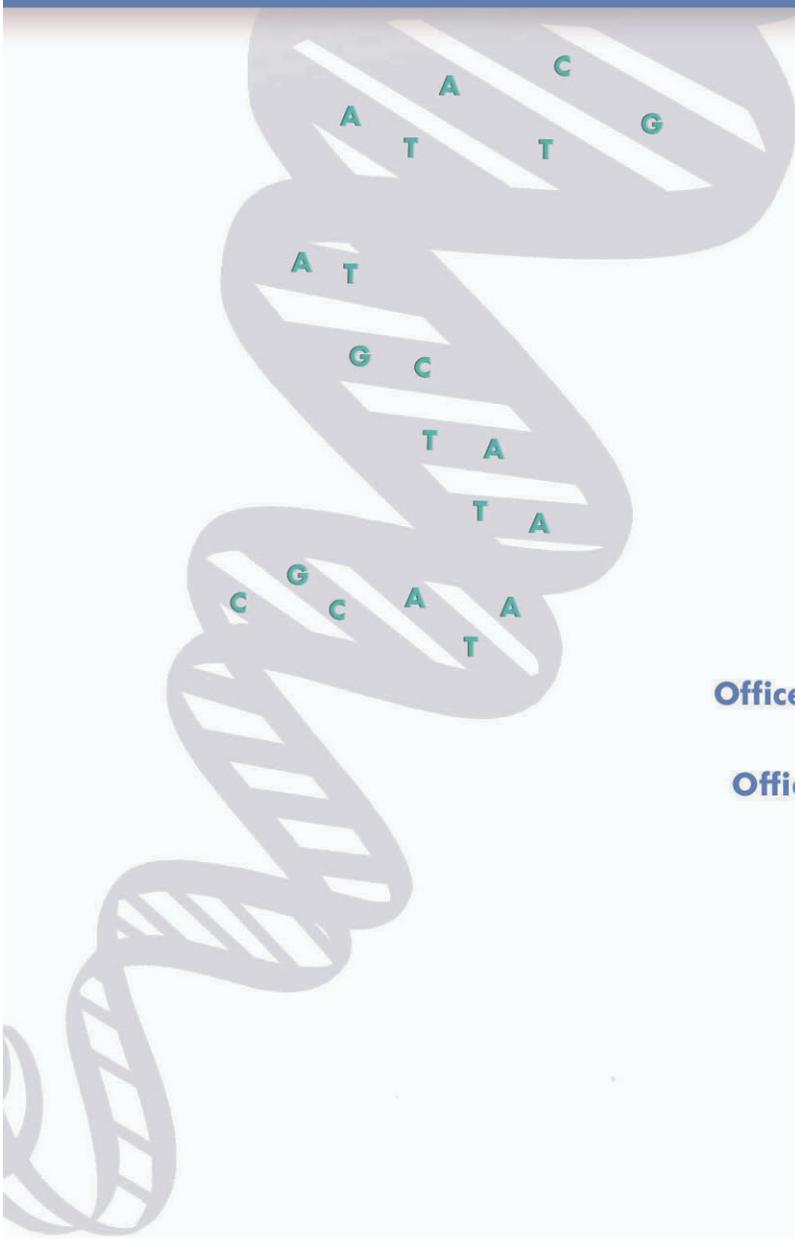




GENOMES *to* LIFE

**BIOLOGICAL SOLUTIONS
FOR ENERGY CHALLENGES**

**U.S. DEPARTMENT OF ENERGY
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS**



Report on Three Genomes to Life Workshops: Data Infrastructure, Modeling and Simulation, and Protein Structure Prediction

**U.S. Department of Energy
Gaithersburg, Maryland
July 22–24, 2003**

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
January 2004**

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





Genomes to Life Program

Contacts for program information

Gary Johnson

U.S. Department of Energy (SC-30)
Office of Advanced Scientific Computing Research
970/225-3794, Fax: 970/223-1415
garyj@er.doe.gov

Marvin Frazier

U.S. Department of Energy (SC-72)
Office of Biological and Environmental Research
301/903-5468, Fax: 301/903-8521
marvin.frazier@science.doe.gov

Publications

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

- <http://DOEGenomesToLife.org>

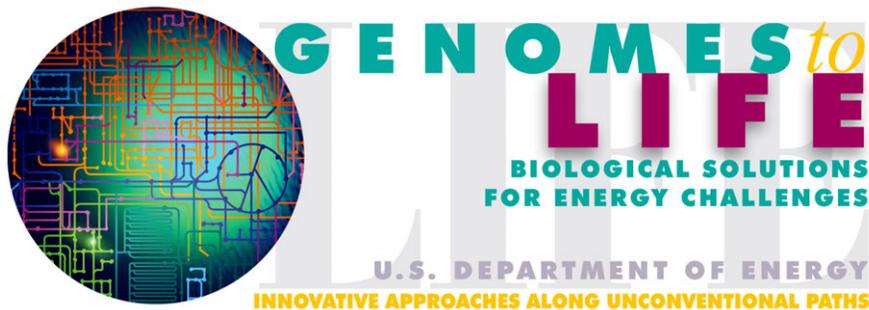
Copies of this report and other Genomes to Life publications:

- Human Genome Management Information System
Oak Ridge National Laboratory
1060 Commerce Park, MS 6480
Oak Ridge, TN 37830

865/576-6669, Fax: 865/574-9888, mansfieldbk@ornl.gov

Electronic versions of this report and additional information:

- <http://doegenomestolife.org/compbio/>



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

Report on Three Genomes to Life Workshops: Data Infrastructure, Modeling and Simulation, and Protein Structure Prediction

**U.S. Department of Energy
Gaithersburg, Maryland
July 22–24, 2003**

Workshop Organizers

**Al Geist and Thomas Zacharia, Oak Ridge National Laboratory
Reinhold Mann and George Michaels, Pacific Northwest National Laboratory
Grant Heffelfinger, Sandia National Laboratories**

**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**



Published January 2004



EXECUTIVE SUMMARY	1
RECOMMENDATIONS	2
REPORT ON THE GENOMES TO LIFE DATA INFRASTRUCTURE WORKSHOP	3
INTRODUCTION	3
SUMMARY OF TALKS AND DISCUSSIONS	4
DATA STANDARDS AND INTEGRATION	4
DATA-MANAGEMENT INFRASTRUCTURE	6
DATA-QUALITY CONTROL	7
DATA-ANALYSIS INFRASTRUCTURE	8
WORKFLOW ENVIRONMENTS FOR DATA COLLECTION	9
TRANSPARENT ACCESS TO DATA AND COMPUTATIONAL RESOURCES	10
CONCLUSIONS	11
REPORT ON THE GENOMES TO LIFE MODELING AND SIMULATION WORKSHOP	12
INTRODUCTION	12
USES OF MODELING AND SIMULATION TO ACCOMPLISH GTL'S GOALS	13
MOLECULAR SIMULATIONS OF PROTEIN FUNCTION AND MACROMOLECULAR INTERACTIONS	13
SIMULATION AND MODELING OF CELLULAR BIOCHEMISTRY	13
DEVELOPMENT OF BETTER QUALITATIVE METHODS	13
SUMMARY OF TALKS AND DISCUSSIONS	14
GENERAL INFRASTRUCTURE NEEDS	14
RECOMMENDATIONS	15
REPORT ON THE GENOMES TO LIFE PROTEIN STRUCTURE PREDICTION WORKSHOP	17
INTRODUCTION	17
STATE OF THE ART	18
GOALS AND CHALLENGES	20
BIOLOGY ISSUES	20
MATH AND COMPUTING SCIENCE ISSUES	21
OTHER ISSUES	22
APPENDIX A: WORKSHOP ATTENDEES	23
JULY 22, 2003—DATA INFRASTRUCTURE WORKSHOP	23
JULY 23, 2003—MODELING AND SIMULATION WORKSHOP	25
JULY 24, 2003—PROTEIN STRUCTURE PREDICTION WORKSHOP	27
APPENDIX B: WORKSHOP AGENDAS	28
JULY 22, 2003—DATA INFRASTRUCTURE WORKSHOP	28
JULY 23, 2003—SIMULATION AND MODELING WORKSHOP	29
JULY 24, 2003—PROTEIN STRUCTURE PREDICTION WORKSHOP	30

