

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

Report on the Computational Biology Workshop for the Genomes to Life Program

**U.S. Department of Energy
Germantown, Maryland
August 7–8, 2001**

Workshop Organizers

**Mike Colvin, Lawrence Livermore National Laboratory¹
Reinhold Mann, Oak Ridge National Laboratory²**

**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 2001

¹On detail to DOE Office of Biological and Environmental Research

²Now at Pacific Northwest National Laboratory

Table of Contents

Executive Summary	1
Introduction	4
Summary of Breakout Discussions	5
Summary of Overview Talks and Discussions	8
Appendices	13
A: Workshop Attendees, August 2001	15
B: Agenda	17
C: Genomes to Life Program Overview	19
Program Planning Workshops for Genomes to Life	Inside back cover

Report on the Computational Biology Workshop for the Genomes to Life Program

**U.S. Department of Energy
Germantown, Maryland
August 7–8, 2001**

Executive Summary

On August 7–8, 2001, a workshop attended by about 40 computational biologists, mathematicians, and computer scientists was held at U.S. Department of Energy (DOE) headquarters in Germantown, Maryland, to determine computational needs for the Genomes to Life (GTL) program. It is one in a series of program planning workshops being held to coordinate the program (see inside back cover). Readers who wish to comment on the contents of this report should send those comments to the workshop's organizers. This workshop had the following specific objectives:

- Translate the GTL goals into requirements for computational biology and identify existing resources relevant to these goals;
- Describe the current state-of-the-art capabilities in relevant computational and biological research areas;
- Identify needs for further development of computational methods, data repositories, data-analysis tools, and modeling and simulation of biological systems under the GTL umbrella;
- Identify high-performance computing infrastructure requirements to accomplish GTL goals; and
- Create a dialog between researchers in the computational and biological sciences.

To accomplish these objectives, the workshop addressed three broad topical areas:

- Biological Data Management, Analysis, and Access
- Computational Prediction of Structure, Function, and Interactions
- High-Level Modeling of Metabolic Pathways and Signaling Networks for Cells and Microbial Communities

These topics were addressed through invited presentations as well as lively discussions in breakout groups and in plenary sessions. The following findings and recommendations were derived from the workshop.

Summary Findings and Recommendations

DOE has a unique opportunity to bring to bear on modern biology its unparalleled experience base, expertise, and unique resources traditionally applied to other science and national security missions. The consensus of the workshop strongly supports DOE's objectives for the Genomes to Life program. DOE fulfills a unique role in this area of microbial research. Neither the private sector nor other federal agencies are positioned to develop the required tools and technologies.

Modeling of Cells and Microbial Communities

Findings

Achieving DOE programmatic goals in environmental remediation, carbon sequestration, and alternative energy feedstocks require integrated models and simulations of metabolic pathways, regulatory networks, and whole-cell functions. In the construction of cellular models, advanced software-development techniques will be necessary because these models are extremely heterogeneous. Relevant simulation levels range from that of individual molecules to molecular complexes, metabolic and signaling pathways, functional subsystems, individual cells, and ultimately cell communities (or organisms). Full-scale modeling and simulations will require petaflop capabilities, as well as a software environment and infrastructure that allow for integration of models at several spatial and temporal scales.

Recommendation

- DOE should support a program of research aimed at accelerating the development of high-fidelity models and simulations of metabolic pathways, regulatory networks, and whole-cell functions.

Biomolecular Simulations

Findings

For selected biological systems of high importance to GTL goals, there is a role for detailed molecular simulations of protein function and interactions. Analyzing protein interactions and the structure and workings of multiprotein complexes in such an organism will require petaflop-scale computing systems.

Recommendations

- DOE should ensure that advanced simulation methodologies and petaflop computing capabilities be available when needed to support full-scale modeling and simulations of pathways, networks, cells, and microbial communities.
- DOE should provide a software environment and infrastructure that allow for integration of models at several spatial and temporal scales.

Functional Annotation of Genomes

Findings

Computational methods will have a major role in the functional annotations of genomes, which is a necessary first step in developing higher-level models of cellular behavior. Significant methods development still is required to achieve the full promise of computational genome annotations. A sustained 2 to 5 teraflops of computing will be necessary for annotations to keep up with estimated rates of microbial sequencing in GTL.

Recommendation

- DOE should support the continued development of automated methods for the structural and functional annotations of whole genomes, including research into such new approaches as evolutionary methods to analyze structure/function relationships.

Experimental Data Analysis and Model Validation

Findings

Understanding functions of microbes and microbial communities depends critically on the ability to develop and validate models and drive simulations based on experimental data. Such analyses also will require breakthrough advances in mathematical methods and algorithms capable of incorporating experimental data produced by a variety of techniques, such as nuclear magnetic resonance, mass spectrometry, X ray, and neutron scattering.

Recommendations

- DOE should develop the methodology necessary for seamless integration of distributed computational and data resources, linking both experiment and simulation.
- DOE should take steps to ensure that high-quality, complete data sets are available to validate models of metabolic pathways, regulatory networks, and whole-cell functions.

Biological Data Management

Findings

Management, representation, analysis, integration, and accessibility of the enormous amount of GTL data are critical to the success of the program. GTL data span many levels of scale and dimensionality, including genome sequences, protein structures, protein-protein interactions, networks, pathways, multimodal molecular and cellular imagery, and complete cell models. Existing biological data repositories often are dispersed, heterogeneous, and isolated from one another—and also may contain data whose use is limited by intellectual-property restrictions.

