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.