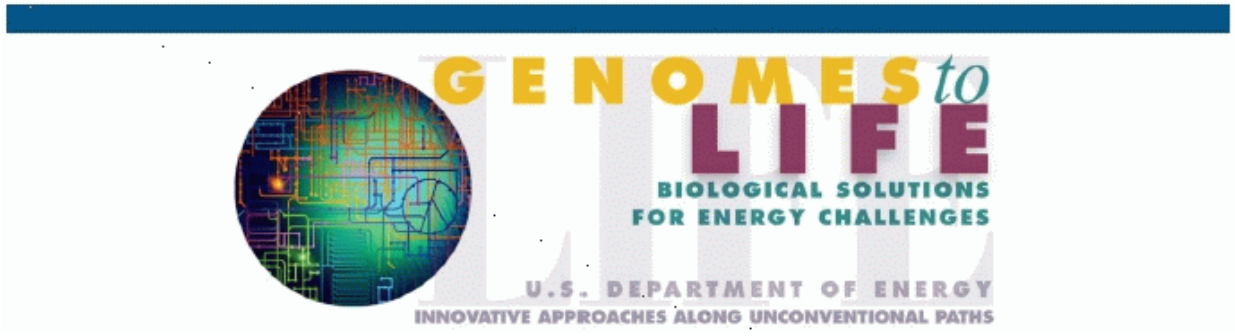

http://DOEGenomesToLife.org/compbio

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, ORNL
Reinhold Mann and George Michaels, PNNL
Grant Heffelfinger, SNL

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

September 2003

Executive Summary

On July 22, 23, 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 the data infrastructure needs for the Genomes to Life (GTL) program with the objective to identify gaps in the present GTL data infrastructure and define the GTL data infrastructure required for the success of the proposed GTL facilities. The second workshop discussed the modeling and simulation needs for the next phase of the GTL program and defined how these relate to the experimental data generated by genomics, proteomics, and metabolomics. The third workshop identified emerging technical challenges in computational protein structure prediction for DOE missions and outlining specific goals for the next phase of GTL. The workshops were attended by representatives from both OBER and OASCR.

The invited experts at each of the workshops made short presentations on what they perceived as the key needs in the GTL data infrastructure, modeling and simulation, and structure prediction respectively. Each presentation was followed by a lively discussion by all the workshop 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 it has to be treated as the first-class component of the GTL program to assure 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 problem of the massive quantities and increased complexity of biological data produced by experiments and computations. Genome-scale collection, analysis, dissemination, and modeling of those data are the key to success of GTL. Localizing these activities within each experimental facility that generates the data will ease integration and organization. However, integration and coordination of these activities across the facilities will be extremely critical to assure 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 the data for analysis, and standardized ways of disseminating the 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 a 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 the 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.
- Establish working groups to define modeling and simulation data and experimentation requirements for validation equivalent to a CASP competition for systems biology.
- Establish stable, production-oriented high-performance computing capabilities for long time-scale biological modeling and simulation computational experiments.