Executive Summary

On July 22, 23, and 24, 2003, three one-day workshops were held in Gaithersburg, Maryland. Each was attended by about 30 computational biologists, mathematicians, and computer scientists who were experts in the respective workshop areas. The first workshop discussed data infrastructure needs for the Genomes to Life (GTL) program with the objective of identifying current gaps and defining the infrastructure required for the success of the proposed GTL facilities. The second workshop discussed modeling and simulation needs for the next phase of the GTL program and defined how these relate to experimental data generated by genomics, proteomics, and metabolomics. The third workshop identified emerging technical challenges in computational protein structure prediction for DOE missions and outlined specific goals for the next phase of GTL. The workshops were attended by representatives from both the Office of Biological and Environmental Research and the Office of Advanced Scientific Computing Research.

Invited experts at each workshop made short presentations on what they perceived as key needs in GTL data infrastructure, modeling and simulation, and structure prediction, respectively. Each presentation was followed by a lively discussion by all attendees. The following findings and recommendations were derived from the three workshops.

A seamless integration of GTL data spanning the entire range of genomics, proteomics, and metabolomics will be extremely challenging but must be treated as the primary component of the GTL program to ensure GTL's chances for success. High-throughput GTL facilities and ultrascale computing will make it possible to address the ultimate goal of modern biology: To achieve a fundamental, comprehensive, and systematic understanding of life. But first the GTL community needs to address the massive quantities and increased complexity of biological data produced by experiments and computations. Genome-scale collection, analysis, dissemination, and modeling of data are key to GTL's success. Localizing these activities within each experimental facility that generates data will ease integration and organization. However, integration and coordination of these activities across the facilities will be extremely critical to ensure high-throughput knowledge synthesis and engage the broader biology community. Ultimately, the success of the data infrastructure will be judged by how well it is accepted by and serves the biology community.

Recommendations

- DOE should lay the groundwork for a GTL data infrastructure composed of a distributed but integrated suite of facilities databases, through the cooperative development of data models and database schemas. This process should seek to identify shared or common data elements, objects, concepts, and identifiers that can lead to metadata types that are sharable across existing GTL projects and future facilities.
- Data-management issues of the infrastructure need to be addressed from the start of GTL. Among those issues are the types of GTL-generated data, support for long-term data curation, data-quality control, mechanisms for accessing data for analysis, and standardized ways of disseminating data to the GTL community.
- The data infrastructure needs to be flexible to allow data analysis and storage strategies to evolve over time in an organized and timely way.
- A data-analysis framework should be part of the data infrastructure and provide transparent access to distributed data sources, analysis tools, and computational resources across the GTL community. The framework should include tools for testing the coherency of disparate bodies of data and allow individual sites to customize data-analysis tools and available databases to match their research needs.
- Mathematical models are needed that are ultimately developed from fundamental biological principles. These models must be tested and verified through integrated wet-lab experimentation using multiple analytical methods and based on well-characterized statistical designs.
- Working groups should be established to define modeling and simulation data and experimentation requirements for validation equivalent to a Critical Assessment of Protein Structure Prediction (called CASP) competition for systems biology.
- Stable, production-oriented high-performance computing capabilities should be established for long time-scale biological modeling and simulation computational experiments.

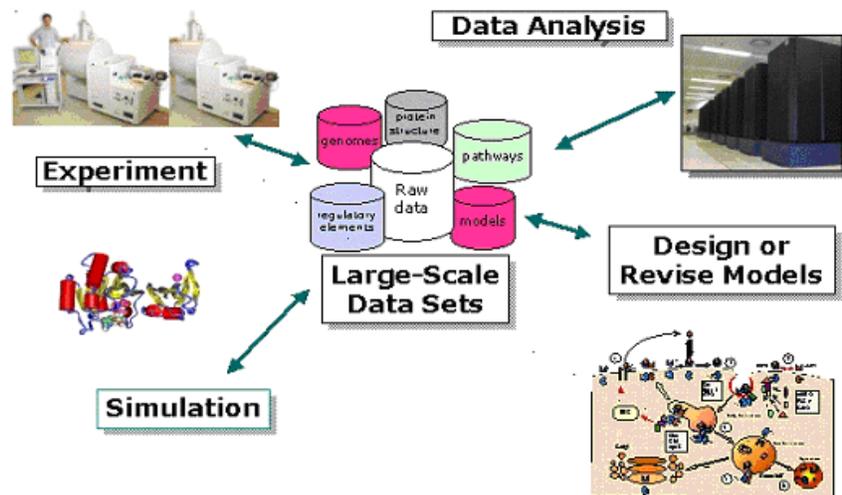


Fig. 1. GTL requires a new synergy between computing and biology, with data at the center of it all.

Report on the Genomes to Life Data Infrastructure Workshop

July 22, 2003

Organizers: Al Geist and Thomas Zacharia
Oak Ridge National Laboratory

Introduction

A workshop was held July 22, 2003, in Gaithersburg, Maryland, to identify the needs and gaps in the existing Genomes to Life (GTL) data infrastructure and to suggest short- and long-term actions to close these gaps. The meeting, supported by DOE's Office of Advanced Scientific Computing Research and Office of Biological and Environmental Research, 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 data research and infrastructure needed to achieve the long-term goals of the GTL program.

The Genomes to Life facilities plan and previous workshop reports place considerable emphasis on developing methods for a large community of biologists to analyze large, distributed biological data sets and develop models and simulations related to complex biological phenomena. They stress the need for integrated approaches to software and hardware infrastructure to accomplish these objectives. An organized approach to coordination and planning in computing will guide data standards, data management, large-scale development of analysis tools, implementation, and support of analysis on specialized hardware environments, including massively parallel computers and distributed grid systems.

To facilitate wide usage of GTL infrastructure in computing, very simple user environments must be created that "know" where to get the necessary data and where an application should run, based on availability and best use of resources, without the user having to specify these details.

Because of the very distributed nature of biology and the biological databases, no one site can hope to cover all the needs of this new science frontier. Some 335 genomic and molecular biology databases currently are distributed around the country, with large quantities of data being added daily.

The data problem is getting much worse as proteomics data are generated from arrays and mass spectrometers. Whereas the genome is static, proteomics data are dependent on time and on initial conditions. Much more experimental biological data (e.g., about conditions)

must be carried with the proteomics data. In addition, proteomics data often have a qualitative part that also must be available (e.g., raw visual data from microarrays). The amount of data that will be generated from microbial community studies promises to be staggering. The data-storage infrastructure and formats for biological data must be established, and the data itself need to be stored and made available to the GTL and broader biological communities. Thus, investments in the data-storage and -access infrastructure need to be made early in the program schedule so GTL facilities and individual experimental and computational groups can benefit from an integrated storage infrastructure and the related analysis tools.

