

<http://DOEGenomesToLife.org/compbio/>

Report on the Mathematics Workshop for the Genomes to Life Program

**U.S. Department of Energy
Gaithersburg, Maryland
March 18–19, 2002**

Workshop Organizers

**David L. Brown, Lawrence Livermore National Laboratory
John Guckenheimer, Cornell University
Esmond G. Ng, Lawrence Berkeley National Laboratory**

**Prepared by the Office of Advanced Scientific Computing Research
and
Office of Biological and Environmental Research
of the
U.S. Department of Energy
Office of Science**

December 2002



Table of Contents

Executive Summary	v
Introduction	1
The Mathematics of Genomes to Life	3
A Grand Challenge in Computational Biology	6
Goal 1: Computational Mathematics Required for Identifying and Characterizing the Molecular Machines of Life	9
Goals 2 and 3: Computational Mathematics in Support of Characterizing Gene Regulatory Networks and Simulating Complex Microbial Communities.....	12
A: Workshop Attendees.....	18
B: Workshop Agenda	20

Report on the Mathematics Workshop for the Genomes to Life Program

Executive Summary

The Mathematics of Genomes to Life

On March 18 and 19, 2002, a group of mathematical, computational, and biological scientists met in Gaithersburg, Maryland, to identify long-term mathematics research needs in biological areas represented by the Genomes to Life (GTL) program under development by the U.S. Department of Energy. Several speakers gave overviews of different ways in which applied and computational mathematics are expected to play an important role in this program. The remainder of the workshop was spent in smaller-group discussions on specific topics, with the goal of designating key areas of mathematics research that will contribute to GTL.

The workshop focused on computational mathematics techniques to identify and characterize the molecular machines of life and characterize gene regulatory networks and the

functional repertoire of complex microbial communities in their natural environments at the molecular level. Effectively modeling these complex biological processes will require substantial developments in many areas of computational mathematics involving discrete, continuous, and stochastic processes. This report summarizes workshop findings in regard to a broad range of mathematical and computational topics and techniques that will be expected to play a role in the long-term research program envisioned for GTL. These topics and techniques include the study of hybrid systems of differential, discrete, and stochastic equations modeling processes with multiple spatial and temporal scales; generalized dynamical systems; statistical modeling; and processes involving noise and uncertainty, differential geometry and topology, graph theory, and mesh generation, among others.

Report on the Mathematics Workshop for the Genomes to Life Program

**U.S. Department of Energy
Gaithersburg, Maryland
March 18–19, 2002**

Introduction

The U.S. Department of Energy (DOE) has the opportunity to bring to bear its unparalleled experience, expertise, and unique resources in computation on the field of modern biology, thus building the foundation for a new, comprehensive, and profound understanding of complex living systems through a new program called Genomes to Life (GTL). DOE's mission requires an understanding of the role of microorganisms in climate change and energy production, the bioremediation of energy and nuclear materials waste, and the health risks of low-dose radiation exposure. Problems of this scale and significance motivate the creation of a program aimed at a complete understanding of microbial systems. This program will build upon the remarkable successes of the Human Genome Project, coupled with DOE's strong foundation in mathematics and computer science required for large-scale scientific simulations, to develop a new computational bioscience program that will enable breakthrough advances in computational techniques for solving complex biological problems and predicting the behavior of complex biological systems.

DOE's current responsibility for remediating 1.7 trillion gallons of contaminated groundwater and 40 million cubic meters of contaminated soil demonstrates the significance and scale of the need for a new computational biology program. The need for groundwater remediation is a result of over 50 years of research, development, and testing of nuclear materials. Current state-of-the-art "pump-and-treat" technology is inefficient, only partially effective, and economically unjustifiable. The development of effective bioremediation techniques promises to provide an alternative approach for groundwater cleanup within DOE that will be both effective and economically feasible. Understanding the behavior and function of the microbes that will play an essential role in bioremediation will be possible, however, only through the development of powerful new computational approaches for modeling, simulating, and understanding the molecular machines of life and gene regulatory networks that govern cell function. DOE's missions to understand the roles of microorganisms in climate change and energy production and the health risks of low-dose radiation exposure also will require extensive development of new computational mathematics approaches. Combined with the revolutionary new technologies in experimental systems biology, computationally based

biology provides the key to the development of the new, comprehensive, and profound understanding of complex living systems that will be essential for DOE to meet its mission-driven challenges during the 21st century.

GTL envisions an aggressive computational and experimental plan for understanding microbial systems focused around four major goals:

Goal 1: Identify and characterize the molecular machines of life—the multiprotein complexes that execute cellular functions and govern cell form.

Goal 2: Characterize gene regulatory networks.

Goal 3: Characterize the functional repertoire of complex microbial communities in their natural environments at the molecular level.

Goal 4: Develop computational methods and capabilities to advance understanding of complex biological systems and predict their behavior.

This workshop represents an early attempt by DOE to identify the mathematical and computational technologies that will be required to support the GTL program. An extensive list of such techniques was compiled, but it should be understood that this is by no means a comprehensive list and that future workshops probably will be needed to fully scope out and understand the complete range of requisite GTL support technologies.

The Mathematics of Genomes to Life

On March 18 and 19, 2002, a group of mathematical, computational, and biological scientists met in Gaithersburg, Maryland, with the purpose of identifying the long-term mathematics research needs in the biological areas represented by the GTL program under development by DOE. Several speakers gave overviews of different areas in which applied and computational mathematics are expected to play an important role in this program. The remainder of the workshop was spent in smaller-group discussions on specific topics, with the goal of identifying the key areas of mathematics research that will contribute to GTL. In particular, they identified many of the anticipated mathematical challenges in GTL in the fields of differential equations, stochastic methods, combinatorial methods, statistical analysis, optimization, and other relevant areas. The agenda of the workshop and a list of participants are attached as appendices.