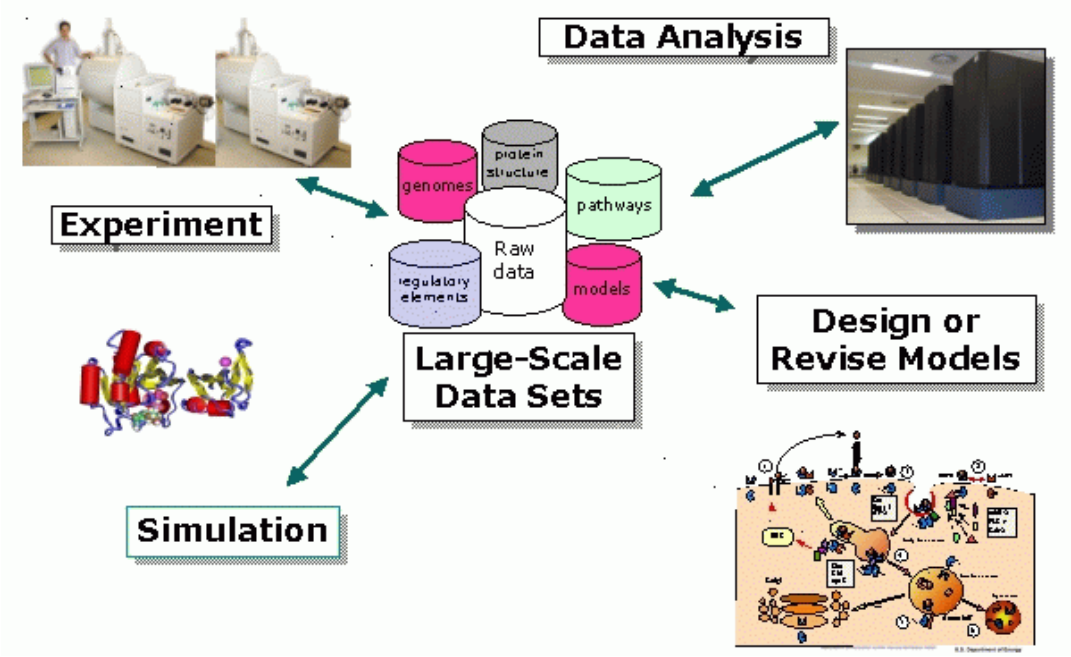

Figure 1. GTL requires and new synergy between computing and biology and data is at the center of it all.

Report on the
Genomes to Life Data Infrastructure Workshop

Organizers: Al Geist and Thomas Zacharia
Oak Ridge National Laboratory
July 22, 2003

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 was supported by DOE's Office of Advanced Scientific Computing Research and Office of Biological and Environmental Research. 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 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. There are currently about 335 genomic and molecular biology databases distributed around the country with large quantities of data being added daily.

The data problem is getting much worse as proteomics data is generated from arrays and mass spectrometers. Whereas the genome is static, proteomics data is time-dependent and dependent on the initial conditions. Much more experimental biological data about conditions, etc. must be carried with the proteomics data. In addition, the proteomics data often has a qualitative part that must be also available, e.g., the raw visual data from microarrays. The amount of data that will be generated from microbial community studies promises to be staggering. It will be necessary to establish the data storage infrastructure and formats for biological data. The data itself needs 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 so that the GTL facilities and

individual experimental and computational groups can benefit from an integrated storage infrastructure and related analysis tools.

Summary of Talks and Discussions

The Department of Energy's (DOE) GTL program has laid out an ambitious plan to use genomic data and high-throughput experimental technologies combined with ultrascale computing resources to study the proteins encoded by the genome throughout the lifecycle of the organism. A series of facilities will be created which 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, NMR, various binding assays, protein chips).

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

Data Standards and Integration in the GTL Data Infrastructure

Data integration has always 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 by technological advances that allow data to be created from the wet-lab at an ever-increasing rate, and by the growing need to combine these data in new and interesting ways.

A core requirement of any large-scale production enterprise such as the GTL Program is the management, manipulation, integration and presentation of the data. With the unique scientific challenges associated with each of the GTL projects and experimental facilities it will not be possible to have a centrally located data infrastructure, due in large part to the distinct research agenda. A seamless integration of GTL data spanning the entire range of genomics and proteomics will be extremely challenging but it has to be treated as the first-class component of the GTL program to assure 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 metadata types that are sharable across the GTL projects and facilities. In this way at the highest level the independent systems can evolve to meet the local conditions of domain experts effectively, while being able to share 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 permit users to mine data from the combined sites.

A key goal of the GTL distributed experimental collaboration is the availability of 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 the basic GTL objects as well as low-level data-interchange and access methods to permit the experimental facilities to communicate effectively. Furthermore, this will permit 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, as well as effective data mining tools for the GTL data resources.