Summary of Talks and Discussions

GTL has laid out an ambitious plan to combine genomic data and high-throughput experimental technologies with ultrascale computing resources to study proteins encoded by the genome throughout the organism's life cycle. A series of facilities will be created that will produce biological data on an unprecedented scale. The GTL program will enable an extensive assembly of experimental and computational devices, including numerous mass spectrometers, imaging devices (X ray, electron, neutron scattering), and complex mixtures of the biophysical characterization devices and tools such as gel electrophoresis, nuclear magnetic resonance (NMR) imaging, various binding assays, and protein chips.

The scale, high-throughput operation, and diversity of types of data and associated analyses present an extreme challenge to traditional approaches to genomic data processing. At present, most software on which analysis pipelines are built is relatively inflexible and not designed for use in a high-throughput environment. In other words, tools for analysis, modeling, and simulation are not adapted readily to variations in processing dictated by the availability of multiple types of experimental data. Moreover, these tools are not designed to function in the distributed computational environment that will be required to support GTL needs, mixing large databases, stand-alone and parallel computers, and remote resources.

Data Standards and Integration

Data integration always has been a foster child of bioinformatics. As a result, integration in the field of genomics is historically spotty at best, with a few monolithic and asymmetric cross-references. A consequence of this poor integration is the propagation of unreliable, incomplete, and noisy information in databases. Many data resources use their own data formats and custom interfaces; navigating between sources and transferring data between interfaces is usually more complicated than a simple mouse click or cut-and-paste operation. The situation is getting much worse, with technological advances that allow data to be created from the wet lab at an ever-increasing rate and with the growing need to combine these data in new and interesting ways.

Core requirements of any large-scale production enterprise such as GTL are the management, manipulation, integration, and presentation of the data. With unique scientific challenges associated with each of the GTL projects and experimental facilities, a centrally

located data infrastructure will not be possible, due in large part to the distinct research agenda. A seamless integration of GTL data spanning the entire range of genomics, proteomics, and metabolomics will be extremely challenging but must be treated as the primary component of the GTL program to ensure GTL's chances for success.

The GTL data-integration enterprise should attempt to lay the groundwork for a distributed but integrated suite of facilities databases through the cooperative development of data models and database schemas. This process should seek to identify shared or common data elements, objects, concepts, and identifiers that can lead to types of metadata that are sharable across the GTL projects and facilities. Independent systems thus can evolve at the highest level to meet effectively the local conditions of domain experts while sharing a common intellectual layer of process and information. This will permit the unique knowledge acquired at each facility to be used across the DOE complex and eventually allow users to mine data from the combined sites.

A key goal of the GTL distributed experimental collaboration is to achieve a common frame of reference (data standards) for both experimental observations of biological phenomena and the representative counterparts within the data model. There is a need to develop a framework of both controlled vocabularies and common ontological definitions of basic GTL objects as well as low-level data-interchange and access methods to permit the experimental facilities to communicate effectively. Furthermore, this framework will allow the development of complex inferential knowledge based on the wealth of experimental data, the construction of data-driven components of large-scale biological modeling and simulation efforts, and effective data-mining tools for GTL data resources.

Recommendations include finding common data needs and patterns among GTL projects, thus leveraging from existing GTL projects and using ongoing biology programs in the community to start the definition phase now and work towards solutions that are capable of evolving.

The recommended plan is to build on what exists to provide useful tools from the beginning and offer analysis end users with familiar interfaces. To ensure that real requirements are met, the plan is that each GTL facility would produce one or two use cases of biology questions to be answered. Biologists in the GTL teams would generate the questions and work on potential solutions jointly with computer scientists.

The GTL data infrastructure should aim to support the following data-standards creation tools:

- Schema-description tools with domain-specific schemas (e.g., lab experiments, microarrays, pathways) as well as standard schemas whenever possible (e.g., MIAME).
- Database-federation tools to use data from multiple independent databases.
- Schema-evolution tools for rapid prototyping of new data types and data transformations.
- Nonstandard data formats including sequences, graphs, three-dimensional structures, and images.

- Data-format interchange by utilizing standard format technology (e.g., XML) as well as schema-interchange tools (e.g., XML translators).
- Operations (e.g., equality, range, and imprecise operators) over nonstandard data, including sequence similarity and pattern-matching and pattern-finding queries.
- Development and deployment of standard ontologies in database systems and ontology tools.

GTL should award efforts for information-integration services and tools and actively promote the development and dissemination of data standards in the larger community. Data-integration design principles should permit the utilization of any form of local integration methods including language-based approaches; flat file, text retrieval, and search engines; data federation and distributed databases; classical data warehousing; centralization; and Web robots and agents. They also should provide mechanisms for all forms of higher-order global integration.

Data-Management Infrastructure

The exponential growth of genomics and proteomics data will far exceed the capabilities and capacities of any single institution. Workshop participants agreed that a distributed but highly coordinated data-management infrastructure is needed. Due to the unique research agenda of each institution, a centralized data infrastructure will be neither possible nor desirable. However, to effectively and efficiently serve GTL data-management needs, regular coordination is needed among sites.

The group emphasized that management issues need to be addressed with high priority from the start of GTL. Among those issues are types of GTL-generated data, means of long-term support of data curation, mechanisms for capturing data (publicly accessible, central vs dispersed repositories, grid-based replicas, federations), mechanisms for filtering data that need to be stored, and ways of disseminating data.

The group emphasized the need for developing middleware components of a distributed search infrastructure to address the scale, heterogeneity, and distributed nature of biological data. Data-integration infrastructure should enable search services to interoperate across domains by providing user-configurable tools for mapping between metadata schemas, performing search queries against multiple data sources, and performing pre- and post-processing queries.

Finally, the group concluded that to make the growing body of biological data available in a form suitable for study and use, a GTL data-management infrastructure must do the following:

- Develop a methodology necessary for seamless integration and interoperation of distributed data resources co-located with major experimental facilities that will enable linking both experiment and simulation.
- Provide mechanisms for automated data deposition and automated and manual data annotation and curation by local and remote experts.

- Develop life sciences enabling database frameworks that provide complex and multidatabase queries, new data models natural to life science, enhanced operations on these data types, and optimized performance.

Data-Quality Control

Quality control emerged repeatedly during the workshop as an obstacle to sharing data across the GTL community. Current databases often are incomplete and contain erroneous information. Furthermore, such spurious information in databases is being propagated increasingly fast. For example, functional information is transferred from proteins annotated in databases to unknown proteins based on their sequence similarity. These transfers, however, can be extremely uncertain and misleading due to the complex evolutionary and structure-function relationships among genes. This functional information is then stored in a database that may be used in other analyses, and the cycle of propagation continues.

The data-quality problem in GTL will get much worse as proteomics data are generated. Whereas the genome is static, proteomics data are dependent on time and on initial conditions. The correct interpretation and summarization of such data will depend on how well additional biological context is being captured. Since experimentally obtained data often provide higher strengths of evidence, quality control in GTL experimental facilities will be even more important. To reduce erroneous information, for example, GTL experimental facilities must have robust analysis tools to statistically validate identified protein complexes as the ones existing in the cell rather than as the artifacts of purification and separation procedures. Likewise, routine checks for completeness of “complexome” coverage will be needed to minimize the amount of incomplete information. Especially problematic could be the experiments capturing transient complexes corresponding to weak binding between subunits but constituting critically important regulatory pathways in the cell.