The March 2002 meeting represents the first time that DOE has taken a close look at the need for new developments in computational mathematics in support of systems biology. While several key areas were identified that will require substantial new investments in computational mathematics research, it should be understood that the topics identified in the current report do not represent a comprehensive survey of all the computational mathematics needs for Genomes to Life, and it is recommended that future workshops be organized to study these issues more thoroughly.

This workshop focused on Goal 4 of the GTL program, which emphasizes the essential significance that scientific computation will

play in meeting the objectives of Goals 1 through 3. With this in mind, participants identified areas of computational mathematics research that will be essential for development of a computational program in support of Goals 1 through 3. The participants also focused on the development of several illustrative examples of representative biological grand-challenge problems whose solution would represent significant advances in our ability to use advanced computational tools to understand and manipulate complex microbial systems. One of these examples is used in this report to illustrate the essential, broad-reaching role that computational mathematics will play in realizing the new systems biology of the 21st century.

Differential Equations

GTL Goals 2 and 3 depend critically on the development, analysis, and effective numerical solution of models describing processes that take place over an enormous range of both temporal and spatial scales. The models will involve systems of ordinary and partial differential equations, which have been the foundation of successful modeling in a great many disciplines. It is clear, however, that ordinary and partial differential equations alone cannot describe most of the complex problems arising in Goals 2 and 3. In addition to the deterministic behavior described by classical differential equations, effective models must incorporate stochastic behavior and uncertain parameters. In particular, we will need to understand how to solve hybrid systems of differential equations, discrete equations, and stochastic equations using computational approaches. Both the theory for understanding the long-time behavior of these highly complex generalized dynamical

systems and the numerical methods for actually solving them are still in their infancy. Furthermore, computational analysis tools are needed for the development and verification of models, including capabilities for sensitivity analysis, parameter estimation, automated bifurcation analysis, and model reduction and verification. In general, these analysis tools have not been developed for these hybrid systems of interest in computational biology.

Stochasticity

Biological systems are inherently stochastic at many levels. In more traditional computational simulation areas such as the study of chemically reacting fluid flows, reactions result from the random collision of molecules in the fluid. Since the number of molecules in the fluid is large, we can average over a large number of collisions to obtain continuous rate equations that accurately describe kinetics. In biological systems, there may be only a modest number of some of the proteins important to a molecular process, so approximation of the reactions with a continuous rate is not appropriate. As with chemically reacting fluid flows, a characteristic of reactions in biological systems is that details of the environment in which a reaction occurs can dramatically alter the rate of reaction.

Noisy Data

Recently developed experimental procedures in biology can generate a large quantity of data on a much larger scale than has resulted from traditional biological experimentation over the past century. A particular feature associated with biological data that differentiates it from the experimental data from other disciplines is that it will be very noisy.

Consequently, new methodologies for dealing with noise and uncertainty will be central to the analysis of experimental data.

Complexity

A third inherent feature of biological modeling is the complexity of biological systems. The molecular processes that control metabolic functions in cells involve complex networks of reactions with feedback and control loops. Also, there are often complex relationships among the different processes, which may operate symbiotically or in competition.

Statistics

GTL Goals 2 and 3 require the linking of massive amounts of data measured at the cellular level to a consistent network model involving all cell components and their actions and modeled mathematically by a system of (possibly stochastic) differential equations. Although large volumes of data will be available, the data will, in general, be very noisy. We need to develop statistical experimental design methodologies aimed at maximizing the utility of experiments, and these methodologies need to be incorporated into the planning phase for experimental studies. Statistics will be used for developing methods and theory for using data to select consistent statistical models and for the estimation of unknown parameters (with confidence measures) in the selected model. In this case, the data consist of genomic DNA sequence data and phenotypes of the cell observed under particular environmental conditions and at a number of time points. The phenotypes of the cell are in the form of gene-specific RNA concentrations, protein concentrations, images of the cell, and the molecules it contains. Statistics plays an important role in the validation of these

biological mathematical models. Statistics can also contribute more generally to methods for incrementally building up a model or selecting between a set of models.

Geometry, Topology, Graph Theory, Grid Generation

Differential geometry and topology are relevant in understanding a number of issues related to molecular machines. Some examples include the structural conformation of closed circular DNA, the free energy associated with supercoiling in closed DNA, the actions of topoisomerases and recombinases, nucleosome winding, and the binding between proteins and DNA. Algebraic and differential topologies provide information about global restraints on a cellular structure and may pinpoint details on the mechanical structure. Computational geometry, geometric topology, and issues of shape spaces, registration, organization, geometry, and analysis are useful in understanding the trajectory, dynamics, and shape changes of the molecular machine. There are also issues of the relationship of the machine to its living environment that may benefit from results in variable geometry and stochastic topology. The continuing interaction of the machine and its environment can lead to a patterned variation in its geometric and topological structure.

There are other related areas of relevance to computational biology. For example, the embedding invariants for graphs have been used in studying topoisomers; random knots

are employed in the study of structures of macromolecules; and the tangle calculus is used in the study of the DNA enzyme mechanism. Since many of the biological systems may have complex interior or exterior geometries, new techniques for geometry representation and grid generation may be needed in their simulations.

Software Challenges

For biologists to make effective use of the anticipated mathematical and computational developments in simulation and computational analysis, these developments must be made available in software tools that should be easy to use. Eventually, problem-solving environments will be required to make the software accessible to the wide community of biologists working in areas related to Goals 1 through 3 of this initiative.