Recommendations

- DOE should support the development of software technologies to manage heterogeneous and distributed biological data sets and the associated data-mining and -visualization methods.
- DOE should provide the biological data storage infrastructure and the multiteraflop-scale computing to ensure timely data updates and interactive problem solving.
- DOE should set a standard for open data in its GTL program and demonstrate its value through required universal use.

General Recommendations

In addition to the specific findings and recommendations above, workshop participants clearly felt that DOE should do the following:

- Continue the development of its GTL computational biology plan through a series of workshops focused on informatics, mathematics, and computer science challenges posed by the GTL systems biology goals.
- Ensure that the computing, networking, and data storage environment necessary to support the accomplishment of GTL goals will be available when needed. This environment should include computing capabilities scaling up through the multiteraflop and into the petaflop range, a storage infrastructure at the multipetabyte level, and a networking infrastructure that will facilitate access to heterogeneous distributed biological data sets by a geographically dispersed collection of investigators. Further definition of this environment should be pursued through a dedicated workshop.
- Establish policies for distribution and ownership of any data generated under the GTL program, prior to commencing peer review of GTL proposals or making any awards that would lead to the creation of such data.
- Support sufficient scope of research to assemble the cross-disciplinary teams of biologists, computational biologists, mathematicians, and computational scientists that will be necessary for the success of GTL.

Introduction

A workshop was held August 7–8, 2001, at the U.S. Department of Energy (DOE) headquarters in Germantown, Maryland, to initiate detailed planning of the computational biology research component of the Genomes to Life (GTL) program. It is one in a series of program planning workshops being held to coordinate the program (see inside back cover). This workshop was supported by DOE's Office of Advanced Scientific Computing Research and Office of Biological and Environmental Research. The goal of the workshop was to begin work towards the following objectives:

- Translate the GTL goals into requirements for computational biology and identify existing resources relevant to these goals;
- Describe the current state of capabilities in relevant computational and biological research areas;
- Identify needs for further development of computational methods, data repositories, data-analysis tools, and modeling and simulation of biological systems under the GTL umbrella;
- Identify high-performance computing infrastructure requirements to accomplish GTL goals; and
- Create a dialog among researchers in the computational and biological sciences.

The workshop included a diverse collection of scientists from DOE laboratories and other organizations (see Appendix A for a complete list of participants). The agenda (see Appendix B) was designed to facilitate discussions on the research and infrastructure needed to achieve the five computational biology aims stated in the GTL program roadmap (see Appendix C for an overview of the GTL program, including goals and expected payoffs):

Aim 1. Develop methods for high-throughput automated genome assembly and annotation

Aim 2. Develop computational tools to support high-throughput experimental measurements of protein-protein interactions and protein-expression profiles

Aim 3. Develop predictive models of microbial behavior using metabolic-network analysis and kinetic models of biochemical pathways

Aim 4. Develop and apply advanced molecular and structural modeling methods for biological systems

Aim 5. Develop the groundwork for large-scale biological computing infrastructure and applications

Breakout sessions during the second day, as well as lively discussions in plenary sessions, addressed three broad topical areas:

- Biological Data Management, Analysis, and Access
- Computational Prediction of Structure, Function, and Interactions
- High-Level Modeling of Metabolic Pathways and Signaling Networks for Cells and Microbial Communities

The above topics also were addressed through invited presentations. The breakout sessions and overview talks are summarized below. Specific findings and recommendations derived from these workshop sessions are presented in the Executive Summary at the beginning of this report.

Summary of Breakout Discussions

Based on the issues raised during the overview talks and discussions, three broad areas were chosen for more detailed analysis by three breakout groups. Each group, consisting of 10 to 15 people, met for 2 hours and then reported back to the entire workshop. These discussions are summarized below.

Biological Data Management, Analysis, and Access

This breakout group addressed an issue that emerged repeatedly during the workshop: the special challenge of data management in achieving the goals of Genomes to Life. A key component of GTL (and systems biology generally) is data integration, and there is a critical need for tools that allow biologists to derive inferences from massive amounts of heterogeneous and distributed biological data. The working group developed a long list of recommendations that provide a general framework for planning in this area that ranged from issues related to data sharing and ownership, to the computational hardware and communication bandwidth necessary to manage biological data. Most critically, this group emphasized that the challenges of data management and integration need to be addressed with high priority from the start of GTL.

Technically, GTL will need a flexible data framework because biology is moving at a fast pace. The types of data will be determined by experiments and also will impact infrastructure requirements. For this reason, the data-analysis and -storage strategies should be allowed to evolve over time in an organized and timely way. Despite this need for flexibility, the program needs a conceptually centralized integration repository—one portal to access data, with principles that define data interfaces.

The working group concluded that this data-management effort is too large to be independently solved within any single program. In particular, GTL should leverage the tools and intellectual output of SciDAC and other efforts in collaborative computing environments and scientific visualization. Investments are needed in integrated databases and new and improved algorithms that scale as the volume of data grows and the GTL program matures. However, the group also stressed that many of the issues in informatics for GTL will be solved by novel applications of existing techniques in computer science, mathematics, and statistics and will not always require fundamental research in these disciplines. For this reason, some mechanism is necessary to recognize and reward collaborative work among disciplines that primarily involves the transfer of established methods.