Recommendations include finding common data needs and patterns between the GTL projects. Thus, leveraging from existing GTL projects, and 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 provide analysis end users with familiar interfaces. To ensure real requirements are met, the plan would be that each GTL would produce one or two use cases of the biology questions to be answered. The biologists in the GTL teams would generate the biology questions and potential solutions worked on 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 (lab experiments, microarray, pathways, etc.) 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
- Non-standard data formats including sequences, graphs, three dimensional structures, images, etc.
- 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 & imprecise operators) over non-standard data including sequence similarity, pattern-matching queries, 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/agents. They should also 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. The workshop participants agreed that a distributed, but highly coordinated, data management infrastructure is needed. Due to the unique research agenda of each institution, it will be neither possible nor desirable to have a centralized data infrastructure. However, to effectively and efficiently, serve the GTL data management needs, regular coordination is needed between the sites.

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

The group emphasized the need for developing middleware components of a distributed search infrastructure that addresses 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 query pre-and post-processing.

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

- Developing 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,
- Providing mechanisms for automated data deposition and automated and manual data annotation and curation by local and remote experts, and
- Developing life sciences enabling database frameworks that provide complex and multi-database queries, new data models natural to life science, enhanced operations on these data types, and optimized performance.

Data Quality Control

Data quality control emerged repeatedly during the workshop. It was identified as an obstacle to sharing data across GTL. Current databases are often 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. However, these transfers 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 analysis and the cycle of propagation continues.

The data quality problem in GTL will get much worse as proteomics data is being generated. Whereas the genome is static, proteomics data is time-dependent and dependent on the initial conditions. The correct interpretation and summarization of such data will depend on how well such additional biological context is being captured. Since experimentally obtained data often provides higher strengths of evidence, the quality control in GTL experimental facilities will be even more important. For example, in order to reduce erroneous information it will be extremely critical for GTL experimental facilities to have analysis tools for robust statistical validation of identified protein complexes as the ones existing in the cell rather than artifacts of the purification and separation procedures. Likewise, routine checks for completeness of the "complexome" coverage will be needed in order to minimize the amount of incomplete information. Especially problematic could be the experiments capturing transient complexes corresponding to the 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 the history of the origin and ownership of data.
- The thorough collection of "meta-data" 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) in order to accurately evaluate the 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.
- Every sequenced organism central to a DOE mission have a Model Organism Database (MOD). 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 1000's of organisms by 2008. Global proteomics is currently generating ~1.0 terabytes (or 10^{12} bytes) a day and scaling up now with 5-10 fold increases per year. This data is not only massive but also very complex. It spans many levels of scale and dimensionality, including genome sequences, protein structures, protein-protein interactions, and metabolic and regulatory networks. The strategic problem is to make biological sense of this data. Current applications allow, at best, data acquisition and cataloguing 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 modelling tools that will enable the transformation from data through information to knowledge.

Experiment templates for a single microbe

class of experiment	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	1125	187.5	168.8	0.225
upper mid	50	3	5	5	3	11250	1875.0	1687.5	2.25
complex	20	5	5	20	3	30000	5000.0	4500.0	6
real interesting	20	5	5	50	3	75000	12500.0	11250.0	15

Profiling method	
Proteomics	Looking at a possible 6000 proteins per microbe assuming ~200 GB per sample
Metabolites	Looking a panel of 500-1000 different molecules assuming ~150GB per sample
Transcription	6000 genes & 2 arrays per sample ~100 MB
	Typically a single significant scientific question takes the multidimensional analysis of at least 1000 biological samples

There remain significant research challenges in *systematic* incorporation of different data types into the analysis in order to construct predictive models of microbial organisms. For example, putative functional sites retrieved using the patterns extracted from motif databases can be false positive. Given the few positions involved in a pattern, the statistical significance of a match can be also low. It is often the case, that additional "context" including a protein structure, protein family, or protein function 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.

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 for predicting optimal experimental design strategies.

The participants concluded that there is a need to develop the next-generation algorithms and tools that will allow biologists to derive inferences from massive amounts of complex, heterogeneous, and distributed biological data. Specifically, there is a need for:

- Developing data analysis and interpretation systems that will provide inference capabilities for establishing relationships across data sources generated by the GTL Program (genomic sequence, gene/protein expression, protein-protein interactions, protein structures and complex structures, and biological pathways) leading to new scientific discoveries.