The workshop emphasized that databases and experimental data repositories should be designed with data-quality control in mind. They should include:

- Data provenance or history of the origin and ownership of data.
- Thorough collection of “metadata” that describes the data itself with a wide range of attributes that must be tracked (e.g., cell type, position in the cell cycle, growth conditions, or computational tools and parameters used) to accurately evaluate experimental or computationally derived data.
- Evidence attribution including source and strengths of evidence (e.g., experimentally verified vs computationally predicted, statistical significance of predictions).
- Automated and manual data annotation and curation as well as systematic detection and correction of annotation errors by local and remote experts.
- A model organism database (MOD) for every sequenced organism central to a DOE mission. MODs are powerful platforms for global analysis of an organism.

Data-Analysis Infrastructure

The GTL program promises to create innovative technologies for high-throughput production of biological data at a rate that will outpace that of any program currently under way. We expect GTL to embark on interesting experiments for thousands of organisms by 2008. Global proteomics is currently generating ~1.0 terabytes (or 10^{12} bytes) a day and scaling up now with 5- to 10-fold increases per year. Not only massive in volume but also very complex, these data span many levels of scale and dimensionality. They include genome sequences, protein structures, protein-protein interactions, and metabolic and regulatory networks. The strategic problem is to make biological sense of the data. Current applications allow, at best, data acquisition and cataloging by organizing the data dump into a tidier pile. However, this does not solve the problem. There is a strong need for “smart” data analysis and modeling tools that will enable the transformation from data through information to knowledge.

Significant research challenges remain in *systematic* incorporation of different data types into the analysis to construct predictive models of microbial organisms. For example, putative functional sites retrieved using patterns extracted from motif databases can be false positive. Given the few positions involved in a pattern, the statistical significance of a match also can be low. Additional “context” including a protein structure, protein family, or protein function often can be utilized to further filter out such false-positive predictions. Therefore, an appropriate “fusion” of various types of data can have significant impact on accomplishing the stated goals of the GTL program.

Experimental Templates for a Single Microbe

Experiment Class	Time Points	Treatments	Conditions	Genetic Variants	Biological Replication	Total Biological Samples	Proteomics Data Volume in TB	Metabolite Data in TB	Transcription Data in TB
Simple (Scratching the Surface)	10	1	3	1	3	90	15.0	13.5	0.018
Moderate	25	3	5	1	3	1,125	187.5	168.8	0.225
Upper Mid	50	3	5	5	3	11,250	1,875.0	1,687.5	2.25
Complex	20	5	5	20	3	30,000	5,000.0	4,500.0	6.0
Real Interesting	20	5	5	50	3	75,000	12,500.0	11,250.0	15.0

Profiling Method

- Proteomics: Looking at a possible 6000 proteins per microbe, assuming ~200 GB per sample.
- Metabolites: Looking at a panel of 500 to 1000 different molecules, assuming ~150 GB per sample.
- Transcription: Six thousand genes and two arrays per sample, ~100 MB.

Typically, a single significant scientific question takes the multidimensional analysis of at least 1000 biological samples.

The data-analysis infrastructure should promote compute-experiment cycles. Performing experiments in silico will offer a clear benefit to complement experimental laboratory methods by providing fast and inexpensive initial analysis to guide further experimentation. In addition to high-sensitivity analytical tools for interpreting experimental data, there is a strong need for developing experience-based systems to predict optimal experimental-design strategies.

Participants concurred on the need to develop next-generation algorithms and tools that will allow biologists to derive inferences from massive amounts of complex, heterogeneous, and distributed biological data to

- Develop data-analysis and -interpretation systems that will provide inference capabilities for establishing relationships across data sources generated by the GTL program (genomic sequence, gene and protein expression, protein-protein interactions, protein structures and complex structures, and biological pathways), leading to new scientific discoveries.
- Create computational tools and capabilities for assimilating, understanding, and modeling data on the scale and complexity of real living systems to build a dynamic knowledge base from this information.
- Enable distributed analysis of ever-increasing databases of diverse biological data for inclusion into simulations models.
- Develop algorithms for integration of noisy, incomplete, and inconsistent data from heterogeneous sources to comprehensively characterize “cellular working parts.”
- Evaluate and optimize the performance of computer-intensive data-analysis algorithms so the targeted computer codes may achieve a higher percentage of peak on systems such as Cray X1 and clusters. These optimized tools would be made available to the broader biological community.

Workflow Environments for Data Collection

Workflow environments should be seen as open extensions to laboratory information management systems (called LIMS) that will be integrated with robotic equipment to capture data in real time and to direct instrument workflow. High-end automation of all steps will be required to reduce experimental costs and to make all data available in real time to GTL researchers and other members of the scientific community. Especially needed is a workflow-based environment that will provide the flexibility and generality required to run complex synchronous and asynchronous scientific experiments and analyses for GTL activities.

Several requirements must be imposed on the development of such workflows:

- They should have fast prototyping capabilities via various interfaces including GUI-based flow-chart formulations like OpenDX or Labview, combined with data-mining algorithms embedded to a programming language like Perl but easier to use by biologists.

- Unlike traditional “Web services” environments, they should work more effectively in the type of computational environment envisioned for GTL, such as Web services, local data, local-parallel and sequential computing, and production-level reliability and fault tolerance.
- Workflow-definition languages should be expressive enough to meet the needs of GTL data acquisition and analysis. Relative merits of different ways of expressing applications within analysis pipelines should be investigated. Specifically, trade-offs in implementation, performance, fault tolerance, and flexibility should be assessed for different forms of workflow components including (local or remote) Web services, data-transformation services, locally invoked “wrapped” executables, and components in the sense of a component model such as the common component architecture (called CCA).

Such workflow capabilities should be developed via close collaboration among biologists and computer scientists to understand and define workflow and to capture the ways in which biologists approach problems.

Transparent Access to Data and Computational Resources

The Internet is by far the method most preferred for disseminating biological data. The informational interface is crucial for communicating the relevance of GTL activities to DOE and the national scientific agenda. By definition, the GTL program will have many users (including remote users) with diverse needs. Some academic researchers, for example, will be interested only in protein complexes related to particular metabolic pathways, while others may be interested in groups of pathways or complexes that show elevated expression level under certain conditions. Users will not be interested in, and will not be able to handle, the enormous flow of raw data produced by GTL. Therefore, a wide array of bioinformatics tools will have to be deployed to process, filter, and present data according to the user’s needs. In some cases, computational post-processing requirements will be quite extensive. As mentioned above, sophisticated semantic and context support will be required.

Thus, accessibility and high-quality presentation of all available biological data to the end user will be critical to GTL’s success. User-friendly interfaces are needed that will allow biologists to effectively access and manipulate vast amounts of data at their disposal. Along with a user-friendly interface, biologists need to know the intrinsic quality of the data (i.e., provenance, completeness, noise). Hence, the integration of front-end interfaces with data-quality control engines must be supported.

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.

Conclusions

Technically, GTL will need a flexible data framework because biology is moving at a fast pace. The types of data involved will be determined by experiments and also will impact infrastructure requirements. For this reason, data-analysis and -storage strategies should be allowed to evolve over time in an organized and timely way.

A number of common issues surfaced in the presentations and subsequent discussions. Most prominent were data integration; data mining; derivation of knowledge from diverse data sources; data management; and challenges associated with data quality, statistical analysis, variability of assays, and, in general, data-set reproducibility.