Modes of Research

Realizing GTL's vision of the new field of systems biology will require substantial development of new mathematics and computational techniques. Close collaboration among computational mathematicians and biologists over a 5- to 10-year time frame will lead to discoveries of how to apply existing mathematical techniques to biological problems. In addition, long-term (more than 10 to 20 years) basic research to develop new mathematical and computational techniques will result in new discoveries in computational biology that we cannot currently anticipate.

A Grand Challenge in Computational Biology

The success of the GTL program will depend on the availability and development of a large variety of mathematical and computational tools. Many of these tools are common in a large number of applications. A representative application is introduced in this section to illustrate the mathematics and computation that are required.

Bioremediation of Radionuclide-Contaminated Groundwater

For more than 50 years, the United States created a vast network of more than 113 facilities for research, development, and testing of nuclear materials. As a result of these activities, subsurface contamination has been identified at over 7000 discrete sites across the DOE complex. With the end of the Cold War threat, DOE has shifted its emphasis to remediation, decommissioning, and decontamination of the immense volumes of contaminated groundwater, sediments, and structures at its sites. DOE is currently responsible for remediating 1.7 trillion gallons of contaminated groundwater, an amount equal to about 4 times the daily U.S. water consumption; and 40 million cubic meters of contaminated soil, enough to fill about 17 professional sports stadiums. Estimates are that more than 60% of DOE facilities have groundwater contaminated with metals or radionuclides. The only contaminants that appear in groundwater more often than metal and radionuclide contaminants are chlorinated hydrocarbons. More than 50% of all soil and sediments at DOE facilities are contaminated with metal and radionuclides, the contaminants found with the highest frequency in soil at all DOE waste sites. Indeed, while virtually all con-

taminants found at industrial sites nationwide can also be found at DOE sites, many metals and especially radionuclides are unique to DOE sites.

Traditional remediation is often inefficient and expensive. Current technology for treatment of groundwater contaminated with metals or radionuclides is “pump and treat,” followed by disposal or reinjection of treated water. This process can be costly and inefficient due to the difficulty of completely removing contaminated groundwater and the sorption of contaminants on mineral surfaces. DOE's Office of Environmental Management (EM), which is responsible for the cleanup, has stated that advances in science and technology are critical for DOE to reduce costs and successfully address these long-term problems.

Bioremediation of metals and radionuclides. The catalytic potential of microorganisms in nature is enormous and yet still relatively untapped for use in environmental cleanup. Bioremediation is the use of microorganisms to decrease, eliminate, modify, or contain hazardous and radioactive wastes to environmentally safe levels. While bioremediation of organic contaminants involves their transformation to benign products such as carbon dioxide, bioremediation of metals and radionuclides involves their removal from the aqueous phase to reduce risk to humans and the environment. Microorganisms can directly transform metals and radionuclides by changing their oxidation state to a reduced form that leads to in situ immobilization. Alternatively, microorganisms can indirectly immobilize metals and radionuclides through the reduction of inorganic ions, which can, in turn, chemically reduce contaminants to less

mobile forms. The long-term stability of these reduced contaminants is as yet unknown. Other mechanisms whereby microorganisms can influence mobility include alteration of pH, oxidation, and complexation.

***Shewanella oneidensis* and uranium.**

S. oneidensis has remarkable versatility in its ability to use various electron acceptors, including oxide, nitrate, fumarate, manganese, iron, and sulfur. This makes the bacterium a strong candidate to substitute contaminants such as uranium or technetium. Researchers are investigating whether *Shewanella* can add electrons (reduce) onto uranium (VI), making it uranium (IV). U (VI) is soluble in water, thus contaminating groundwater and spreading the uranium beyond the original source. When transformed into U (IV) (reduced), it is less soluble and more easily contained.

Uranium is reduced through an oxidation/reduction reaction that transfers electrons from electron donors such as organic material to electron acceptors such as sulfates, or, in this case, uranium. Energy is gained in the process that drives the metabolism of the cell.

Biologists need to know much more about the molecules that transfer those electrons and the pathway that contains these molecules, because little is known about the mechanisms. Multiple copies of similar proteins are approximate substitutes for each other. For example, it is believed that there are genes for at least seven C-type chromosomes in *Shewanella*. The pathway that transports the electrons from the electron donor, such as organic matter, to the electron acceptor, such as uranium, involves several individual reactions. Little is known about how many different pathways are possible from the genes in the genome and how these pathways are regulated in response to vary-

ing environmental conditions, such as the availability of various electron acceptors and donors.

Goal 1: Machines

The number of cytochromes (electron transfer agents) in microbes can range from a few to over 200, and the cytochromes from bacterium to bacterium differ in structure, localization, and perhaps function. Knowing more about the geometry and other characteristics of the cytochromes will enable deeper understanding of how they work and perhaps answer the following questions: What affects the specificity and activation of a cytochrome? Which cytochromes that have evolved for iron or nitrate can adapt to uranium or technetium? If we were to engineer a cytochrome for chromium, what criteria would we use?

Much is unknown regarding the actual mechanism of transferring electrons from molecule (usually but not always a protein) to molecule along the electron transport pathway. Some proteins are anchored to the membrane; others are thought to move about more freely. Because the intermembrane space (periplasm) is not electrically conductive, it is thought that the proteins need to be in close contact with each other for the electron to be transported. These interactions among proteins and other molecules, still largely unknown, are defined as molecular machines and are an example of the object of Goal 1.

