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput.Phys.* 22, 403-434.
- (3) Shimizu, T. S. (2002) "The spatial organisation of cell signalling pathways - a computer-based study" Ph. D. Thesis, University of Cambridge.
- (4) Bray, D. (2002) Bacterial chemotaxis and the question of gain *Proceedings of the National Academy of Sciences USA* 99:7-9.

Genomic Object Net: XML Visualization of Simulation Results from Biological Modeling on Hybrid Functional Petri Net

Hiroshi Matsuno¹

matsuno@sci.yamaguchi-u.ac.jp

Atsushi Doi¹

atsushi@ib.sci.yamaguchi-u.ac.jp

Sachie Fujita¹

fujita@ib.sci.yamaguchi-u.ac.jp

Makiko Sasaki²

maki-s@ims.u-tokyo.ac.jp

Yuichi Hirata²

hirata@ims.u-tokyo.ac.jp

Satoru Miyano²

miyano@ims.u-tokyo.ac.jp

¹ Faculty of Science, Yamaguchi University, 1677-1, Yoshida, Yamaguchi-shi, Yamaguchi 753-8512, Japan

² Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Keywords: Genomic Object Net, XML, visualization, hybrid functional Petri net

1 Introduction

In [2], we showed that hybrid Petri net (HPN) provides the promising basic architecture for representing biological processes with an example of description and simulation of λ phage genetic switch mechanism. We launched the project “Genomic Object Net Project” whose aim is to develop the software tool which can be used by biologist easily and intuitively.

As a first step, in [3], we introduced the new mathematical expression enhanced from HPN “hybrid functional Petri net (HFPN)” for realizing biopathways naturally and intuitively. Furthermore, by using Genomic Object Net Assembler [3], we showed how the concept of HFPN is suitable for describing and simulating biological processes through the realizations of circadian rhythms in *Drosophila melanogaster*, glycolytic pathway of *Escherichia coli* with the lac operon gene regulatory mechanism, and apoptosis induced by the protein Fas.

Recently, we developed a tool “Genomic Object Net Visualizer” based on XML technology which enables us to visualize simulation results produced by Genomic Object Net Assembler. By using this tool, users in biology/medicine can view simulation results on their own aspects.

In this software demonstration, we will present several visualization examples of simulation results including lac operon gene regulatory network, circadian rhythms in *Drosophila*, and Delta-Notch lateral inhibition mechanism.

2 Genomic Object Net Visualizer

“Genomic Object Net Visualizer” is developed on the basis of XML technology. Users can realize visualizations of simulation results of aimed biological phenomenon by creating XML document in which CSV files produced by Genomic Object Net Assembler are included as basic data for simulations.

Figure 1 shows a visualization example of circadian rhythms in *Drosophila*. The right window describes the concentration behaviors of PER protein and dCLK protein, and these two graphs are driven by CSV files produced by Genomic Object Net Assembler where HFPN circadian rhythm model is realized. The left upper window shows an animation of PER and TIM proteins and these complexes