An important step is to address and resolve serious issues concerning data resources and access methods. The current state of the art for biology is less than desirable. There are myriad data silos and a few monolithic, asymmetric cross-references. A consequence of this poor data integration is the propagation of spurious information in databases. Many data resources have limited, idiosyncratic querying capabilities that are designed mostly for browsing human data. There is a lack of accepted standards for defining, querying, and transmitting common data objects; neither are there effective strategies for discouraging data hoarding (delayed releases of data are not uncommon). Ultimately, the success of GTL will be judged by how well the program is accepted and serves groups within DOE and, just as important, the broader life sciences community. To achieve this success, the GTL program needs a new paradigm on data ownership in which the data are openly available.

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.

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 primary component of the program from the beginning.

Report on the Genomes to Life Modeling and Simulation Workshop

July 23, 2003

Organizers: Reinhold Mann and George Michaels

Introduction

Biological modeling and simulation are key to the next phase of Genomes to Life (GTL). Most dynamic features of metabolomics and protein interactions within microbes are impossible to measure experimentally today. Modeling and simulation offer the potential to explain both experimental observations as well as to help guide future experiments.

Experiments in turn help validate the simulated models in a symbiotic cycle of computation and experiment. Because of its leadership in biological and computational science and its vast computational infrastructure, the U.S. Department of Energy (DOE) is uniquely positioned to make fundamental contributions to modern cellular biology. A focused research effort, however, is essential to accomplishing the goals of GTL.

To help identify and characterize this research effort, a workshop supported by DOE's Office of Advanced Scientific Computing Research and Office of Biological and Environmental Research was held in Gaithersburg, Maryland, on July 23, 2003. The workshop focused on defining the modeling and simulation needs for the next phase of the GTL program in sufficient detail to guide R&D activities. The main objectives of the modeling and simulation workshop were the following:

- Provide a clear definition of how modeling and simulation relate to experimental data generated by genomics, proteomics, and metabolomics. The connection to biological relevance and the integration of modeling and simulation with experiment are important. Well-characterized experimental data sets are needed that can drive modeling and simulation benchmarking.
- Discuss potential benchmark paradigm problems that could lead to identifying sufficient detail of the specific mathematical and computational problems to be addressed. Discuss metrics for models linked to experimental data. Of particular interest will be biologically relevant modeling and simulation problems that drive efficient use of terascale computer systems.
- Provide a clear definition of the role for high-performance ultrascale computing.

Uses of Modeling and Simulation to Accomplish GTL's Goals

Participants gave the following recommendations for three areas in which DOE should invest to accomplish its GTL goals.

Molecular Simulations of Protein Function and Macromolecular Interactions

Molecular simulations of protein function are necessary in many situations where direct observations are difficult or impossible. Typical simulations of cellular biochemistry require substantial input of protein behavior, some of which is difficult to obtain experimentally. For example, the binding and unbinding rates of proteins in complexes can affect the more “important” functions of those complexes (e.g., signal transduction). The alternatives seem to be to develop new experimental technologies, exploit old experimental technologies more thoroughly, and develop molecular dynamics simulations to try to avoid the experimental avenue.

Simulation and Modeling of Cellular Biochemistry

Modeling of cellular biochemistry will increasingly involve accurate, or at least plausible, models of cellular structures, volumes, and gross mechanics. This increasingly important spatial component has concomitant visualization needs and opportunities. Just constructing a large-scale simulation with complex spatial structures demands flexible visualization tools, while the value of powerful visualization tools in analyzing results of such simulations has already been established. This is another area where the computational strengths and expertise of the national laboratories can be applied, both by making high-performance software available and by providing computational infrastructure for its actual use in extremely large scale applications.

Development of Better Qualitative Methods

Tools in this category prove their worth daily in biology and should not be overlooked by the GTL program simply because they may not seem like simulation or even, in the conventional sense, “applied math.” Some of these tools, such as hidden Markov models, offer built-in inferential capabilities. Their application to system behavior, as exemplified in the theory of qualitative ODEs and dynamic Bayesian networks, are again modeling technologies that are in their infancy compared to simulation technologies of high-energy physics or to their current use in sequence analysis. These modeling techniques present an opportunity for applying DOE expertise that should not be neglected because of their unorthodox modeling approaches.

Summary of Talks and Discussions

The flood of experimental data being generated contains little that can be used with existing modeling methods. There's a mismatch between the experiment needs or design approaches by modelers versus those of biologists. For example, yeast now has >20,000 measured protein-protein, protein-DNA, protein-small molecule interactions. Similar networks soon will be available for a variety of bacteria and the worm, fly, mouse, and human. The pressing need is for computational models and tools able to integrate molecular interaction networks with molecular states on a cellular scale.

Simulation-driven experimentation is missing. Mathematical models are needed that ultimately are developed from fundamental biological principles. These models must be tested and verified through integrated wet-lab experimentation using multiple analytical methods and based on well-characterized statistical designs. Presently there is a lack of data to validate models and simulations and a lack of whole genome and proteome data to construct large-scale models. Most existing models are for small-gene or protein systems.

It still is a significant challenge to infer regulatory networks from metabolites, expression data, or protein-protein interactions. Modeling-integration frameworks are critically needed that allow multiple cellular system models to be easily combined into a single simulation.

The participants suggested that a competition similar to Critical Assessment of Protein Structure Prediction (called CASP) but focused on computational challenges faced by the Genomes to Life program would inspire the community and provide metrics for success.

A number of tools were identified as critical to the next phase of GTL, and support for their development should be established. These include analysis tools that test the coherency of disparate bodies of data; network inference tools because most problems of concern will come down to modeling the interactions of a number of different interacting species; and good visualization tools to allow experts in biology to find patterns or artifacts in large data sets not easily detected in other ways. Development of modeling and simulation toolkits and libraries would provide a means of integrating and distributing these and other tools needed within the GTL facilities and across the GTL program.

Many of today's molecular biophysics simulations are limited by the quality of the force fields. Research is needed in the creation of high-quality force fields for biophysics simulations. Multiscale mathematical research is needed on a wide range of dynamical systems both spatial and temporal. This finding concurs with recommendations from the *Report on the Mathematics Workshop for the Genomes to Life Program, March 18–19, 2002*.

General Infrastructure Needs

Participants identified a number of high-priority needs common across the GTL community and vital to exploring biology problems. New types of databases (both hardware and operating system) must accommodate large data volumes, high schema complexity, and rapid query retrieval. Along with this, research should be done on new scalable storage hardware and software systems that can accommodate petabyte-scale data volumes and provide rapid

analysis, data query, and retrieval. Rapid retrieval will require environments for large-scale data analysis on clusters and massively parallel programming technology for tools, libraries, and repositories. Support will be vital for the development of reusable component and middleware analysis codes. One computational challenge of reverse engineering is to rigorously solve a network model that best matches known data and knowledge of the biology modeled. Data mining is an essential first step in solving the reverse-engineering problem. Much existing information is hidden in the often-noisy, incomplete, and sometimes-conflicting data. Computational prediction and modeling and data collection through experiments should be one integrated process; computation should be a key driver for designing experiments.

Networking and computing hardware also are required across the community, along with robust network technologies to support GTL facility-oriented community access, analysis, and archival activities. Stable computing power (i.e., in a production-oriented environment) is needed to run long time-scale biological simulations as well as real-time experiment drivers for the GTL facilities.