- Creating computational tools and capabilities to assimilate, understand, and model the data on the scale and complexity of real living systems, to build a dynamic knowledge base from this information.

- Enabling distributed analysis of ever-increasing databases of diverse biological data for inclusion into simulations models.

- Developing algorithms for integration of noisy, incomplete, and inconsistent data from heterogeneous sources to comprehensively characterize "cellular working parts".

- Evaluating and optimizing the performance of computer-intensive data analysis algorithms so that the targeted computer codes may achieve higher percent of peak on systems such as Cray X1 and clusters. Then making these optimized tools available to the broader biological community.

Workflow Environments for Data Collection

The workflow environments should be seen as open extensions to LIMS systems that will be integrated with robotic equipment to capture data in real-time and 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 the scientific community. There is a strong need in developing a workflow-based environment that will provide the flexibility and generality required to run the complex synchronous and asynchronous scientific experiments and analyses for the GTL activities. There are several requirements imposed on the development of such workflows:

- They should provide 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 plus local data

plus local parallel and sequential computing plus production level reliability and fault-tolerance.

- Workflow definition languages should be expressive enough to meet the needs of GTL data acquisition and analysis processing. Relative merits of different ways of expressing the applications within analysis pipelines should be investigated. Specifically, the trade-offs in terms of 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, or components in the sense of a component model, such as the Common Component Architecture (CCA).

The development of such workflow capabilities should be done in close collaboration with biologists and computer scientists in order to understand and define workflow and to capture the ways the biologists approach the problems.

Transparent Access to Data and Computational Resources.

The Internet is by far the most preferred method for disseminating biological data. The informational interface is crucial for the relevance of the GTL activities to DOE and the national scientific agenda. By definition the GTL Program will have many users (including remote users) with diverse needs. For example, some academic researchers 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, the computational post processing requirements will be quite extensive. As mentioned above, sophisticated semantic and context support will be required.

Thus, the 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 the vast amount of data at their disposal. Along with a user-friendly interface, the biologist needs 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.

Combining visualization and data mining for powerful exploratory and pattern recognition capability. Case study: Combining visualization and data mining applied to mass spectrometer proteomic data used for early cancer detection in collaboration with NIH and NCI.

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.

Data Infrastructure Workshop Conclusions

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.

There were a number of common issues that surfaced in 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 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 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. 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 nor 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 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.

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 first-class component of the program from the beginning.

Report on the Genomes to Life Modeling and Simulation Workshop

Organizers: Reinhold Mann and George Michaels
Pacific Northwest National Laboratory
July 23, 2003

Introduction

Biological modeling and simulation are key to the next phase of Genomes to Life (GTL). Most dynamic features and 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 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. But a focused research effort is an essential step toward accomplishing the goals of the Genomes to Life program.

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 the 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 to:

- 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 integration of the modeling and simulation with experiment are important. There need to be well characterized experimental data sets that can drive modeling and simulation benchmarking
- Discuss potential benchmark paradigm problems that could lead to sufficient detail of the specific mathematical and computational problems to be addressed. And discuss metrics for models linked to experimental data. Of particular interest will be those biologically relevant modeling and simulation problems that drive efficient use of tera-scale computer systems.
- Provide a clear definition of the role for high-performance ultra-scale computing.

Uses of Modeling and Simulation to Accomplish GTL's Goals

Participants gave the following recommendations for areas in which DOE needs to 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, such as signal transduction. The alternatives seem to be to develop new experimental technologies, exploit old experimental technologies more thoroughly, and to 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 compared with 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

There's a flood of experimental data being generated; however, there's a paucity of data that can be used with the existing modeling methods. There's a mismatch between the experiment needs or design approaches by modelers versus biologists. For example, Yeast now have >20,000 measured protein-protein, protein-DNA, protein-small molecule interactions. Similar networks will soon be available for a variety of bacteria and the worm, fly, mouse, and human. There is a pressing need 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 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. Presently there is a lack of data to validate models and simulations and a lack of whole genome/proteome data to construct large-scale models. Most existing models are for small-gene or protein systems.