Goal 2: Gene Regulatory Networks

The redox reaction conducted by *Shewanella* is the result of a long chain of reactions that strip energy off the substrates as electrons are passed from one molecule to the next. A mathematical description of that pathway and other pathways that provide substrates

or carry away products helps us understand what affects the reaction we care about (reducing uranium). Some pertinent questions include the following: Why does one species reduce uranium and another not? Why do electron acceptors sometimes get used in an order that differs from their chemical reduction potential? Is there something about alternative pathways that would help explain that? How are these electron pathways regulated? With the variety of acceptors and donors, the large number of analogous proteins such as cytochromes, and so forth, the combinatorial aspects of understanding and predicting pathways multiplies rapidly.

Goal 3: Microbe Communities

Shewanella will compete against other dissimilar metal-reducing bacteria (DMRB) in the community of microbes that live in the

contaminated zone (the substrates are used for energy supply rather than absorbed into the biomass). We will need to decide whether some species of DMRB, such as *Shewanella*, are preferable for our purposes to others, such as *Geobacter* or *Desulfovibrio*. If so, how do we adjust the environment to favor the bacteria we prefer? How do we find out which organisms are present and what genes they have without the ability to culture more than about 1%? What other species are needed for the metabolism of the community to be adjusted so that we maximize the reduction of uranium? For example, can we alter the environment so that other species reduce the nitrate if present, allowing *Shewanella* to use the uranium as the primary acceptor (otherwise it might use the nitrate)?

GOAL 1: Computational Mathematics Required for Identifying and Characterizing the Molecular Machines of Life

Discovering the existence and understanding the structure and function of molecular machines are among the first but crucial steps towards the prediction of the behavior of complex biological systems. The identification and characterization of molecular machines require the use of a wide range of mathematical techniques, such as geometry, topology, and statistics. Geometry and topology are useful in studying the shapes and structures of DNA molecules and their interactions with protein molecules. Statistics is applicable in the handling of noise, uncertainties, and design of experiments.

Understanding the geometric and topological properties of DNA. The geometric and topological properties of DNA molecules are critical aspects of their function. Protein-ligand binding, protein docking, protein modulation of DNA expression, altering the function of DNAs, and many other aspects of protein function depend on precise geometric alignment and topological arrangements. Computational geometry and topology are expected to be useful in understanding and modeling such phenomena. Carefully controlling the local shape (geometry) of the DNA double helix is crucial for cellular metabolism. Some proteins bind to DNA, bending it to make distal sites on the DNA spatially juxtapose, thereby facilitating binding of other enzymes that require a pair of sites on which to operate. Other proteins must disassemble the double helix in a gene to facilitate expression of that gene, and, when finished, the original DNA must be reassembled. The double helix also must be taken apart and reassembled to facilitate DNA repair. Enzymes (such as topoisomerases and recombinases), which

manipulate the geometry and topology of cellular DNA, perform many important cellular processes (including segregation of daughter chromosomes, gene regulation, DNA repair, and generation of antibody diversity). In the topological approach to enzymology, circular DNA is incubated with an enzyme, producing an enzyme signature in the form of DNA knots and links. By observing the changes in DNA geometry (supercoiling) and topology (knotting and linking) due to enzyme action, the enzyme binding and mechanism often can be characterized. Once the geometry of protein-DNA binding and topology of the enzyme function (breakage, rotation and reconnection of DNA strands) are known, the results of knowledge of the protein and DNA-protein crystallography can be productively interpreted. The geometric and topological information shows where to look on the protein crystal for the moving parts of the molecular machine, and this information can then be used to design drugs that enhance or inhibit enzyme mechanism.

It is necessary to build analytical models of protein-DNA binding, as well as computational models to simulate the interaction of flexible 3D proteins and highly 1D DNA. To study some of the cellular functions, mathematical models of 3D shape also are needed to allow local and detailed description of geometry of protein and DNA surfaces and the local changes in shape. However, the geometry of such biological assemblies can be difficult to describe concisely because of their complex, curvilinear boundaries. New geometric modeling techniques could be very useful. Discretizations of interiors or exteriors of these complex geometries are sometimes necessary for simulations. New

mesh-generation techniques for such problems are needed. Furthermore, these shape models should be able to describe and compute large-scale changes in protein conformation, which molecular machines use to facilitate their mechanism. Simulation of such machines will entail stochastic geometry and topology (i.e., random perturbations of local geometry and global topology) to determine any moving parts in the machines.

A particular challenge associated with the geometry of biomolecules arises from their flexibility. Useful geometric models and mesh-generation methods must be able to handle the significant structural deformations that are intrinsic to the functioning of many pathways. For example, protein-ligand binding often requires the protein to change shape as the ligand nears the binding site. Unfortunately, these systems are often of large scale and therefore computationally expensive. Methods for reducing the space of allowable deformations would have a big impact on the ability to simulate systems for long periods of time. Techniques that exploit the structure of these systems to identify the presumably small number of interesting degrees of freedom would be of great value. Possible approaches include computational topology and ideas from the study of graph rigidity.

Modeling of biomolecules. Understanding the interactions of large biomolecules requires accurate models of energetics. In principle, these can be computed precisely via quantum mechanics, but this is feasible only for small systems. For large systems, classical mechanics is generally employed, with sophisticated potential fields describing the energetics of atomic interactions. The computation of low-energy geometries becomes a nonlinear, global optimization problem in a high-dimensional space and for which there

are an enormous number of local minimizers. Improvements in optimization techniques for such problems will enable larger and more realistic simulations and also help with the validation and improvement of the empirical energetics models.

Combinatoric techniques for predicting protein structures. Since there are only about 300 known folding motifs in proteins, methods to predict the tertiary structure of proteins from sequence data have been useful in assigning function to newly discovered proteins. Many proteins are built as dimers or tetramers from similar modules, and the interaction of these modules determines the active site responsible for the protein function. The knowledge of orthologous and paralogous proteins also can be helpful in assigning function and identifying the proteins involved in the molecular machines of life. Current techniques such as threading and clustering have enjoyed limited success in understanding the intrinsic structures, but new techniques from combinatorics and geometry remain to be discovered and developed to solve these challenging problems.