Recommendations

Participants identified infrastructure requirements that span the entire GTL community as well as some that are specific to GTL modeling and simulation. They recommended the following actions:

- Support and develop plans for storage, community access, and analysis of the sometimes-vast amounts of GTL facility-oriented experimental data produced by a variety of high-throughput technologies. To this, we will gradually have to add similar data coming from simulations, and we will have to develop analyses that test the coherency of disparate bodies of data.
- Establish stable, production-oriented high-performance computing capabilities for long time-scale modeling and simulation computational experiments.
- Mathematical models are needed that are ultimately developed from fundamental biological principles and incorporate the above analyses for whole-cell simulation uniting genomics, proteomics, and metabolomics complexities. DOE's advanced computing and systems biology facilities and expertise put the agency in a unique position to help develop new modeling and simulation theories and to implement these principles in ways that leverage some of the world's most powerful computers.
- These models must be tested and verified through integrated wet-lab experimentation using multiple analytical methods and based on well-characterized statistical designs. A specific call should be issued for model- and simulation-driven experimentation.
- Develop algorithms for scalable stiff and differential-algebraic integrators, multiobjective constrained optimization, and multiparameter bifurcation and sensitivity analysis, statistical graph models, stochastic optimization, and computationally intensive operations.

- 
- Conduct research on model analysis methods including model abstraction, version management, model transport, reduction, parametric sensitivity, and parameter development using collaborative data filtering for data constraints: large matrix manipulation, optimization.
 - Develop plans for establishing a modeling and simulation infrastructure centered on metabolism for both improved understanding and engineering of metabolic systems.
 - Support development of hybrid simulation systems that would integrate methods for mixed deterministic and stochastic, mixed discrete and continuous, mixed differential and algebraic, or mixed-scale simulations.
 - Establish working groups to define modeling and simulation data and experimentation requirements for validation equivalent to a CASP competition for systems biology.

Report on the Genomes to Life Protein Structure Prediction Workshop

July 24, 2003

Organizer: Grant Heffelfinger

Introduction

Prediction of three-dimensional structures of proteins from their amino acid sequences via computational methods is a well-studied problem in modern computational biology. This is due not only to the problem's technical challenge, but, more significantly, to its importance. Proteins and protein complexes make up the biological machinery that carries out the biological functions in a cell; understanding the functional mechanisms of biological activity requires knowing the fundamental atomic structure and dynamic behavior of proteins and protein complexes. Ultimately, a protein's structure provides much more functional information than its amino acid sequence.

The arrival of high-throughput genomic sequencing has led to an explosion of genomic information, but experimental methods for solving protein structures, including X-ray crystallography or NMR, remain slow and expensive. Furthermore, many proteins are expressed at very low rates, making them difficult to obtain in experimentally useful quantities. Other proteins are difficult to crystallize (a requirement for X-ray crystallography methods) due to physical-chemical attributes (e.g., low solubility) associated with their function. For example, membrane proteins are largely insoluble yet are thought to comprise 30% of all proteins! Such limitations also often apply to protein complexes; thus, experimentally resolving the structure of a single protein complex often requires many months of work. Finally, microbial genomes now can be sequenced and annotated within days, providing the amino acid sequences of a microbe's proteome, but establishing functional annotation of the proteome is a key bottleneck in high-throughput microbial biology.

In the meantime, computational protein structure prediction has become increasingly powerful with the availability of a growing number of solved protein structures (due to the successes of experimental methods) as well as the realization that in nature, the number of unique structural folds is quite small compared to the number of proteins. Thus, many proteins can be accurately modeled based on homologous structures via threading or homology modeling; the potential applicability of such techniques is estimated to be as high as 50 to 60% of all proteins in a newly sequenced microbial genome. Furthermore, computationally predicted structures lower in resolution than experimental measurements have significant

utility (e.g., to suggest protein functions and mechanisms or for genome-scale annotation work). More accurate structure predictions, on the other hand, provide the basis for protein complex structure prediction and understanding of the dynamics of protein complexes.

For all these reasons, computational methods of predicting protein structure are widely seen to hold the most promise for estimating the structure of most proteins in all genomes at various *levels of resolution*. However, significant new mathematical, computer-science, and high-end computing tools and capabilities are needed to enable these methods to realize their potential. More specifically, computational structural genomics presents a class of challenging computational problems involving searching an enormous conformational space. High-performance computing, sophisticated new algorithms, and parallel implementations are key to addressing these challenging problems. For these reasons, the U.S. Department of Energy, with its collection of high-performance computing facilities, will play a key and unique role in addressing the challenging issues of computational prediction and modeling of protein and complex structures, especially for the proteomes of microbes with relevance to DOE's missions in energy production, global climate-change mitigation, and environmental cleanup.

State of the Art

Computational prediction and characterization of protein structures and complexes can be classified into the following categories: (1) predicting the structure of individual proteins, (2) predicting the structure of protein complexes, and (3) understanding the dynamics of protein complexes. While protein structure prediction provides a foundation for all three, understanding the dynamic behavior of protein complexes is key to understanding their functions and fundamental mechanisms and often employs methods drawn from computational molecular biophysics and biochemistry.

In general, computational methods for elucidating molecular structure and processes can be classified into three major categories depending on the similarity of the target (the protein for which a structure is desired) to proteins with known structure: (1) comparative modeling, (2) threading, and (3) *de novo* or *ab initio* structure prediction. Because these methods have varying levels of computational complexity, their boundaries are becoming more and more blurred as each class of methods employs techniques and ideas from other classes, ultimately yielding hybrid methods.

Comparative modeling involves carrying out sequence alignment between the target protein and one or more other template proteins or proteins with known structure. The three-dimensional structure for the target protein is then constructed from the coordinates of the template protein. For regions where there are few or no overlaps or gaps in the sequence alignment, coordinates are obtained from other models. Statistical analysis has shown that comparative modeling can provide reliable atomic coordinates with a low root mean square deviation (*rmsd*) from a high-quality, experimentally obtained structure for about 20% of all proteins in a genome. Furthermore, in analyses of the fourth community-wide Critical Assessment of Protein Structure Prediction (called CASP), Moulton et al. and Schonbrun et al.

observe that the key element to the success of comparative modeling is sequence alignment: “loop modeling and further refinement are futile without a reasonably accurate initial alignment.” In addition, multiple sequence alignment and multiple proteins used as templates for different regions of the target sequence may improve results. Interestingly, however, using molecular dynamics or molecular mechanics to refine structures predicted by comparative methods often has increased, rather than decreased, the *rmsd* from the experimentally derived structure. Most agree that a systematic investigation is needed to obtain fundamental insight into why this is true and thus suggest methods for improving comparative modeling. Finally, and perhaps most telling, comparative modeling still generally predicts structures that are closer to the best-available template used for sequence alignment than to the experimentally derived structure. In other words, in most cases, the *rmsd* between a structure predicted by comparative modeling and the experimentally derived structure is larger than that between the comparative modeling prediction and the best-available template.

In threading, a suitable fold from a library of structures is employed as the query sequence, yielding an alignment between the query protein and the fold. This class of methods currently is applicable to some 50 to 70% of all proteins, as long as a protein has a structural homolog and analog in the space of known proteins. For this reason, to date, threading has been useful mainly for identifying structural folds and predicting backbone structures.

Unlike homology modeling and threading, both of which rely on a known structure template, *ab initio* structure prediction involves predicting structure utilizing physical principles of protein structure. The key advantage of this approach is that it does not require a structural template for a whole protein, making it broadly applicable. However, because this method is computationally demanding, many recent *ab initio* approaches use knowledge-based methods in combination with high-quality force fields. For example, one can use the alignment derived from fold recognition in comparative modeling or assemble partial structures predicted by threading before applying *ab initio* methods. Such combinations currently comprise the primary successes of such techniques.