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

The participants suggested that a competition similar to Critical Assessment of Protein Structure Prediction (CASP) but focused on the 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 that support for their development should be established. There is the need for analysis tools that test the coherency of disparate bodies of data. Network inference tools because most of the problems of concern will come down to modeling the interactions of a number of different interacting species. Visualization tools since good visualization tools allow experts in biology to find patterns or artifacts in the large data sets not easily detected 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. Multi-scale mathematical research on a wide range of dynamical systems both spatial and temporal is needed. This finding concurs with recommendation from the GTL Report on the Mathematics Workshop, March 18 and 19, 2002.

General Infrastructure needs identified in the Modeling and Simulation workshop.

The participants identified a number of needs common across the GTL community and vital to exploring biology problems. As such there solution was suggested to be one of the highest priority efforts for DOE. There is a need for new types of databases (both hardware and operating system) that can accommodate large data volumes and high schema complexity and rapid query retrieval. Along with this research is needed on new scalable storage hardware/software systems that can accommodate petabyte-scale data volumes and provide rapid analysis, data query, and retrieval. With rapid retrieval will come the need for environments for large-scale data analysis on clusters and massively parallel programming technology for tools, libraries, and repositories. Support is needed for the development of re-usable component/middleware for analysis codes. One computational challenge of reverse engineering is to rigorously solve a network model that best matches known data/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/modeling and data collection through experiments should be one integrated process; computation should be a key driver for designing experiments.

Networking and computing hardware are also required across the community. Robust network technologies are needed 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 that the GTL facilities will require.

Recommendations from the modeling and simulation workshop

The modeling and simulation workshop participants identified infrastructure needs that span the entire GTL community as well as some needs specific to GTL modeling and simulation. They recommended support for 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 complexity. It's advanced computing and systems biology facilities and expertise put DOE in a unique position to help develop new modeling and simulation theory and to implement it 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 for model/simulation-driven experimentation is needed.
- Develop algorithms for scalable stiff/differential-algebraic integrators, multi-objective constrained optimization, and multi-parameter bifurcation and sensitivity analysis, statistical graph models, stochastic optimization, and computationally intensive operations.
- 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 both for 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, and 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

Organizer: Grant Heffelfinger
Sandia National Laboratories
July 24, 2003

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. Protein and protein complexes are the biological machinery which carry-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 complexes dynamic behavior. Ultimately a protein's structure provides much more functional information then its amino acid sequence.

The arrival of high throughput genomic sequencing has led to an explosion of genomic information yet 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 quantifies. Other proteins are difficult to crystallize (a requirement for x-ray crystallography methods) due to their 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 can now be sequenced and annotated within days providing the amino acid sequences of a microbe's proteome yet establishing functional annotation of the proteome as a key bottleneck in high throughput microbial biology.

In meantime, computational protein structure prediction has become increasingly powerful with the availability of an increasingly large 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 technique and the potential applicability of such techniques is estimated to be as high as 50-60% 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 dynamics of the protein complexes.

For all of 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 address these challenging problems. For these reasons, the US Department of Energy (DOE), 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/complex structures, especially for the proteomes of microbes with relevance to DOE's missions in energy production, global climate change mitigation, and environmental cleanup.

The State of the Art

Computational prediction/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) and understanding the dynamics 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/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 become 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 is little or no overlap 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 (CASP), Moult et al. (1) and Schonbron et al. (2) 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 using multiple proteins as templates for different regions of the target sequence may improve results but interestingly, using molecular dynamics or molecular mechanics to refine structures predicted by comparative methods has often increased, rather than decreased, the *rmsd* from the experimentally-derived structure. Most agree that a systematic investigation is needed to obtain fundamental insight as to why this is true and thus suggest methods for how comparative modeling can be improved. Finally, and perhaps most telling, comparative modeling still generally predicts structures which are closer to the best available template used for the sequence alignment than 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 is currently applicable to some 50-70% 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 mainly useful 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 *ab initio* methods are 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 ab inito methods.