Novel experimental technologies for investigating large molecules continue to advance rapidly. Similar to recently developed genomic and structural biology techniques, automatable, rapid, and inexpensive technologies for protein sequencing are being developed. Methods to quickly and accurately interpret these experimental results are essential. As experimental methodologies evolve, the analytical tools required to understand and interpret the results need to be developed as well. Some examples include isotope-labeled versions of nuclear magnetic resonance (NMR) to measure molecular distances and techniques such as MALDI (matrix-assisted laser dissociation-ionization) and SELDI (surface-

enhanced laser dissociation and ionization) time-of-flight mass spectrometry techniques that are being developed to sequence proteins rapidly from complex mixtures.

Combinatorial techniques have already found applications in these novel experimental methodologies. For example, interpretation of NMR data employs matching in bipartite graphs to assign peaks to residues. De novo protein sequencing with mass spectrometry has employed algorithms for finding most likely paths in associated graphs. Both types of experimental data benefit from fast database searches to find similar patterns in known molecules.

Statistical analysis. In addition to using topological and graph-theoretic approaches to reduce the space of all 3D configurations of proteins to a lower dimensional space, statistical methods can be considered as well. One can employ appropriate data transformations, data-reduction techniques such as principal component analysis, and clustering techniques. This type of dimension reduction will help to model the microorganism at the cellular level. In particular, the issue of selecting the true underlying dimension (e.g., the number of clusters) is an important problem. It is also important to establish relations between the sequence of a protein and its structure. For example, there are many more sequences than structures, which makes it important to determine what sequence components are predictive of the structure. This corresponds to a regression problem having thousands of features but also a high-dimensional outcome. Much work on methodologies needs to be done in this area.

In the characterization of multiprotein complexes, protein sites that are binding domains for other proteins and the kind of structures that allow binding to occur need to

be understood. Sequence data, structural data, and binding information for all pairs of proteins are determined experimentally. The statistical problem is to find the motifs of the proteins and the properties of their 3D structures predicting the binding event. However, there is an additional layer of complexity: both proteins are provided as sequence data and structure data, while the outcome is the single indicator of binding of the proteins. Multivariate prediction/regression models might need to incorporate this special coupling of covariates in order to allow an optimal search for the significant features of the sequence and structures of the motifs.

Biological experiments generating high-throughput data (e.g., microarray data) are subject to experimental noise; repetition of the same experiment will generate a different, but probably similar, data set. It is important to design experiments that are the most informative in answering particular questions of interest. It is also desirable to be able to determine an optimal number of times an experiment might need to be repeated for sufficient confidence in the answer. In other words, access to sample size formulas is needed to provide enough power to detect enough effects of a certain size. In addition, each analysis needs to provide precision estimates and reproducibility measures of findings in the data. Bootstrapping is a powerful but compute-intensive statistical approach that can provide such confidence measures. The method involves repeatedly simulating data from an empirical approximation of the true data-generating distribution. Given the high-dimensional setting in biological applications, research is needed to investigate and refine computational inference procedures such as bootstrapping and Monte-Carlo cross-validation for testing, estimation, and reliability assessment.

GOALS 2 and 3: Computational Mathematics in Support of Characterizing Gene Regulatory Networks and Simulating Complex Microbial Communities

GTL Goals 2 and 3 seek to characterize gene regulatory networks and also the functional repertoire of complex microbial communities in their natural environments at the molecular level. While these two goals pertain to biology at very different spatial and temporal scales, the computational mathematics required to model and understand the relevant processes have many aspects in common. In particular, the understanding of stochastic differential equations and the development of computational methods for solving them will play a pivotal role in accomplishing both goals. For this reason, we discuss the two goals together in this section.

Gene regulatory networks govern which genes are expressed in a cell at any given time, how much product is made from each one, and the cell's responses to diverse environmental cues and intracellular signals. Mathematical tools must be developed to discover the architecture, function, and dynamics of these networks from experimental data, to make useful computational models of them, and to solve the systems of mathematical equations and thus simulate the behavior of the regulatory networks. Similarly, complex microbial communities play an extremely important role in several of DOE's most distinguishing missions. These communities catalyze such crucial environmental processes as the recycling of carbon, nitrogen, and trace elements; the transformations of contaminants from toxic to benign forms; and the transformation of reduced and oxidized forms of carbon. Mathematical descriptions of microbial communities must

be developed based either on discrete models or in terms of (possibly stochastic) differential equations.

Modeling Gene Regulatory Networks. The first step in modeling a particular gene regulatory network is to discover the architecture and function of the network. The state of a cell can be characterized by the concentrations of each of its components (the DNA, RNA, protein molecules, and multiprotein molecules). The development of the cell's state over time can be described by (1) a causal graph that tells for each molecule (represented as a node in the graph) which molecules (i.e., the parents) influence the molecule (arrows are shown from parent nodes into this node) and by (2) an actual parametric description of the functional relation between each node and its parents. Directed, weighted edges reflect the influence of one molecule upon another. In cases where several molecules collectively influence another, a hypergraph model is more appropriate. The structure of these hypergraphs can be used to understand the fundamental properties of a regulatory network, like the rate-limiting step in a complex cascade of reactions.

Matroids, Petri nets, and polyhedral optimization have found applications in understanding the cycle structures in biochemical pathways such as the Krebs cycle. These combinatorial techniques can lead to understanding in situations where the gene regulatory network is not well characterized. They also can be potentially used to design “what

if” questions to develop metabolically engineered versions of the networks for bioremediation, manufacture of pharmaceuticals, and related applications.