Finally, this subgroup concluded that a number of tasks in data management and annotation will have very large computational demands and, therefore, that high-performance computing resources must be available to the biology user community involved in data assembly, annotation, and curation. For some applications, compute-cycle requirements can be predicted, but for others the nature of the problems requires advancements in methods, so the algorithms and high-performance computing requirements are not yet clear.

Ultimately, the success of GTL will be judged by how well the program is accepted and serves groups within DOE and, just as importantly, the broader life sciences community. To achieve this success, the GTL program needs a new paradigm on data ownership in which the data is openly available.

Computational Prediction of Structure, Function, and Interactions

This subgroup focused on three aspects of GTL that will involve molecular-level simulations and prediction: high-throughput protein-structure prediction for genome functional annotation; integrated experimental and computational approaches to structures and function for hard-to-

isolate proteins and complexes; and, for a selected set of proteins and protein complexes critical to the GTL program, advanced molecular simulations of biochemical activity.

Prediction of protein function will involve the use of a number of methods, including structure prediction by comparative modeling and threading, “Rosetta”-type methods, and those based on phylogeny. All of these methods will need extensive further development to be applied automatically, especially for large, multidomain proteins. There also is a need for research into such new approaches as evolutionary methods to analyze structure/function relationships.

Another issue emphasized by the subgroup was that because all current methods for annotating structure and function require *finished* genome sequences, either resources must be devoted to completely finishing the genomes or computational approaches must be developed to effectively annotate unfinished sequences.

Whole-genome functional annotation will require significant computer resources. For example, estimates based on recent high-throughput protein-threading studies predict that a one-half-teraflop computer could thread 200 genes per day, so that threading of a whole bacterial genome would take from 2 to 4 weeks. Assuming that the GTL program will involve sequencing 20 bacteria per year, then 2 to 5 teraflops of sustained computing time will be required to keep up with that sequencing rate. Further, more advanced annotation methods will require significantly more computer resources, and there are certain types of protein structures (e.g., membrane-bound proteins), for which wholly new structure- and function-prediction methods will be necessary.

Ultimately, reliable, high-throughput determination of protein and protein-complex structures and functions will require computational methods capable of integrating several sources of experimental data, such as mass spectrometry (MS), protein arrays, crosslinking, nuclear magnetic resonance (NMR), and others. In many cases, even relatively sparse data can be used to derive constraints that speed up optimization approaches significantly and render them more accurate. High-throughput MS experiments involving complexes and crosslinkers pose significant informatics and computational challenges.

An important driver for high-performance computing systems will be modeling and simulation to predict the behavior of complexes for specific sets of proteins chosen from network analyses and other experiments. The computational requirements for such simulations are the best characterized among all of the areas of computational biology; moreover, many of these simulation methods are already implemented on teraflop-scale computers. Pure computing power is the major limitation on the size and accuracy of many biochemical simulations, which will involve data and models of protein-protein interactions, ligand-protein interactions, electron-transfer interactions, and membrane characteristics. Molecular dynamics (MD) and quantum mechanics-based molecular modeling will push high-end computing and require development of more effective scalable algorithms.

Finally, the subgroup emphasized the need for the GTL program to push the envelope for biophysical modeling, in particular, to develop the ability to predict the actual behavior of proteins and protein complexes for a selected set of biological processes chosen for their importance to GTL goals.

High-Level Modeling of Metabolic Pathways and Signaling Networks for Cells and Microbial Communities

The ultimate goal of such research would be physically complete models of a cell that would be developed based on a mix of empirical and computed data. Such models ultimately would be able to predict how a cell’s genome and environmental factors combine to yield its phenotype.

Models, therefore, would be powerful tools for both scientific discovery and the design of pathways or even whole microorganisms with novel capabilities. Such models have many drivers within the DOE mission areas, including environmental remediation, carbon sequestration, and alternative energy feedstocks.

The subgroup enumerated a series of specific scientific and engineering scenarios, including the engineering of modified *Deinococcus radiodurans* to clean up aromatic hydrocarbons in a radiation-intensive environment; elucidation of intercellular communication pathways in bacterial communities; and understanding the roles of cyanobacteria and diatoms for carbon sequestration. An ultimate culmination of such modeling methods would be the ability to automatically generate a complete description of a bacterium (as currently found in *Bergey's Manual*) using only DNA sequence data from an environmentally collected sample.

The subgroup emphasized that achieving predictive capabilities will require overcoming many technical challenges. For example, cell modeling involves a more complex collection of components and materials than existing models of climate or mechanical systems. Many of the developments needed involve research in computer science and mathematics. New mathematical methods are needed for analysis of raw biological data for inclusion in models and the subsequent statistical design of experiments to validate those models. As described in the previous section, there are major research challenges related to database query and database design in support of modeling, as well as the development of effective databases to capture modeling output and the models themselves.