Finally, comprehending the dynamics of protein complexes is essential for such specific phenomena as protein self-assembly, protein-protein interactions or docking, and understanding how molecular machines work. The state of the art in developing and applying computational methods to address these challenges varies greatly with the specific challenge. For example, computational methods applied to protein docking can be classified as rigid-body docking or flexible docking, depending on whether or not the models allow the proteins' docking regions to move or flex during the docking process. The conformations of docking proteins are well known to experience significant changes, particularly at the docking interface, but capturing such molecular phenomena is computationally expensive. While rigid docking has reached some level of maturity for practical applications, flexible docking remains beyond our reach. An increasing amount of data is pointing to flexible docking as the underlying mechanism for such important and fundamental cellular processes as signaling. Meanwhile, like the protein structure prediction problem, hybrid methods continue to emerge as potentially useful approaches to the flexible docking problem; these include using

sequence-based structure predictions for the protein-interaction surfaces, followed by molecular models of the flexible docking process, much like comparative modeling followed by ab initio refinement for protein structure prediction.

Goals and Challenges

Workshop participants discussed technical challenges to computational protein structure determination and worked to identify specific goals. Goals and challenges were grouped into three categories depending on their drivers: (1) biology issues, (2) math and computing science issues (computer science, computational science, and high-performance computing), and (3) other issues.

Biology Issues

During the course of the workshop, two specific metrics were advanced. “Successful” methods should be able to:

- Predict structures with 2Å *rmsd* for proteins with 200 residues given 40% amino acid alignment with proteins in the database, and
- Correctly predict contacts and hydrogen bonds.

In addition, seven specific challenges were identified:

- 1. Accurate Predictions of Protein Backbone Structures.** The fundamental challenges identified for predicting backbone structures were the percentage of correctly predicted contacts and hydrogen bonds, especially given the lack of adequate sequence alignment.
- 2. Membrane Proteins.** Experimentally obtaining the structure of membrane proteins is very challenging, given the difficulty of crystallizing them. Furthermore, not only do membrane proteins play significant roles in important cellular processes (e.g., cell signaling), they are thought to comprise some 30% of the proteome of any given cell. For these reasons, computational and experimental methods (e.g., solid state NMR, optical approaches) are needed for a variety of applications beyond predicting structure and understanding dynamic molecular processes. Examples include predicting membrane type from a membrane protein’s amino acid sequence and elucidating a membrane protein’s location in and orientation to the membrane, once again subject to the two identified performance metrics discussed above.
- 3. “Refining Refinement” or Force-Field Development.** The intuitively appealing approach of employing more substantial or more accurate molecular information to improve results is still evolving. Ultimately, the general goal that “40% amino acid sequence alignment is sufficient for 2Å *rmsd* for proteins with 200 residues or less” was advanced as the ultimate metric for success in the improvement of refinement methods. However, one shorter-term metric also advanced was that “refinement uniformly improves the results of coarse-grained models.” Finally, participants agreed that refinement methods should be able to accurately predict the thermodynamics of model systems.

- 4. Obtaining and Coupling with Needed Experimental Data.** The success of computational methods could be enhanced significantly with the availability of more and varied types of experimental data such as NMR or cryo-EM. Once again, the ultimate usefulness of such data should be judged in the context of the two performance metrics.
- 5. Protein Assembly, Docking, Molecular Machines.** This broad category of challenges captures the essential need to use computational methods to accurately predict biological function and yield fundamental mechanistic understanding of these molecular processes.
- 6. Functional Annotation or Exploiting Evolutionary Relationships.** This is seen as a challenge in terms of current levels of confidence in predictions of such methods.
- 7. Exploiting Peptides Toward the Prediction of Function and Potential Binding Sites.** This is an essential computational goal to enable the successful development of high-throughput experimental proteomics methods where the identification of associations is a key requirement.

Math and Computing Science Issues

Six specific math and computing science issues were identified:

- 1. Global Optimization, Sampling, and Statistics.** Components of this broad area need additional investigation. Examples of such work include mathematical proofs for discrete representations of model systems and a single performance metric, “random starting conditions give uniform results (reproducibility).”
- 2. Force Fields, Including the Incorporation of Polarizability.** This area of development is a significant need, with the ultimate success metric of “parameterizing a new force field on 20 residue proteins giving *correct* results,” with *correct* quantified not only in terms of the two performance metrics discussed in the previous section but also the correct prediction of secondary structure. Workshop participants agreed that the development and implementation of new force fields would require tackling significant math and computing science issues ranging from mathematical methods to parallel implementations.
- 3. Incorporating Experimental Data, Knowledge-Based Methods.** This challenge is driven not only by biology but also by math and computer science, primarily due to integration methods that employ significantly different algorithms and approaches. Once again, performance metrics identified for biology-driven challenges are appropriate for judging progress in this area.
- 4. Algorithm Development, Simulation Methods, and Parallel Implementations.** This math and computing science challenge is advanced primarily in the context of developing new model methods and suitable mathematical representations. Such approaches are likely to range from knowledge-based protein structure prediction methods to computationally intense models employing detailed physical and chemical descriptions and data as well as combinations.

- 
- 5. Domain Parsing (e.g., Large Proteins).** Defined by the need for handling multiple domains within single large proteins, parsing poses significant math and computing science challenges because of difficulties in (1) precipitating a “starting point” for modular structures that can be folded, ultimately enabling a “divide and conquer” approach to structure prediction for large proteins, and (2) using experimental data to prioritize modular determination.

Other Issues

Finally, seven other challenges were identified that fall outside the definitions of “biology-driven” and “math- and computing science–driven.” All were viewed as issues relative to the application of computational tools to science and engineering tasks well beyond biology. As such, these issues were left to other venues for further discussion:

- Assessing Model Quality or Confidence
- Methods Verification
- Open Source Software Development Practices
- Implications of New Algorithms to High-Performance Computing Hardware Architectures
- Operating Systems Issues (e.g., job submissions, parallel I/O)
- Communication Needs Within the Research Community
- Code Portability.

Appendix A: Workshop Attendees

July 22, 2003—Data Infrastructure Workshop

Organizers: Al Geist and Thomas Zacharia

<u>Name</u>	<u>Affiliation</u>	<u>E-Mail Address</u>
Gordon Anderson	PNNL	gordon.anderson@pnl.gov
Stephen Elbert	IBM	selbert@us.ibm.com
Krzysztof Fidelis	LLNL	fidelis1@llnl.gov
Ed Frank	ANL	efrank@mcs.anl.gov
Marvin Frazier	DOE	marvin.frazier@science.doe.gov
Al Geist	ORNL	gst@ornl.gov
Nathan Goodman	ISB	natg@shore.net
Debbie Gracio	PNNL	debbie.gracio@pnl.gov
Grant Heffelfinger	SNL	gsheffe@sandia.gov
Gary Johnson	DOE	garyj@er.doe.gov
Peter Karp	SRI	pkarp@ai.sri.com
Arthur Katz	DOE	arthur.katz@science.doe.gov
Eugene Kolker	Biotech	ekolker@biotech.org
Mike Knotek	DOE	m.knotek@verizon.net
Phil Locascio	ORNL	locasciop@ornl.gov
Bertram Ludaescher	SDSC	ludaesch@sdsc.edu
Natalia Maltsev	ANL	maltsev@mcs.anl.gov
Reinhold Mann	PNNL	mannrc@pnl.gov
Michael McGuigan	BNL	mcguigan@bnl.gov
Noelle Metting	DOE	noelle.metting@science.doe.gov
George Michaels	PNL	george.michaels@pnnl.gov
Nagiza Samatova	ORNL	samatovan@ornl.gov