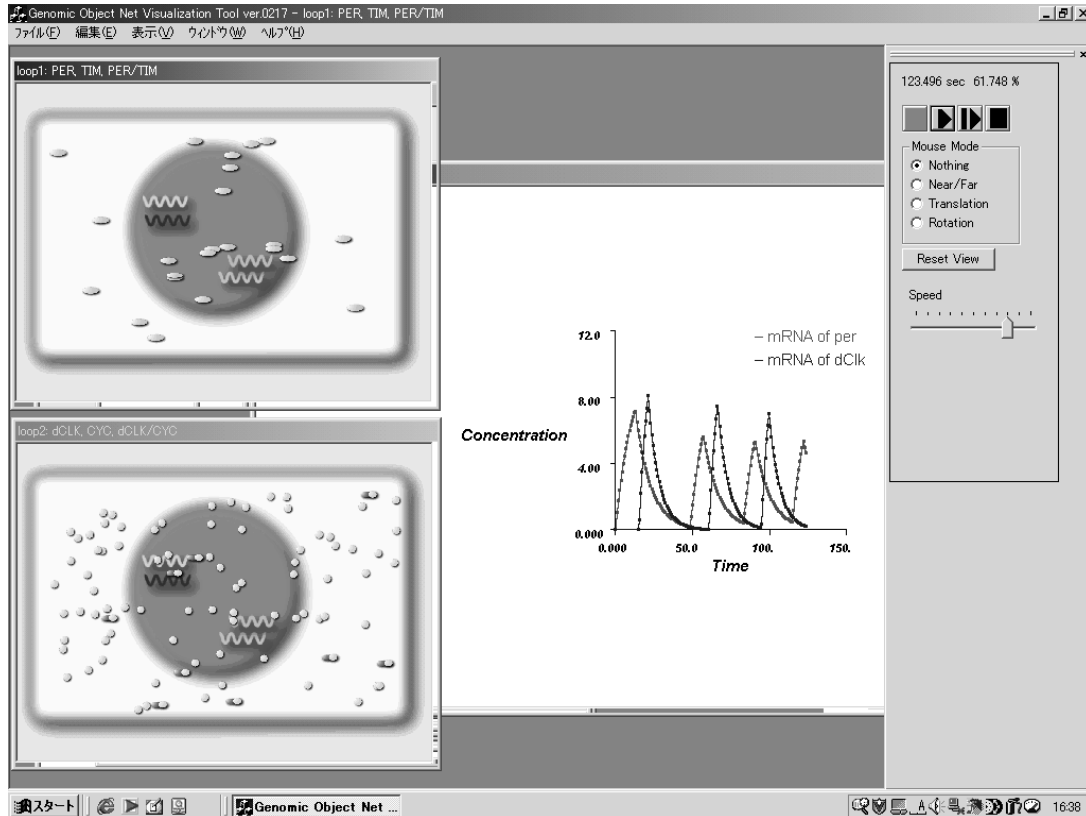


Figure 1: View simulation results as an animation by GON Visualizer.

whose numbers are changed according to these concentration levels. The right lower window shows an animation of dCLK and CYC proteins and these complexes whose number is changed according to that concentration level. Please refer to [1] for the details about genes including *per*, *tim*, *dClk*, and *cyc* which participate in gene regulatory mechanism of biological rhythms in *Drosophila*.

Acknowledgments

This work is partially supported by the Grand-in-Aid for Scientific Research on Priority Areas (C) “Genome Information Science” and Grand-in-Aid for Scientific Research (B) (No.12480080) from the Ministry of Education, Culture, Sports, Science and Technology in Japan

References

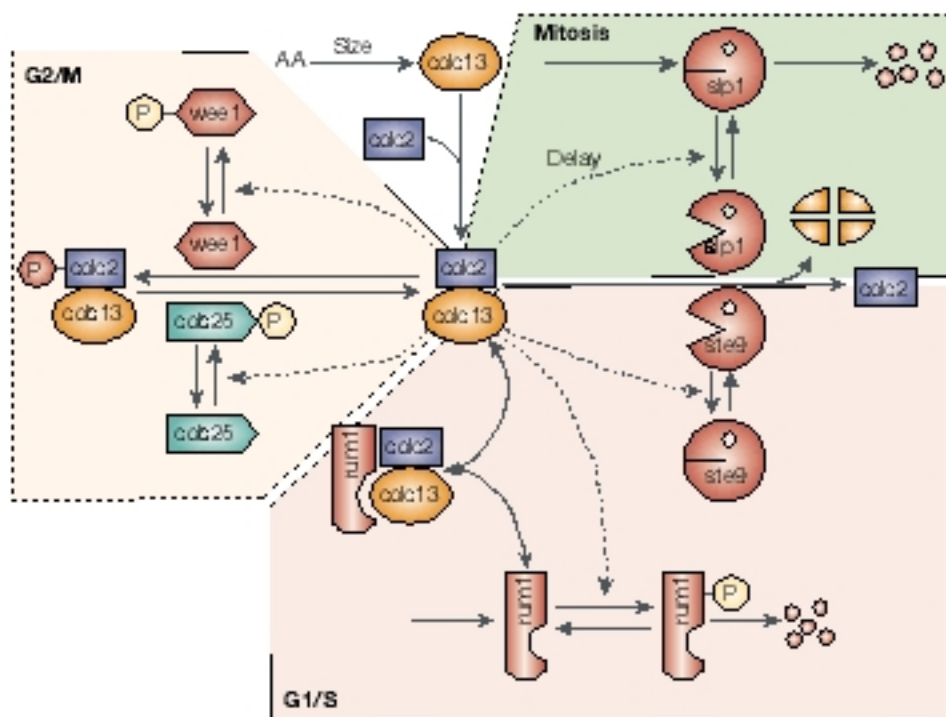
- [1] Hardin P.E., From biological clock to biological rhythms. *Genome Biology* 2000, **1**(4), reviews1023.1-1023.5, 2000.
- [2] Matsuno, H., Doi, A., Nagasaki, M., and Miyano, S., Hybrid Petri net representation of gene regulatory network. *Pacific Symposium on Biocomputing 2000*, 338–349, 2000.
- [3] Matsuno, H., Doi, A., Tanakan, Y., Aoshima, H., Hirata, Y., and Miyano, S., Genomic Object Net: basic architecture for representing and simulating biopathways, *submitted*, 2001.
- [4] Matsuno, H, Doi, A., Hirata, Y., and Miyano, S., XML documentation of biopathways and their simulations in Genomic Object Net, *Genome Informatics*, 12, 2001.