Finally, understanding the dynamics of protein complexes is essential for specific phenomena such 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 one of two approaches: rigid-body docking and flexible docking, depending on whether or not the models allow the docking regions of the proteins to move or flex during the docking process. While it is well-known that the conformations of docking proteins can experience significant conformational changes, particularly at the docking interface, capturing such molecular phenomena is computationally expensive, hence the usefulness of rigid docking. However, while rigid docking has reached some level of maturity for practical applications, not does flexible docking remains largely beyond our reach, but an increasing amount of data points to flexible docking as the underlying mechanism for important and fundamental cellular processes such as signaling. Meanwhile, like the protein structure prediction problem, hybrid methods continue to emerge as potentially useful approaches to the flexible docking problem such as using sequence-based structure predictions for the protein-interaction surfaces followed by molecular models of the flexible docking process, much like the of comparative modeling followed by *ab initio* refinement for protein structure prediction.

Goals & Challenges

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

Biology-Driven Challenges and Goals

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

1. predict structures with 2A *rmsd* for proteins with 200 residues given 40% amino acid alignment with proteins in the database, and
2. correctly predicted 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 due 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 (e.g., solid state NMR, optical approaches, etc.) methods are needed for a variety of applications beyond predicting structure and understanding dynamic molecular processes. Examples of such needs 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 and/or more accurate molecular information to improve results is still evolving. Ultimately, the general goal of "40% amino acid sequence alignment is sufficient for 2A *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 was advanced as well, "refinement uniformly improves the results of coarse-grained models." Finally, it was 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 significantly enhanced 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 was identified to capture the essential need for employing computational methods to accurately predict biological function and yield fundamental mechanistic understanding these molecular processes.
6. **Functional Annotation or Exploiting Evolutionary Relationships** was seen as a challenge in terms of the current levels of confidence in the predictions of such methods.
7. **Exploiting Peptides Toward the Prediction of Function and Potential Binding Sites** was identified as an essential goal for computational methods to enable the successful development of high throughput experimental proteomics methods where the identification of associations is a key requirement.

Math & Computing Science

Six specific math and computing science issues were identified:

1. **Global Optimization, Sampling, and Statistics** were identified as a broad area needing additional investigation. Examples of needed work put forth included mathematical proofs for discrete representations of model systems and a single performance metric was discussed, "random starting conditions give uniform results (reproducibility)."
2. **Force-fields, Including the Incorporation of Polarizability** were seen as a significant need with the ultimate success metric of "parameterizing a new force-field on 20 residue proteins gives '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** was seen not only as a biology-driven challenge but also a math and computing science challenge, primarily due to the challenges of integrating methods employing significantly different algorithms and approaches. Once again, the performance metrics identified for the biology-drive challenges were seen to be appropriate for judging progress in this area.
4. **Algorithm Development** and
5. **Simulation Methods/Parallel Implementations** were advanced as a math and computing science challenge, primarily in the context of developing new models methods as well as 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.
6. **Domain Parsing (e.g., Large Proteins)**, as defined by the need for handling multiple domains within single large proteins was seen as posing significant math and computing science challenges. This is due to challenges: 1) precipitating a "starting point" for modular structures that CAN be folded, ultimately enabling a "divide & conquer" approach to structure prediction for large proteins, and 2) using experimental data to prioritize modular determination.

Additional Challenges to Computational Protein Structure Prediction

Finally, seven additional challenges were identified which fall outside the definitions of "biology-driven" and "math and computing science-driven." All were viewed to be issues relative to the application of computational tools to science and engineering challenges well beyond biology. As such, these challenges were left to other venues for further discussion.