To create these extremely heterogeneous cellular models, advanced software-development techniques will be necessary. Relevant simulation levels range from individual molecules to molecular complexes, metabolic and signaling pathways, functional subsystems, individual cells, and, ultimately, cell communities. Any general cell-level model will involve a variety of components and “subgrid” models. Effective abstractions are needed for multiple modeling hierarchies. The subgroup concluded that the actual simulations would involve the use of collections of “community” codes, requiring robust interfaces for component coupling. Ultimately, such models will be most effective when integrated into problem-solving environments for integrating experimental data required to determine simulation parameters and to validate simulation results. Finally, the simulation codes need to be scalable from desktops to the largest machines.

Computationally, no single architecture is appropriate for all aspects of predictive cell modeling. Whole-cell models will require tightly coupled parallel architectures, with smaller-component models running on workstations and whole-cell simulations on petaflop-scale systems. Whatever the form of the distributed computing infrastructure and data resources, they have to allow interactive access to both experimental groups and modeling groups. There may be a role for special-purpose hardware—for example, processors designed to allow very efficient integer operations.

Finally, the subgroup emphasized that a major issue in the development of such models is the interface between modeling and experiment. In particular, there will have to be a close coupling between the collection of cell data and its use in models, as well as validation of the models against very high quality experimental data sets.

Summary of Overview Talks and Discussions

A series of overview talks were presented with the goal of summarizing the current state of the art most relevant to the five aims of computational biology as stated in the GTL roadmap. Although these talks covered different topics (see summaries below), there were a number of common issues that surfaced in all the presentations and subsequent discussions. Most prominent were the

issues of data integration, data mining, derivation of knowledge from diverse data sources, data management, and synthesis of information from a large number of scientific publications.

Aim 1. Develop Methods for High-Throughput Automated Genome Assembly and Annotation

Genome assembly relies on mature approaches and algorithms. Current implementations can assemble whole mammalian genomes in a matter of tens of hours or less, using currently available computers. A recent assembly of all current public mouse genome sequences (approximately $2.7\times$ coverage) took 8 hours using the NERSC Phase II system. Continuing development needs to be done in the area of highly repetitive sequence domains, and with respect to assembling sequences from mixtures of microorganisms.

Sequence annotation and comparative analyses across multiple genomes are recurring computational tasks that require a high-performance computing infrastructure to ensure that regular information updates are part of the most current annotation and to facilitate interactive exploratory genome analyses. For example, the genome analysis resource established at Oak Ridge National Laboratory (ORNL) is making extensive use of the computing resources at the ORNL Center for Computational Sciences, which include multiteraflop systems by IBM and Compaq. Annotation goes far beyond finding coding regions in genome sequences. Finding regulatory elements is an unsolved research problem in even the simplest genomes and is expected to involve significant computational and mathematical challenges. Some analysis of regulatory regions can be accomplished by large-scale genome comparisons.

There remain significant research challenges in high-level annotation, including assignment of functions to every gene found in whole-genome sequences. This is particularly difficult because the pathway databases are incomplete and the microbial genomes encode for metabolic pathways about which there is very little biochemical data. At this time, most of the genes found in new genomic sequences do not have assigned functions. Some functions can be inferred by computational structure determination and protein folding, but a wide range of research problems remains to be solved in this area. Challenges in large-scale genome annotation easily could outpace the development of high-performance computer hardware and the software environments for effectively using that hardware. Within the next 5 years, genome sequences likely are to be completed at rates 10 to 100 times the current pace. High-throughput analytical approaches, as well as the informatics capabilities to manage the data and information for easy access by the biological research community, will present significant research challenges in this time period.

Aim 2. Develop Computational Tools to Support High-Throughput Experimental Measurements of Protein-Protein Interactions and Protein-Expression Profiles

The presentation focused primarily on high-throughput analysis of gene-expression profiles and relatively less on protein expression or protein-protein interactions. An enormous amount of data is being produced by experiments involving microarrays of oligonucleotides, cDNAs, and proteins/antibodies—all involving various tissues, exposures, other experimental conditions, and time-course studies. There are challenges associated with data quality, statistical analysis, variability of assays, and, in general, data-set reproducibility. Several analysis methods have been applied to microarray data sets. Various clustering approaches, singular-value decomposition, and pattern-recognition methods including several classes of neural net-based methods have been used. All current approaches fail to integrate into the analysis the often-substantial body of pre-existing knowledge, and most fail to account for experimental errors.

The situation is similar for the analysis and management of other types of biological data, such as mass spectral expression data or yeast two-hybrid data on protein-protein interactions. For all high-throughput experimental methods in biology, significant work is required to develop the

tools for statistical analysis, interpretation, annotation, and curation of the data. Furthermore, the full promise of these experimental methods will be achieved only if methods can be developed for integrating the different data types. For example, MS, NMR, and crystallography generate complementary data on proteins and complexes that, if integrated appropriately, can have significant impact on accomplishing the stated goals of the GTL program.

Aim 3. Develop Predictive Models of Microbial Behavior Using Metabolic-Network Analysis and Kinetic Models of Biochemical Pathways

One of the ultimate goals of the GTL program is predictive modeling of microbes and microbial communities. This presentation described the current state of the art for data-driven approaches to deriving metabolic networks from “parts lists” of enzymes involved in the pathways. The approach presented involves subjecting metabolic networks to known constraints that lead to descriptions of a solution space that shows how and under what conditions and particular biochemical behavior the reactions will occur. Constraints include capacity, maximum flux, connectivity, systemic stoichiometry, and physical/chemical factors (e.g., osmotic pressure, enzyme kinetics, and regulation). The red blood cell metabolic network was presented as an example, with 32 reactions, 29 external signals, and 19 metabolites. Recent work also has shown that using genome data and other information to predict many of the characteristics of *Saccharomyces cerevisiae* is possible. Shifts in gene-expression profiles can be predicted with 75% to 80% accuracy.