The model must be developed so that it is consistent with the observed experimental data. The challenge is to reconstruct the gene regulatory network from experimental measurements on the dynamics of all the genes. With recent advances in DNA microarray technology, measurement of gene-expression levels has become possible on a genomic scale under various biochemical and environmental conditions. However, much remains to be done to improve the quality of the measurements and to reduce the cost of these experiments. There is also a need to develop statistical methods to extract meaningful information from such noisy data.

Constructing the causal graphs that represent gene regulatory networks is done through careful experimentation and painstaking manual inference. Automatic or semiautomatic methods for inferring the architecture of regulatory networks will be required in the future. One proposed method uses hierarchical clustering algorithms. Unfortunately, the sensitivity of this approach to noise is not understood, and, in addition, only the correlations and not the causal relations among the various genes and signaling pathways can be identified. Another drawback of this class of methods is that they may miss alternative pathways that are followed during stressed conditions. While this class of methods may therefore not be suitable for explaining and predicting the dynamical behaviors of the underlying systems, they may still be useful for organizing data and for discovering patterns.

Methods for reverse-engineering gene networks must be developed to identify network structure from microarray measurements, recovering causal relations, interactive effects, and alternative pathways in addition to dominant pathways. Concepts from system-identification theory, applied neural networks, genetic algorithms, and Bayesian models may prove valuable in this context. The usefulness of these methods has been demonstrated in recovering small networks and in distinguishing among several competing models. However, for reconstructing network structures de novo, they are not very efficient in terms of both data requirements and computational cost.

Another approach is to model gene regulatory networks with systems of ordinary differential equations and to estimate the connectivity parameters using various statistical methods such as singular value decomposition and robust regression. These methods are efficient in recovering the architecture for large networks such as those that occur in natural genomes. Much in the spirit of systems biology, the goal of these methods is to extract the gene regulatory networks on a global scale and to do so efficiently to identify individual subnetworks in a first draft of the entire network's topology, upon which further, more local analysis can be based.

While promising, these efficient large-scale methods are only in their early stages of development, and much research will be required before they will become routine tools for systems biology. For example, in eukaryotic cells, significant time delays may result from biochemical species crossing the nuclear membrane. Such species may interact in small numbers so that biological noise becomes significant. Moreover, some proteins

may form complexes whose effects cannot be deduced from their components. Existing methods, therefore, will need to be adapted to include time delays, stochasticity, and combinatorial effects.

Once a model of the network has been constructed, its behavior can be modeled. From a functional point of view, a regulatory network consists of a long chain of chemical reactions modeled by ODEs. Reactions involving molecules present in high concentrations might be described by a deterministic differential equation model, whereas reactions involving molecules present in moderate concentrations might be described by a mesoscale stochastic differential equation model. The modeling of reactions involving molecules present in low concentrations requires a purely stochastic approach. A gene regulatory network typically will include all of these situations simultaneously. Research is needed in formulating and understanding the behavior of such a model and where it can be simplified without sacrificing its intended predictive capabilities. The long-term success of GTL relies on our ability to effectively model phenomena described by systems of stochastic evolution equations.

However, the basic mathematical theory of these types of systems does not provide an adequate foundation for understanding the behavior of complex biological systems. Furthermore, our knowledge of numerical approximation of stochastic systems lags considerably behind our knowledge of deterministic systems. A critical mathematical issue in developing methodologies for simulating these types of systems is the large number of stochastic degrees of freedom that must be modeled. Simplistic Monte Carlo approaches will rapidly become computationally intractable. Approaches such as the polynomial

chaos techniques currently under development offer the possibility of simulating stochastic systems more efficiently than Monte Carlo; however, these types of approaches are currently limited to a modest number of stochastic variables.

In contrast to reaction models used in other areas of science and engineering, models for gene regulatory networks are not nearly so well established. As described above, the graph that defines the stoichiometry of the reaction network can be determined, with some uncertainty, by statistical analysis of bioarray data. Parameters such as reaction rates and probabilities are known only approximately. Hence it is also important to be able to assess the effect of these types of uncertainties in the system structure and parameters on its behavior. Sensitivity analysis, the tool by which one can quickly ascertain the parameters to which the system is most sensitive, will greatly facilitate model development and analysis.

Sensitivity analysis for ordinary and partial differential equation systems has been highly successful and widely used in many areas of science and engineering. Extending this computational tool to the class of generalized dynamical systems outlined above, however, is not a straightforward problem. Furthermore, the structural stability (i.e., the stability of the qualitative behavior of the system with respect to perturbations to its input) needs to be understood. Tools for bifurcation analysis also are needed and will face similar theoretical and computational challenges for this quite-general class of problems. Gene regulatory systems are complex feedback control systems. Tools and methodologies from this area will be invaluable in understanding their structure but face similar challenges due to the hybrid multiscale nature of this class of problems.

Research leading to understanding of multiscale phenomena will be required. In general, long-term basic mathematics research is needed for developing methodologies for multiscale problems. Multiscale phenomena appear in a number of forms in GTL. There are disparate time scales that must be represented in simulations. We need to be able to represent complex metabolic processes that are regulated by gene regulatory networks. We would like to develop simplified representations of subnetworks that effectively model the behavior of that submodel in large systems. In studying microbial communities, we would like to be able to develop models that represent the aggregate behavior of the community in much the same spirit as field theories in physics. Substantial additional research is required to develop appropriate methodologies for addressing the multiscale aspects of modeling microbial communities. Coarse-graining techniques will be necessary in the incorporation of information into biological models describing processes occurring at the larger scale. Research is required for development of techniques for coarse-graining and theory to assess how well and under what conditions the coarse-grained models retain the required predictive capabilities. The coarse-grained models may themselves be discrete or hybrid systems.