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

Appendix A: Workshop Attendees

July 22, 2003 Data Infrastructure Workshop

Name	Institution	Email Address
Natalia Maltsev	ANL	maltsev@mcs.anl.gov
Eugene Kolker	Biatech	ekoller@biatech.org
Peter Bond	BNL	bond@bnl.gov
Cathy Wu	Georgetown	wuc@georgetown.edu
Nat Goodman		natg@shore.net
Arie Shoshone	LBL	arie@lbl.gov
Dan Rokhsar	LBL	dsrokhsar@lbl.gov
Nagiza Samatova	ORNL	6sn@ornl.gov
Ed Uberbacher	ORNL	ube@ornl.gov
Gordon Anderson	PNL	Gordon.Anderson@pnl.gov
Deb Gracio	PNL	Debbie.Gracio@pnl.gov
Helen Berman	Rutgers	berman@rcsb.rutgers.edu
Peter Karp	SRI	pkarp@ai.sri.com
Bruno Sobral	Virginia Tech	sobral@vbi.vt.edu
Nancy Slater	LBL	naslater@lbl.gov
Mike Knotek	DOE	m.knotek@verizon.net
Ari Patrinos	DOE	Ari.Patrinos@science.doe.gov
Marv Frazier	DOE	Marvin.Frazier@science.doe.gov
David Thomassen	DOE	David.Thomassen@science.doe.gov
Dan Drell	DOE	Daniel.Drell@science.doe.gov
John Houghton	DOE	John.Houghton@science.doe.gov
Arthur Katz	DOE	Arthur.Katz@science.doe.gov
Gary Johnson	DOE	Gary.Johnson@sceince.doe.gov
Ed Frank	ANL	efrank@mcs.anl.gov
Bertram Ludaescher	SDSC	ludaesch@sdsc.edu
Stephen Elbert	IBM	selbert@us.ibm.com
Krzysztof Fidelis	LLNL	stuff@llnl.gov
Al Geist	ORNL	gst@ornl.gov
Grant Heffelfinger	SNL	gsheffe@sandia.gov
George Michaels	PNL	GeorgeMichaels@pnnl.gov
Reinhold Mann	PNL	ReinholdMann@pnl.gov
Thomas Zacharia	ORNL	zac@ornl.gov
Phil Locascio	ORNL	l15@ornl.gov

July 23, 2003 Modeling and Simulation Workshop

Name	Institution	Email Address
Carl Anderson	BNL	cwa@bnl.gov
Adam Arkin	LBNL	APArkin@lbl.gov
Mike Banda	JGI	MJBanda@lbl.gov
Roger Brent	TMSI	brent@molsci.org
Hamid Bolouri	ISB	hbolouri@systemsbiology.org
Mike Colvin	LLNL	colvin2@llnl.gov
John Doyle	Cal Tech	doyle@cds.caltech.edu
Steve Elbert	IBM	selbert@us.ibm.com
Jean Loup Faulon	SNL	jfaulon@sandia.gov
Kristof Fidelis	LLNL	fidelis1@llnl.gov
Trey Ideker	MIT-Whitehead	trey@wi.mit.edu
Deb Gracio	PNNL	debbie.gracio@pnl.gov
Mike Knotek	Consultant	m.knotek@verizon.net
Phillip Locasio	ORNL	i15@ornl.gov
Larry Lok	TMSI	lok@molsci.org
Natalia Maltsev	ANL	maltsev@mcs.anl.gov
Juan Meza	LBNL	JCMeza@lbl.gov
Ion Moraru	U. Conn.	moraru@panda.uchc.edu
Danny Rintoul	SNL	mdrinto@sandia.gov
Christophe Schilling	Genomatica	cschilling@genomatica.com
Scott Studham	PNNL	scott.studham@pnl.gov
Shankar Subramaniam	UCSD	shsubramaniam@ucsd.edu
Jim Tiedje	Michigan	tiedjej@pilot.msu.edu
Ed Uberbacher	ORNL	ube@ornl.gov
Daniel Van der Lilie	BNL	dvl@bnl.gov
Ying Xu	ORNL	xuy1@ornl.gov
Ari Patrinos	DOE-BER	ari.patrinos@science.doe.gov
David Thomassen	DOE-BER	david.thomassen@science.doe.gov
Marv Frazier	DOE-BER	marvin.frazier@science.doe.gov
Gary Johnson	DOE-ASCR	garyj@er.doe.gov
Dan Drell	DOE	daniel.drell@science.doe.gov
Arthur Katz	DOE	arthur.katz@science.doe.gov
Chuck Romine	DOE-ASCR	romine@er.doe.gov
John van Rosendale	DOE-ASCR	johnvr@er.doe.gov
Marvin Stadolsky	DOE-BER	marvin.stadolsky@er.doe.gov
Noelle F. Metting	DOE-BRE	noelle.metting@science.doe.gov