Modeling and simulation of genetic regulatory networks

A variety of methods for the modeling and simulation of genetic regulatory networks have been proposed, such as approaches based on differential equation models and stochastic models. These models provide detailed descriptions of genetic regulatory networks, down to the molecular level. In addition, they can be used to make precise, numerical predictions of the behavior of regulatory systems. Many excellent examples of the application of these methods to prokaryote and eukaryote networks can be found in the literature. For example:

Cell-cycle control system in fission yeast and equations describing the dynamics of the proteins cdc2-cdc13 and rum1. J.J. Tyson et al. (2001), *Nat. Rev. Mol. Cell. Biol.*, 2:908-916.

In many situations of biological interest, however, the application of differential equation and stochastic models is seriously hampered. In the first place, the biochemical reaction mechanisms underlying regulatory interactions are usually not or incompletely known. In the second place, quantitative information on kinetic parameters and molecular concentrations is only seldom available, even in the case of well-studied model systems.



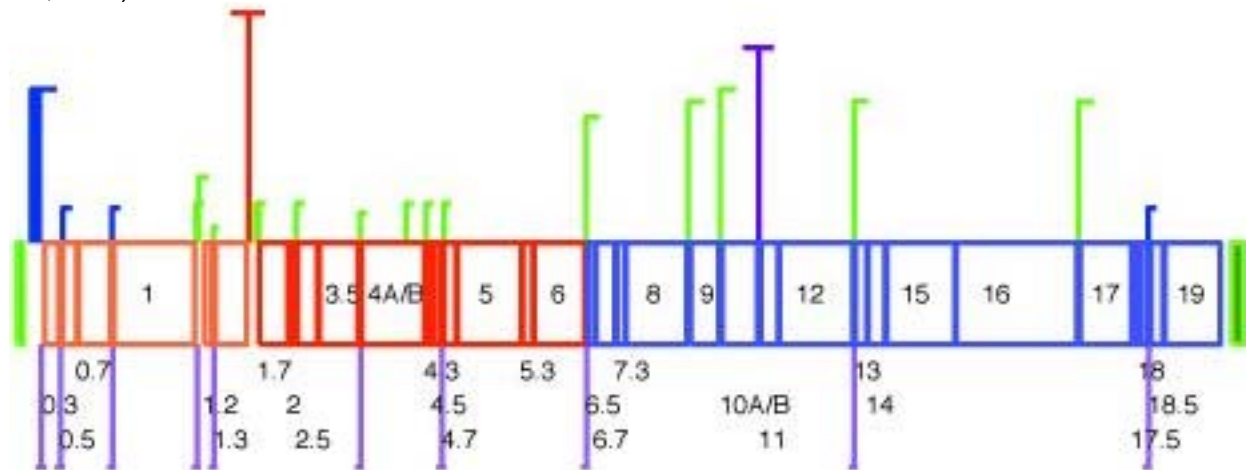
Literature

- H. de Jong (2002), Modeling and simulation of genetic regulatory systems: A literature review, *Journal of Computational Biology*, 9(1):69-105.
- J. Hasty, D. McMillen, F. Isaacs, J.J. Collins (2001), Computational studies of gene regulatory networks: in numero molecular biology, *Nature Review Genetics*, 2(4):268-279.
- P. Smolen, D.A. Baxter, J.H. Byrne (2000), Modeling transcriptional control in gene networks: Methods, recent results, and future directions, *Bulletin of Mathematical Biology*, 62(2):247-292.
- D. Thieffry, H. de Jong (2002), Modélisation, analyse et simulation des réseaux génétiques, *Médecine/Sciences*, 18(4):492-502.

Model of the T7 Genome

The genetic map of T7 is based on the sequence of 39,937 bp [3] and contains 56 known or potential genes. Genes are all transcribed from the same strand of DNA and are ordered by number sequentially from the genetic left end, which is the first end of the genome to enter the infected cell. Thus gene 0.3 is the first gene to be expressed and 19.5 the last. Integral numbered genes are unconditionally essential (except that gene 2 amber mutants grow on *E. coli* C) whereas most non-integer genes are non-essential or conditionally essential, exceptions being 2.5, 6.7, and 7.3. Coding sequences occupy almost 92% of the genome, most of the remainder containing the terminal repeats, origins of replication, promoters, and RNase III processing sites or other recognizable genetic signals. T7 genes are described as close-packed but there are five instances of potential overlapping genes that are read in a different frame: genes 4.1 and 4.2 almost