Robust parameter estimation. To create realizations of the biological models, it is necessary to estimate relevant model parameters from experimental data. One major area for mathematical and statistical research is the development of robust parameter-estimation procedures for the noisy environments characteristic of biological systems. Noisy parameter estimation arises in both the development of models for molecular processes in cells and in assembling the

submodules into an integrated model for the overall community. Approaches for parameter estimation need to be integrated with methodologies for sensitivity analysis and principal component analysis. These types of tools provide the functionality to identify key parameters that have a large effect on the overall dynamics and to identify critical subprocesses of the system.

Modeling microbial communities. Constructing models of microbial communities in their natural environment at a molecular level involves a number of steps. Models of cellular molecular processes, such as those discussed above, will be used. These models must include not only how the process works in isolation but also how gene regulatory networks regulate the process and how environmental factors modulate its behavior. These models must then be synthesized into models for each of the types of cells living in the community. Cellular models must then be integrated into a model for the community's environment. As GTL research progresses, developing tools not only to simulate the community but also to control its behavior will be required.

Once an initial model for the community has been developed, we can begin to simulate and analyze the behavior of the community. Initial simulations with the model will be used to validate the model. The validation of the model will involve close feedback with the individual component modules that are part of the model as the individual subsystems are refined to improve comparisons with databases of experimental data.

Once models for the various subprocesses have been developed, they will be integrated into a large stochastic model for the entire community. The models for subprocesses will range from stochastic differential equations

to discrete stochastic processes as well as hybrid models that include aspects of both. There are a number of strategies for developing models for cellular communities that represent different levels of both spatial and biochemical fidelity. Although several promising approaches are being developed in the biological and biomedical communities for modeling multicellular systems, substantial additional development will be required for them to adequately model a natural microbial community.


Analysis tools will be required. Another critical area of mathematical and statistical research needed for developing simulation capability for microbial systems is to develop the tools needed for analysis of both computational and simulation data. For experimental data, the tools should be capable of dealing with large but noisy data sets. We will need to develop tools for making justifiable statistical inferences from this type of data. For computational data we will need to analyze large volumes of data that represent, in some fashion, an ensemble of solutions to the system. Here, the volume of data is sufficiently large that we will need to develop mathematical tools for analysis rather than relying on visualization paradigms.

Ignoring algorithmic and computational issues, systems of stochastic differential equations can be linked to a statistical model (i.e., a probabilistic model for the actual observed data). Subsequently, the unknown parameters in this model (i.e., the coefficients in the stochastic differential equations) are estimated with maximum likelihood estimation. The uncertainty of parameter estimates also needs to be provided so that tests can determine which effects in each molecule-specific stochastic differential equation are

significant and thus should be included in the model. Model selection is a relevant statistical research area.

The statistical models result in Bayesian models by treating the unknown coefficients in the stochastic differential equations as random variables with a prior distribution. Bayesian models have the attractive property of being able to incorporate naturally prior information on the actual parameter values.

Modeling the complete cell with a huge system of stochastic differential equations corresponds with fitting the complete underlying probabilistic system as a whole to the observed data. The corresponding statistical models are referred to as parametric models. In semiparametric models, contrary to fully parameterized models, only a component of the whole system is modeled and no assumptions are made about the rest of the underlying system generating the data. All observed data are used, which makes this semiparametric approach truly different from just modeling a few system components and ignoring all data on other components, a typical approach followed in current literature. Contrary to fully parameterized models, in a semiparametric model the modeled component is only estimated and fitted, resulting in much better finite sample performance of parameter estimates. Second, by modeling the whole process, it is much more likely that the model is misspecified and nonidentified, so that none of the parameter estimates and corresponding confidence intervals can be trusted. In particular, semiparametric modeling can be used to minimize the modeling of many hidden unobserved variables not of interest and thereby is a practical way of dealing with many system variables that are only partly observed. Finally, estimation methods in



semiparametric models are much less computer intensive than fitting completely parameterized models. For these reasons, the use of semiparametric models will represent an important approach to achieving the final goal of modeling (learning) the whole system. In particular, it will allow reliable findings in

an incremental fashion (the preferable learning method). Although semiparametric models are routinely and successfully applied in medical research, epidemiologic studies, and so on, they have not been developed and used in biological research.

Appendix A: Workshop Attendees

Steve Ashby Lawrence Livermore National Laboratory	Sorin Istrail Celera
Tom Bartol Salk Institute	Gary Johnson DOE/OASCR
Douglas Baxter University of Texas, Houston	Mads Kaern Boston University
John Bell Lawrence Berkeley National Laboratory	Arthur Katz DOE/OBER
Craig Benham University of California, Davis	Louis H. Kauffman University of Illinois, Chicago
Bill Bosl Lawrence Livermore National Laboratory	Peter Kirchner DOE/OBER
David Brown Lawrence Livermore National Laboratory	Mike Knotek DOE/OBER
Angela Cheer University of California, Davis	Frank Larimer Oak Ridge National Laboratory
Yung-Sze Choi University of Connecticut, Math Center	Natalia Maltsev Argonne National Laboratory
Eric Darve Stanford University	Reinhold Mann Pacific Northwest National Laboratory
David Dixon Pacific Northwest National Laboratory	Noelle Metting DOE/OBER
Drew Endy Massachusetts Institute of Technology	Juan Meza Sandia National Laboratories
Marv Frazier DOE/OBER	William Miner DOE/OASCR
Leon Glass McGill University	Jorge More Argonne National Laboratory
Andrey Gorin Oak Ridge National Laboratory	Richard Mural Celera
Bill Gropp Argonne National Laboratory	Esmond Ng Lawrence Berkeley National Laboratory
John Guckenheimer Cornell University	Joe Oliveira Pacific Northwest National Laboratory
John Harer Duke University	Ed Oliver DOE/OASCR
Bruce Hendrickson Sandia National Laboratories	Linda Petzold University of California, Santa Barbara
John Houghton DOE/OBER	Ali Pinar Lawrence Berkeley National Laboratory
Fern Hunt NIST	Walt Polansky DOE/OASCR