As this systems-level approach to understanding bacterial systems develops, several questions must be addressed. What are the biological design variables? Can biological systems be modeled in the same detail as physical/chemical systems? How do physical/chemical principles and approximations developed for modeling nonliving systems apply to the simulation of living systems? Are numerical values for parameters, such as enzyme-catalyzed reaction rates, known, or even knowable, since such properties change with time and environmental conditions, and from individual to individual?

Remaining challenges include the incorporation of kinetics and regulatory controls in current modeling approaches. Some molecules, including certain proteins and chemical signals, occur in such small numbers in the cell that they cannot be described accurately in terms of continuous concentrations. Instead, they must be described using discrete numbers of molecules, an approach that requires more complex mathematics and extensive statistical sampling; this approach is better simulated on novel architectures. Other challenges arise with questions of optimality criteria used in biological systems. For example, *Bacillus subtilis* is not optimized for growth while *Escherichia coli* does appear to be.

The discussions made clear that reaching the ultimate goal of predictively modeling such complex biological systems as cells requires many fundamental advancements, ranging from a better understanding of nonequilibrium processes, to the collection of complete data sets describing the properties of a cell. Hence, at this time, the limiting issue is not the availability of computational resources. Finally, the point was made that a significant amount of needed computing work actually requires integer arithmetic and rule-based systems. Participants recognized that vendors in the high-performance computing arena are unlikely to produce special-purpose hardware, but there may be opportunities to encourage vendors to optimize future processors for integer operations.

Aim 4. Develop and Apply Advanced Molecular and Structural Modeling Methods for Biological Systems

This talk described the current state of the art in the whole range of molecular-simulation methods, from the computational prediction of protein structure based on experimental data, to

first-principles simulations of biochemical processes. The presentation began with a description of the wide range of size and time scales involved in biological systems, pointing out that different simulation approaches would be appropriate at different levels of description. These methods include, at the highest levels, qualitative network analyses of biological pathways (e.g., with Petri nets) and quantitative network analysis (e.g., using the Monte Carlo approach), and range all the way down to molecular simulations of protein-protein interactions and quantum mechanical (QM) predictions of chemical reaction energies.

The talk included an overview of methods for predicting protein structures and also discussed many of the challenges to first-principles predictions of protein structure. These challenges include the long time scales (milliseconds to seconds) and very subtle energetics (often less than 10 kcal/mole) for protein folding. Nevertheless, empirically based methods including comparative modeling and “threading” often can successfully predict protein structure based on sequence similarities to proteins for which structures are known.

The talk went on to describe the two principal approaches to modeling biological processes at the molecular level. The most accurate are QM methods, which involve approximately solving the Schroedinger wave equation for the electronic motion of electrons in atoms and molecules. There is a large hierarchy of methods for solving the electronic Schroedinger equation, ranging from those that scale almost linearly in the number of atoms to much more accurate methods that scale as the seventh power of the number of atoms in the system. Although the best of these methods can achieve accuracies for energies and structures as good as or better than experimental methods, they are too computationally costly to be applied to most biochemical processes. Research is needed to develop versions of these methods that scale less steeply with system size, or to develop ways to empirically correct less costly methods.

The other approach to modeling molecular systems uses the much less accurate classical (ball-and-spring) force fields to describe the atomic interactions, but this method can be applied to much larger systems and much longer time scales. Such approaches include both MD, in which the motion in time of each atom is simulated, and Monte Carlo, in which a large ensemble of atomic configurations is randomly generated and sampled.

A continuing challenge is the long-time MD for slow events (actually involving multiple time scales). The issue of reaching macroscopic time scales from MD simulations cannot be solved solely by increases in hardware—the number of processors. Development of theoretically sound, time-coarsening methodologies is needed to permit dynamics-based methods for traversing much longer time scales. Another related high-priority research area is the development of improved force fields for MD, such as those that include polarization effects.

There are many areas of active research aimed at improving molecular-simulation methods. Promising emerging methods include mixed QM/molecular mechanics methods that may allow accurate QM methods to be applied only in the regions where they are necessary, such as in enzyme-active sites, while the larger system is modeled classically. Another area of active research is first-principles molecular dynamics simulation, which involves using a fully QM description of the atomic interactions and electronic structure calculations (Car-Parinello approaches). These methods have been demonstrated to yield extremely accurate properties for water, solvated ions, and very small biochemical systems, but they are limited computationally to very short time scales and system sizes.

The talk concluded by exploring the developments necessary to transform biology into a “systems science.” Systems biology as described in the GTL roadmap requires significant expertise and resources that cross traditional disciplinary boundaries. Also needed is the development of new theories and mathematics, as well as the development of new algorithms, their implementation on high-performance computer systems, and extensive use of large, distributed, and heterogeneous

databases with wide availability to make software and computer systems usable. Ultimately, this will lead to a new model for biological analysis that will involve a cycle beginning with the computational synthesis of available biological information to formulate specific biological hypotheses that will drive new experiments or, in some cases, specific computational simulations in place of experiments. The data from the new experiments will feed back into the next round of synthesis and hypothesis development.