July 24, 2003 Protein Structure Prediction Workshop

Name	Institution	Email Address
Carl Anderson	BNL	cwa@bnl.gov
Charlie Brooks	Scripps	brooks@scripps.edu
Valerie Daggett	U. Washington	Daggett@u.washington.edu
Tom Darden	NIH	darden@t-rex.niehs.nih.gov
Jim Davenport	BNL	jdaven@bnl.gov
David Deerfield	PSC	Deerfield@psc.edu
Steve Elbert	IBM	selbert@us.ibm.com
Jean Loup Faulon	SNL	jfaulon@sandia.gov
Kristof Fidelis	LLNL	fidelis1@llnl.gov
Angel Garcia	LANL	axg@lanl.gov
Al Geist	ORNL	gst@ornl.gov
Mike Knotek	Consultant	m.knotek@verizon.net
Phillip Locasio	ORNL	i15@ornl.gov
Leslie Kuhn	Michigan St.	kuhnl@msu.edu
Ben McMahon	LANL	mcmahon@lanl.gov
John Moult	U. Maryland	moult@umbi.umd.edu
Ruth Nussinov	National Cancer Inst	ruthn@ncifcrf.gov
Carlos Simmerling	Stony Brook	simmerling@sunyb.edu
Jeffery Skolnick	U. Buffalo	skolnick@buffalo.edu
Fred Stevens	ANL	fstevens@anl.gov
Chang-Sung Tung	LANL	ct@lanl.gov
Bill Wedemeyer		Bill_wedemeyer@usa.net
Ed Uberbacher	ORNL	ube@ornl.gov
Todd Yeates	MBI	yeates@mbi.ucla.edu
Ying Xu	ORNL	xuy1@ornl.gov
David Thomassen	DOE-BER	david.thomassen@science.doe.gov
Marv Frazier	DOE-BER	marvin.frazier@science.doe.gov
Gary Johnson	DOE-ASCR	garyj@er.doe.gov
Dan Drell	DOE	daniel.drell@science.doe.gov
Arthur Katz	DOE	arthur.katz@science.doe.gov
John van Rosendale	DOE-ASCR	johnvr@er.doe.gov
Marvin Stadolsky	DOE-BER	marvin.stadolsky@er.doe.gov
Noelle F. Metting	DOE-BRE	noelle.metting@science.doe.gov

Appendix B Agendas

July 22, 2003 Data Infrastructure Workshop

Tuesday, July 22, 2003

8:30	Welcome, Introductions, and Mission of the Workshop (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 goals for the community in GTL data infrastructure
10:00	Break
10:30	Half the participants present their vision of the key data issues for GTL and describe how it complements, adds to, 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	Summarize key points made by the participants
3:00	Break
3:30	Discuss the creation of whitepaper incorporating the results of the workshop
5:00	End

July 23, 2003 Simulation and Modeling Workshop

Wednesday, July 23, 2003

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 their single slide on their vision of the key modeling and simulation issues for GTL.Each followed by short discussion by attendees
12:00	Working Lunch (provided)
1:00	Continued one slide presentations
2:30	Summary of key points made by the participants
3:00	Break
3:30	Discussion of process for development of workshop report, assignments for workshop participants
5:00	Adjourn

July 24, 2003 Computational Protein Structure Prediction Workshop

Thursday, July 24, 2003

7:30	Continental Breakfast
8:00	Welcome, Introductions, and Mission of the Workshop
8:15	Overview of the 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 & discussions of the key technical challenges of computational protein structure prediction for GTL
11:45	Lunch
12:15	Single slides: visions & discussions of the key technical challenges of computational protein structure prediction for GTL
2:00	Summary of key points made by the participants
2:30	Break
3:00	Discuss the creation of whitepaper incorporating the results of the workshop
4:00	Adjourn