Arie Shoshani	LBNL	shoshani@lbl.gov
Nancy Slater	LBNL	naslater@lbl.gov
Bruno Sobral	Virginia Tech	sobral@mail.vt.edu
Michael Tereskinski	DOE	michael.tereskinski@science.doe.gov
David Thomassen	DOE	david.thomassen@science.doe.gov
Ed Uberbacher	ORNL	uberbachered@ornl.gov
John Westbrook	Rutgers	jwest@rcsb.rutgers.edu
Cathy Wu	Georgetown	wuc@georgetown.edu
Thomas Zacharia	ORNL	zachariat@ornl.gov

July 23, 2003—Modeling and Simulation Workshop

Organizers: George Michaels and Reinhold Mann

<u>Name</u>	<u>Affiliation</u>	<u>E-Mail Address</u>
Carl Anderson	BNL	cwa@bnl.gov
Adam Arkin	LBNL	aparkin@lbl.gov
Michael Banda	LBNL	mjbanda@lbl.gov
Paul Bayer	DOE	paul.bayer@science.doe.gov
Hamid Bolouri	ISB	hbolouri@systembiology.org
Roger Brent	TMSI	brent@moisc.org
Mike Colvin	LLNL	colvin2@llnl.gov
John Doyle	Cal Tech	doyle@cds.caltech.edu
Daniel Drell	DOE	daniel.drell@science.doe.gov
Stephen Elbert	IBM	selbert@us.ibm.com
Jean Loup Faulon	SNL	jfaulon@sandia.gov
Krzysztof Fidelis	LLNL	fidelis1@llnl.gov
Al Geist	ORNL	gst@ornl.gov
Debbie Gracio	PNNL	debbie.gracio@pnl.gov
Grant Heffelfinger	SNL	gsheffe@sandia.gov
Trey Ideker	MIT-Whitehead	trey@bioeng.scsd.edu
Gary Johnson	DOE	garyj@er.doe.gov
Arthur Katz	DOE	arthur.katz@science.doe.gov
Mike Knotek	DOE	m.knotek@verizon.net
Phil Locascio	ORNL	locasciop@ornl.gov
Larry Lok	TMSI	lok@moisc.org
Natalia Maltsev	ANL	maltsev@mcs.anl.gov
Reinhold Mann	PNNL	mannrc@pnl.gov
Noelle Metting	DOE	noelle.metting@science.doe.gov
George Michaels	PNL	george.michaels@pnnl.gov
Ion Moraru	U Conn Health Ctr	mararu@panda.uchc.edu
Mark Rintoul	SNL	rintoul@sandia.gov



Christophe Schilling	Genomatica	cschilling@genomatica.com
Scott Studham	PNNL	studham@pnl.gov
Michael Tereskinski	DOE	michael.tereskinski@science.doe.gov
David Thomassen	DOE	david.thomassen@science.doe.gov
Ed Uberbacher	ORNL	uberbachered@ornl.gov
Ying Xu	ORNL	xuy1@ornl.gov

July 24, 2003—Protein Structure Prediction Workshop

Organizer: Grant Heffelfinger

<u>Name</u>	<u>Affiliation</u>	<u>E-Mail Address</u>
Carl Anderson	BNL	cwa@bnl.gov
Thomas Darden	NIEHS	darden@niehs.nih.gov
Daniel Drell	DOE	daniel.drell@science.doe.gov
Stephen Elbert	IBM	selbert@us.ibm.com
Krzysztof Fidelis	LLNL	fidelis1@llnl.gov
Angel Garcia	LANL	axg@lanl.gov
Al Geist	ORNL	gst@ornl.gov
Grant Heffelfinger	SNL	gsheffe@sandia.gov
Gary Johnson	DOE	garyj@er.doe.gov
Arthur Katz	DOE	arthur.katz@science.doe.gov
Mike Knotek	DOE	m.knotek@verizon.net
Jerry Li	NIGMS	lij@nigms.nih.gov
Leslie Kuhn	Michigan State	kuhnl@msu.edu
Phil Locascio	ORNL	locasciop@ornl.gov
George Michaels	PNL	george.michaels@pnnl.gov
John Moulton	UMD	moulton@umbi.umd.edu
Ruth Nussinov	SAID Frederick	ruthn@ncisgi.ncicrf.gov
Steve Plimpton	SNL	splimpton@sandia.gov
Carlos Simmerling	SUNY Stony Brook	carlos.simmerling@sunysb.edu
Jeffrey Skolnick	U Buffalo	skolnick@buffalo.edu
Fred Stevens	ANL	fstevens@anl.gov
Michael Tereskinski	DOE	michael.tereskinski@science.doe.gov
Chang-Shung Tung	LANL	ct@lanl.gov
Ed Uberbacher	ORNL	uberbacher@ornl.gov
John VanRosendale	DOE	johnvr@er.doe.gov
William Wedemeyer	U Washington	bill_wedemeyer@usa.net
Todd Yeates	UCLA	yeates@mbi.ucla.edu

Appendix B: Workshop Agendas

July 22, 2003—Data Infrastructure Workshop

- 8:30 Welcome, introductions, and workshop mission (Gary Johnson)
- 8:45 Where we are today, previous meetings, and proposed GTL facilities (Al Geist)
- 9:30 Open discussion of state of the art and potential near-term community goals in GTL data infrastructure
- 10:00 Break
- 10:30 Half the participants present their vision of key data issues for GTL and describe how it complements or contradicts the discussion so far (5 minutes each). Each followed by short discussion by attendees (5 minutes).
- 12:00 Working lunch
- 1:00 Second half of participants present
- 2:30 Summary of key points made by participants
- 3:00 Break
- 3:30 Discussion of creation of whitepaper incorporating workshop results
- 5:00 End

July 23, 2003—Simulation and Modeling Workshop

- 8:00 Continental breakfast
- 8:30 Welcome, introductions, and workshop goals
- 8:45 Summary of Genomes to Life program
- 9:15 Biological drivers for modeling and simulation
- 9:45 Roundtable discussion of state of the art and potential near-term goals for GTL modeling and simulation
- 10:30 Break
- 10:45 Participants present single slide on their vision of the key modeling and simulation issues for GTL followed by short discussion by attendees
- 12:00 Working lunch (provided)
- 1:00 One-slide presentations continue
- 2:30 Summary of key points made by participants
- 3:00 Break
- 3:30 Discussion of process for development of workshop report, assignments for workshop participants
- 5:00 Adjourn

July 24, 2003—Protein Structure Prediction Workshop

- 7:30 Continental breakfast
- 8:00 Welcome, introductions, and workshop mission
- 8:15 Overview of GTL program and four proposed facilities
- 8:30 Computational protein structure prediction: An overview
- 9:00 Discussion
- 9:30 Break
- 9:45 Single slides: Visions and discussions of key technical challenges of computational protein structure prediction for GTL
- 11:45 Lunch
- 12:15 Single slides: Visions and discussions of key technical challenges of computational protein structure prediction for GTL
- 2:00 Summary of key points made by participants
- 2:30 Break
- 3:00 Discussion of whitepaper incorporating workshop results
- 4:00 Adjourn



Program-Planning Workshops for Genomes to Life

A series of program-planning workshops is being held to help plan and coordinate Genomes to Life. Meeting reports are placed on the Web as soon as they become available (<http://doegenomestolife.org/pubs.shtml>). To learn more about the program, please see the Web site or use the contact information on the inside front cover for Gary Johnson or Marvin Frazier.

Office of Advanced Scientific Computing
Office of Biological and Environmental Research
SC-30 or SC-72 / Germantown Building
U.S. Department of Energy
1000 Independence Avenue, S.W.
Washington, DC 20585-1290