Aim 5. Develop the Groundwork for Large-Scale Biological Computing Infrastructure and Applications

In addition to addressing requirements in terms of compute cycles and connectivity of computational resources for GTL, an important step is to address and resolve serious issues concerning data resources and access methods. The current state of the art in this arena for biology is less than desirable. There are a myriad of data silos and a few monolithic, asymmetric cross-references. A consequence of this poor data integration is the propagation of spurious information in databases; for example, there is the not-infrequent situation where gene A has a low level of similarity with gene B in another organism, and researchers find a gene similar to gene A and then claim it has the function of gene B. Many data resources have limited, idiosyncratic querying capabilities that are designed mostly for browsing human data. There are no third-party annotation mechanisms in common use. The distributed annotation system effort (<http://stein.cshl.org/DAS>) under development shows great promise to remedy this deficiency. There is a lack of accepted standards for defining, querying, and transmitting common data objects nor are there effective strategies for discouraging data hoarding (delayed releases of data are not uncommon).

GTL will span the entire range of genomics—including sequence, proteins, expression function, and pathways—and the resolution of the data problems outlined above is paramount to the success of GTL. Scaling is a huge challenge for GTL, but scaling of data volume is only one part of the problem. An equally difficult challenge will be the seamless integration of such data resources as genomic sequence, protein analysis, genomic and protein expression arrays, and pathway information. Accomplishing the scaling among multiple laboratories will be even harder. Integration in the field of genomics is historically spotty at best, and GTL will bring in different disciplines, each with its own agenda.

“What databases does GTL need to build?” This is not the problem as much as a real need to establish a free and level market for data, so that GTL has a chance to scale and succeed. With such a free market for data, open competition could establish the needed data resources and integration. Free-market design principles for GTL data resources should include:

- Establishment of a common data-release policy for all GTL program awards;
- Equal compliance by all awardees with policies concerning data release/sharing;
- Development of common ontologies for basic GTL objects;
- Establishment of common, low-level data-interchange methods;
- Establishment of a common set of GTL URLs to allow automated query access by all GTL sites; and
- Definition of only the basic data objects, interchange methods, and access methods, with the market providing all higher forms of integration.

Data integration should be competed openly, not with the establishment of monolithic sites. GTL should provide grants for information-integration services and tools, and it should actively participate in genomics standards/integration efforts in the larger community. Traditional integration methods may have merit for some aspects of GTL: language-based approaches; flat

file, text retrieval, and search engines; data federation and distributed databases; classical data warehousing; centralization; and web robots/agents. Each method will benefit from free-market data access.

Collaboratories and computational grids collect resources under a common set of middleware. The details of specific distributed resources are not apparent. Biology already has grids that come from a natural method of scientific investigation (i.e., inference from many data sources and analyses). However, the biology community neglected to use computer science terminology for this environment. An explicit GTL grid would encompass data and computational resources as well as collaboration technologies. Common technologies would enable annotation jamborees and other intensely interactive and computer-enabled biological investigations without scientists having to be physically at one site. A GTL grid would include several experimental devices, such as mass spectrometers, NMR systems, light and neutron sources, and other experimental facilities. This grid would tightly couple the experimentalists with computational experts and resources.

Application software infrastructure is equally important. The GTL program should create a free market for GTL software, with open sources and access available to consortium members.

The computer science community believes petaflop machines will be possible and personal teraflop machines will be available in the next 5 years. The amount of computing machinery that will be available as distributed resources will be amazing. As the GTL program develops, the next generation of computers—possibly with hundreds of thousands, or millions, of processors—will become operational. Biologists like other user communities will face problems related to algorithms that will scale to petaflops. The problem of systems integration will become more important than in any biology program before GTL. The Defense Advanced Research Projects Agency has been dealing with issues of systems integration in several of its programs (e.g., Bio-Spice and Image Understanding). This paradigm is worth evaluating for GTL. Moreover, with a revolution in broadband networking expected over the next 5 years, raw, long-haul bandwidth may not be a limiting factor for the success of GTL.

Several developments are under way with respect to standards. For example, researchers at the University of Washington and the California Institute of Technology are writing CellXML to simulate cell functions. The successful example of the U.S. Department of Defense enforcing a hardware design standard indicates that an agency can make a huge difference toward developing a culture of interoperability. GTL needs to be more than the sum of independent, lab-centric projects bolted together. DOE could impact significantly a set of interoperability standards for the biology community. GTL's chances for success will be seriously compromised if its informatics and computational biology infrastructure is not treated as a first-class component of the program from the beginning.