Alex Pothen
Old Dominion University

Padma Raghavan
Pennsylvania State University

Chuck Romine
DOE/OASCR

Joel Saltz
Ohio State University

Lukasz Salwinski
University of California, Los Angeles

Steven Salzberg
TIGR

Nagiza Samatova
Oak Ridge National Laboratory

Robert D. Skeel
University of Illinois, Champaign-Urbana

Sylvia Spengler
NSF

Mike Steuerwalt
NSF

Joel Stiles
Pittsburgh Supercomputing Center

Dan Strahs
New York University

De Witt Summers
Florida State University

Peter Swain
Rockefeller University

Pieter Swart
Los Alamos National Laboratory

Harold Trease
Pacific Northwest National Laboratory

Ilya Vakser
State University of New York, Stony Brook

Mark Vanderlann
University of California, Berkeley

John van Rosendale
DOE/OASCR

Massimo Vergassola
Rockefeller University

Mike Viola
DOE/OBER

Alex Vologodskii
New York University

Scott Weidman
National Academy of Science

Alan Willse
Pacific Northwest National Laboratory

Matthew Wright
Harvard University

Steven Yeung
Boston University

Appendix B: Workshop Agenda

Mathematics Workshop for the Genomes to Life Program

Gaithersburg Hilton
620 Perry Parkway, Gaithersburg, MD

March 18–19, 2002

Organizers

David Brown	Lawrence Livermore National Laboratory
John Guckenheimer	Cornell University
Esmond G. Ng	Lawrence Berkeley National Laboratory

Program Managers

Gary Johnson	MICS
John Houghton	BER

Purpose

The purpose of the workshop is to bring together leaders in biology and mathematics to identify the long-term mathematics research needs in the biological areas represented by GTL. Several speakers will give overviews of different areas in which applied and computational mathematics are expected to play an important role in this program. Most of the workshop will be spent in smaller group discussions on specific topics, with the goal of identifying the key areas of mathematics research that will contribute to GTL. The outcome of this meeting will be the preparation of a report that describes the anticipated mathematical challenges in GTL in the fields of differential equations, stochastic methods, combinatorial methods, statistical analysis, optimization, and other relevant areas. DOE will use this report in the development of the GTL program. Similar reports are being produced by workshops on Computational Infrastructure, Computer Science, and Imaging, also being held during the first part of the 2002 calendar year.

Monday, March 18, 2002

8:15–8:30 a.m.	Welcoming Remarks, Workshop Logistics
8:30–9:00 a.m.	Gary Johnson and John Houghton: DOE Vision – OASCR and OBER
9:00 a.m.	Overview Presentations The presentations are meant to collectively give an overview of the mathematical needs for the Genomes to Life Program, emphasizing breadth instead of depth.
9:00–9:45 a.m.	DeWitt Sumners, FSU: Topological Models in Cellular Biology
9:45–10:30 a.m.	Steven Salzberg, TIGR: Genome Comparisons: Detecting Large-Scale Rearrangements and Single Nucleotide Polymorphisms
10:30–10:45 a.m.	Break
10:45–11:30 a.m.	John Guckenheimer, Cornell: Multiple Time scales in Dynamical Models

11:30 a.m.–12:15 p.m.	David Dixon, PNNL: Mathematical and Computational Needs for GTL—A Systems Biology Perspective
12:15–1:15 p.m.	Lunch: AMS Program Perspective
Lunch Speaker	Chuck Romine, OASCR, MICS
1:15–2:00 p.m.	Drew Endy, MIT: Math-Driven Experiments will Bring Genomes to Life
2:00– 4:45 p.m.	Breakout Sessions The purpose of the first set of breakouts is to identify math areas that are relevant to the GTL program and potentially can lead to breakthroughs.

Breakout Sessions

- Differential equations (ODEs, DAEs, PDEs, dynamical systems)
 - Stochastic methods
 - Combinatorial Methods (discrete methods, optimization)
 - Statistical Analysis
 - None of the above
- | | |
|----------------|--|
| 4:45–6:00 p.m. | Quick Summary Presentations from Breakout Groups |
| 6:00 p.m. | End of the First Day |

Tuesday, March 19, 2002

8:15–8:30 a.m.	Logistics
8:30–10:15 a.m.	General Discussions About 20 minutes per breakout group; provide a detailed summary report of the first day's breakout sessions. Open discussion.
10:15–10:30 a.m.	Break
10:30 a.m.	Breakout Discussions (same groups) The purpose of the second set of breakouts is to identify potential breakthroughs in biology over the next 10 years that might be enabled by advances in mathematics.
Lunch will be served.	
1:00–1:30 p.m.	Preparation of Summaries by Breakout Session Coordinators
1:30–3:30 p.m.	Summary Presentations from Breakout Groups
3:30–3:45 p.m.	Break
3:45–5:30 p.m.	Discussion on the Preparation of Report for DOE. The report should include both mathematics and biology perspectives. Volunteers from both fields are needed.
5:30 p.m.	Workshop Adjourns