entirely within gene 4 coding sequences, gene 18.7 lies within gene 18.5, and both genes 19.2 and 19.3 overlap the gene 19 sequence. In addition, by means of an in-frame internal initiation, gene 4 codes for two proteins, gp4A and gp4B. Further, programmed ribosomal frameshifting to the +1 frame during translation of gene 0.6 and to the -1 frame during translation of genes 5.5 and 10 affords, respectively, gp0.6B, a 168 residue 5.5-5.7 [3] fusion product and gp10B [4]. No biological function for these frame-shifted products is known but T3 contains a different yet comparable shifty sequence at the same relative position in gene 10 that also leads to a gp10B product [5]. The gene 10 homologues of other T7-like phages are also thought to make two products; gp10B may therefore be important under some conditions of infection.

Three classes of T7 genes have been recognized: class I genes are those expressed from about 2 min to 8 min after infection at 30C, class II genes from about 6 min to 15 min and class III genes from 8 min till lysis that occurs after ~ 25 min [6]. The ten class I genes are transcribed by *E. coli* RNA polymerase, their functions are mainly to subvert the bacterium into a phage-producing factory; the only essential phage gene is gene 1, coding for the T7 RNA polymerase. The latter enzyme transcribes the class II genes, mainly involved in phage DNA metabolism, and the class III genes, whose functions are predominantly morphogenetic. (adapted from I.J. Molineux, in *Encyclopedia of Molecular Biology*, T.E. Creighton, Ed., Wiley, NY, 1999)



T7 Genome in silico

In the simulation we represent the T7 genome at the level of genetic elements (genes, promoters, terminators, RNaseIII sites). More specifically, in T7v2.5 each genetic element is tracked using a number between 0 and 73; these are given in a key on the original website. Where, on the wild-type genome, genetic elements overlap we use poly-element blocks (i.e., we do not break up overlapping elements). In the wild-type genome the order of numbers increases from 0 to 73. By changing the order of numbers you change the order of the genetic elements. The names for the genetic elements fall into several classes. These are: EcXX = *E. coli* RNA polymerase promoters, ϕ XX = T7 RNA polymerase promoters, RXX = RNaseIII sites, italicized numbers = genes, and TE and T ϕ are the terminators for the *E. coli* and T7 RNA polymerase, respectively. The wild-type genome is shown above. The numbered boxes are genes (orange are class I, red are class II, blue are class III), the blue and green "periscopes" are *E. coli* and T7 RNA polymerase promoters, respectively, the red and purple "T's" are *E. coli* and T7 RNA polymerase terminators, and the inverted magenta "T's" are the RNaseIII sites. The promoter and terminator element's heights are directly proportional to their in vivo strengths/activities.