Appendices

Appendix A: Workshop Attendees, August 2001

Appendix B: Agenda

Appendix C: Genomes to Life Program Overview

Appendix A: Workshop Attendees, August 2001

#	Last	First	Affiliation	E-mail
1	Anderson	Carl	BNL	cwa@bnl.gov
2	Arkin	Adam	LBNL	aparkin@lbl.gov
3	Bayer	Paul	DOE	paul.bayer@science.doe.gov
4	Branscomb	Elbert	LLNL	ewbranscomb@lbl.gov
5	Cary	Robert	LANL	rb Cary@lanl.gov
6	Colvin	Mike	LLNL	colvin2@llnl.gov
7	Critchlow	Terence	LLNL	critchlow1@llnl.gov
8	Daniel	Hitchcock	DOE	Daniel.Hitchcock@science.doe.gov
9	David	Thomassen	DOE	David.Thomassen@science.doe.gov
10	Dixon	David	PNNL	david.dixon@pnl.gov
11	Dunning	Thom	NC SCC	thom.dunning@ncsc.org
12	Fidelis	Krzysztof	LLNL	fidelis@llnl.gov
13	Frazier	Marvin	DOE	Marvin.Frazier@science.doe.gov
14	Geist	Al	ANL	gst@ornl.gov
15	Gilna	Paul	LANL	pgil@lanl.gov
16	Heffelfinger	Grant	SNL	gsheffe@sandia.gov
17	Houghton	John	DOE	john.houghton@science.doe.gov
18	Johnson	Gary	DOE	garyj@er.doe.gov
19	Knotek	Mike	DOE	m.knotek@gte.net
20	Larimer	Frank	ORNL	larimerfw@ornl.gov
21	Locascio	Phil	ORNL	locasciop@ornl.gov
22	Lubeck	Olaf	LANL	olubeck@lanl.gov
23	Makowski	Lee	ANL	lmakowski@anl.gov
24	Maltsev	Natalia	ANL	maltsev@mcs.anl.gov
25	Mann	Reinhold	ORNL	mannrc@ornl.gov
26	Mansfield	Betty	ORNL	mansfieldbk@ornl.gov
27	Marvin	Stodolsky	DOE	marvin.stodolsky@science.doe.gov
28	Melius	Carl	LLNL	melius1@llnl.gov
29	Oliver	Ed	DOE	Ed.Oliver@science.doe.gov
30	Palsson	Bernhard	UCSD	palsson@ucsd.edu
31	Patrinos	Ari	DOE	Ari.Patrinos@science.doe.gov
32	Rokhsar	Dan	JGI	DSRokhsar@lbl.gov
33	Samatova	Nagiza	ORNL	samatova@cs.utk.edu
34	Selkov	Evgeni	ANL	selkov@megapathdsl.net
35	Simon	Horst	LBNL	HDSimon@lbl.gov
36	Slezak	Tom	LLNL	slezak@llnl.gov
37	Stevens	Rick	ANL	stevens@mcs.anl.gov
38	Stevens	Walt	DOE	Walter.Stevens@science.doe.gov
39	Trewhella	Jill	LANL	jtrewella@lanl.gov
40	Uberbacher	Ed	ORNL	ube@ornl.gov
41	Wiley	Steve	PNNL	Steven.Wiley@pnl.gov
42	Wooley	John	UCSD	jwooley@ucsd.edu
43	Worley	Brian	ORNL	wor@ornl.gov

Appendix B

Agenda Computational Biology Workshop Genomes to Life Program

August 7–8, 2001
Room A-410, DOE Germantown Headquarters
Germantown, Maryland

Tuesday, August 7, 2001

Goals and Computational Needs of the GTL Program

Moderator: John Wooley, UCSD

8:00 – 9:00	Arrival, Badging, Coffee and Pastries	
9:00 – 9:15	Welcome and Introduction	Oliver, Patrinos
9:15 – 9:30	Review of Workshop Goals and Agenda	Johnson, Colvin, Mann
9:30 – 9:45	Overview of presentations	Wooley
9:45 – 10:15	High-throughput automated genome assembly and annotation	Rokhsar
10:15 – 10:45	Discussion	Uberbacher
10:45– 11:00	Break	
11:00 – 11:30	Analysis of protein-protein interactions and protein-expression profiles	Cary
11:30 – 12:00	Discussion	Branscomb
12:00 – 1:00	Lunch	
1:00 – 1:30	Predictive models of microbial behavior, and models of biochemical pathways	Palsson
1:30 – 2:00	Discussion	Wiley
2:00 – 2:30	Advanced molecular and structural modeling methods for biological systems	Dixon
2:30 – 3:00	Discussion	Heffelfinger
3:00 – 3:15	Break	
3:15 – 3:45	Large-scale biological computing infrastructure	Slezak
3:45 – 4:15	Discussion	Stevens
4:15 – 5:00	Charge to Breakout Groups	Johnson, Colvin, Mann
5:00	Adjourn	

Wednesday, August 8, 2001

GTL Computational Research Priorities and Infrastructure Needs

Moderator: Thom Dunning, NC SCC

8:30 – 10:30 Breakout Group Discussions

Topics:

1. Data analysis, management, validation, representation and integration, (e.g. genome annotation, expression array analysis)
2. Metabolic pathway reconstruction and simulations, and modeling cells and cell communities
3. Methods predicting macromolecular structure, function, and interactions (including support of experimental. methods) Guideline questions:
 - 1) What are the key next steps in this area to reaching the GTL goals?
 - 2) What are the highest research priorities in this field?
 - 3) What are the major technical roadblocks to achieving this goal?
 - 4) How far is the goal from the current state-of-the-art?
 - 5) What is the research effort necessary to achieve it?
 - 6) What is the mix of research disciplines needed to reach this goal?
 - 7) What other agencies and companies are sponsoring closely related research?

10:30 – 10:45 Break

10:45 – 12:00 Reports back from breakout groups

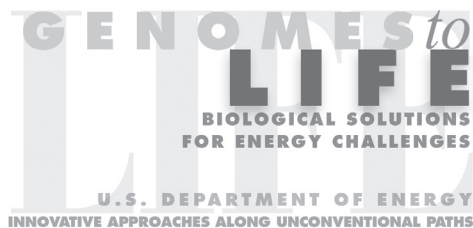
12:00 – 1:00 Lunch

1:00 – 2:00 Discussion of mathematics and computational research priorities for GTL

2:00 – 3:00 Discussion of GTL computational infrastructure requirements

3:00 – 3:30 Wrap-up

3:30 Adjourn



Program Overview



Office of Science

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

December 2001

Built on the continuing successes of international genome-sequencing projects, the Genomes to Life program will take the logical next step: a quest to understand the composition and function of the biochemical networks and pathways that carry out the essential processes of living organisms. The roadmap published in April 2001 sets forth an aggressive 10-year plan designed to exploit high-throughput genomic strategies and centered around the four major goals outlined in the chart at right.

The Genomes to Life program reflects the fundamental change now occurring in the way biologists think about biology, a perspective that is a logical and compelling product of the Human Genome Project (HGP). The new program will build on HGP achievements, both by exploiting its data and by extending its paradigm of comprehensive, whole-genome biology to the next level. This approach ultimately will enable an integrated and predictive understanding of biological systems—an understanding that will offer insights into how both microbial and human cells respond to environmental changes. The applications of this next level of understanding will be revolutionary.

The current state-of-the-art instrumentation and computation enable and encourage the immediate establishment of this ambitious and far-reaching program. The strategic alliance created between DOE's offices of Advanced Scientific Computing Research (ASCR) and Biological

and Environmental Research (BER) will develop the infrastructure to meet these challenges. Concurrent technology development also will be needed to reach all goals within the



next decade. Substantial efforts will be devoted, for example, to improving technologies for characterizing proteins and protein complexes, localizing them in cells and tissues, carrying out high-throughput functional assays of complete cellular protein inventories, and sequencing and analyzing microbial DNA taken from natural environments.

The Genomes to Life program complements and augments the Department of Energy's (DOE) Microbial Cell Project, launched in FY 2001. The goal of this established project is to collect, analyze, and integrate data on individual microbes in an effort to understand how cellular components function together to create living systems, particularly those with capabilities of interest to DOE.

DOE is strongly positioned to make major contributions to the scientific advances promised by the biology of the 21st century. Strengths of DOE's national laboratories include major facilities for DNA sequencing and molecular structure characterization, high-performance computing resources, the expertise and infrastructure for technology development, and a legacy of productive multidisciplinary research essential for such an ambitious and complex program. In the effort to understand biological systems, these assets and the Genomes to Life program will complement and fundamentally

Genomes to Life Program

GTL was developed in response to a 1999 charge by the DOE Office of Science to the Biological and Environmental Research Advisory Committee to define DOE's potential roles in post-HGP science. The resulting August 2000 report, *Bringing the Genome to Life*, set forth recommendations that led to the roadmap published in April 2001. The FY 2002 budget for GTL is \$19.5 million.

GTL Publications

Documents, meeting reports, and image gallery are downloadable via the Web:



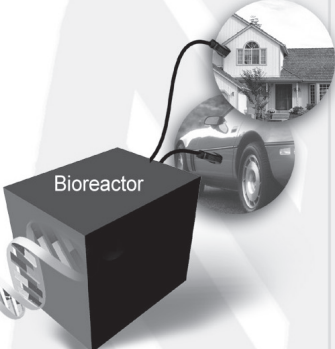


- DOEGenomesToLife.org

Requests for future publications:

- Human Genome Management Information System
865/576-6669, Fax: /574-9888
mansfieldbk@ornl.gov

enable the capabilities and efforts of the National Institutes of Health, the National Science Foundation, and other agencies and institutions around the world.

P A Y O F F S F O R T H E N A T I O N

<i>Human Health Protection</i>	<i>Energy Security</i>	<i>Environmental Cleanup</i>
 <p>Enhance bioterror agent detection and response</p>  <p>Clarify human susceptibility to energy-related materials</p>	 <p>Enable U.S. energy security</p> <ul style="list-style-type: none"> • Launch major new American industry in bioenergy 	 <p>Stabilize atmospheric carbon dioxide to counter global warming</p>  <p>Save billions of dollars in toxic waste cleanup and disposal</p>

Office of Advanced Scientific Computing Research • Office of Biological and Environmental Research

Program Planning Workshops for Genomes to Life

A series of program planning workshops is being held to coordinate Genomes to Life. Several took place in 2001, and others are anticipated for 2002. Meeting reports are placed on the Web as soon as they become available (DOEGenomesToLife.org). To learn more about the program, please see the Web site or use the [contact information](#) for Marvin Frazier or Gary Johnson.

2001 GTL Workshops

- June 23 Role of Biotechnology in Mitigating Greenhouse Gas Concentrations
- August 7-8 Computational Biology
- September 6-7 Computational and systems Biology: Visions for the Future
- December 10-11 Mass Spectrometry Technologies

2002 GTL Workshops (all dates subject to change)

- Jan 22-23 Computing and Networking Infrastructure
- March 6-7 Computer Science
- March 12-13 Mathematics
- Early 2002 Imaging Technologies