GENOMES *to* LIFE
BIOLOGICAL SOLUTIONS
FOR ENERGY CHALLENGES
U.S. DEPARTMENT OF ENERGY
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

http://DOEGenomesToLife.org/compbio/

Report on the Computer Science Workshop for the Genomes to Life Program

U.S. Department of Energy
Gaithersburg, Maryland
March 6–7, 2002

Workshop Organizers
Ray Bair, Pacific Northwest National Laboratory
Gary Johnson and John C. Houghton, U.S. Department of Energy
Peter D. Karp, SRI International
Rick Stevens and Bill Gropp, Argonne National Laboratory

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

January 2003

Office of Science
U.S. DEPARTMENT OF ENERGY

Table of Contents

Report on the Computer Science Workshop for the Genomes to Life Program[1]

U.S. Department of Energy
Gaithersburg, Maryland
March 6–7, 2002

Executive Summary

On March 6–7, 2002, the U.S. Department of Energy (DOE) sponsored a 2-day workshop on computer science for the Genomes to Life (GTL) program. About 50 researchers from universities, national laboratories, research institutions, DOE, and industry attended. The objective was to bring together experts in computational and experimental biology with researchers in bioinformatics and computer science to address the following objectives:

- Discuss computational-science research issues and approaches,
- Identify key computer science challenges for DOE's GTL program, and
- Develop recommendations about how computer science can contribute to major thrusts in the GTL program.

Each day began with presentations by speakers from government agencies, academia, and industry. The intention was not only to outline the daunting challenges in systems biology but also to inspire workshop participants to formulate a program for advanced biology research. Five breakout groups were established: (1) genome annotation, (2) protein expression and proteomics, (3) technical text mining for biological data, (4) simulation tools for cell networks, and (5) interoperability facilitation. At the end of the day, a representative from each group presented a status report on the state of the art, approaches and obstacles in the specific area, and a list of recommendations for future work. These presentations were followed by an open discussion of cross-cutting issues and next steps.

Issues

- Participants agreed that high-performance computing has fundamentally changed the way biologists do science.
- The use of parallel computing systems has enabled high-throughput genome analysis and even comparative analysis.

[1] This report was produced from the best available notes and does not represent a verbatim or consensus document of the workshop.

- Protein chips have offered a viable technology for proteomics studies.
- Modern search engines are allowing access to unprecedented amounts of biological data.
- Cell models have made possible quantitative predictions of metabolic pathways.
- A number of frameworks and tools are starting to support component-based software development.

Nevertheless, a number of outstanding issues were identified.

Genome annotation. A clear need exists for computing systems that automatically produce genome annotations, including protein-function predictions, at a much higher accuracy than currently possible. Annotation of such other features as operons, promoters, transcription factor binding sites, SNPs, and protein complexes must be automated as well. More effective methodologies are needed to validate function predictions, encode the expertise of human annotation experts, and apply confidence levels at multiple levels of granularity. Systematic revisions of outdated genome annotations are required to correct predictions or generate predication for previously unidentified proteins.

Protein expression and proteomics. The proteome is far more complex than the genome; for example, there are at least 300,000 proteins encoded by only about 30,000 genes. Integrating data and uncovering associated regulatory networks will require new methods for pattern discovery and for assigning confidence measures to the resulting computational models. Furthermore, specialized visualization systems will be essential for displaying protein-interaction networks, mapping data to pathways, and examining computational results from cluster analysis.

Technical text mining for biological data. A huge amount of critical biology literature is simply not able to take advantage of modern search-engine technology. Further research is needed to understand how—and indeed whether—relevant technology from text-data mining and natural language processing can be applied effectively to biology. Tools also must support semantic interoperability; key issues involve lexicography, semantics, syntax, and recovery of information implicit in context.

Simulation tools for cell networks. Today’s simulations are limited to subsets of processes in individual cells or simple cellular interactions. For more comprehensive modeling, fundamental research issues must be addressed, including representation of multiple levels of spatial and temporal scales in cellular systems and coupling of modeling and simulation with mathematical analysis and experimental databases.

Facilitating interoperability. Component-based architectures are essential for a cross-disciplinary project such as Genomes to Life. Groups such as the forum on Common Component Architecture (CCA) have been developing standards to support a scalable component-based architecture; their work should be adopted and extended. Equally important is further research in data discovery and analysis; the scale and heterogeneity of GTL data sources will require interoperability within and without the GTL program, extensible schemas, and multimodal representations.

Computer Science Challenges

To address these problems, workshop participants formulated a number of specific challenges that require computer science advances, broadly summarized here by topic. The group recommended development of the following:

Data Representation

- Next-generation genome-annotation system with accuracy equal to or exceeding the best human predictions
- Mechanism for multimodal representation of data

Analysis Tools

- Scalable methods of comparing many genomes
- Tools and analyses to determine how molecular complexes work within the cell
- Techniques for inferring and analyzing regulatory and signaling networks
- Tools to extract patterns in mass spectrometry data sets
- Tools for semantic interoperability

Integration Methods

- Methods for integrating dissimilar mathematical models into complex and integrated overall models
- Tools for semantic interoperability

Visualization

- Tools to display networks and clusters at many levels of detail
- Approaches for interpreting data streams and comparing high-throughput data with simulation output

Models

- High-performance, scalable algorithms for network analyses and cell modeling
- Methods to propagate measures of confidence from diverse data sources to complex models

Validation

- Robust model and simulation-validation techniques
- Methods for assessing the accuracy of genome-annotation systems

Standards

- Good software-engineering practices and standard definitions (e.g., CCA)
- Standard ontology and data-exchange format for encoding complex types of annotation

Databases

- Large repository for microbial and ecological literature relevant to GTL
- Big relational database derived by automatic generation of semantic metadata from the biological literature
- Databases that support automated versioning and identification of data provenance
- Long-term support of public sequence databases

Projects

- Series of challenge evaluations to track the state of the art in text processing, data mining, and annotation, as applied to biology
- Collaboratory pilot project in biology (similar to SciDAC projects)

Introduction

Built on the continuing successes of international genome-sequencing projects, the U.S. Department of Energy is developing the Genomes to Life (GTL) program, which takes the logical next step toward understanding the composition and function of the biochemical networks and pathways that carry out the essential processes of living organisms. GTL sets forth an aggressive plan designed to exploit high-throughput genomic strategies and centered around four major goals:

- Identify and characterize the molecular machines of life—the multiprotein complexes that execute cellular functions and govern cell form.
- Characterize gene regulatory networks.
- Characterize the functional repertoire of complex microbial communities in their natural environments and at the molecular level.
- Develop computational methods and capabilities to advance understanding of complex biological systems and predict their behavior.

The Genomes to Life program involves a new approach to biology. It combines large experimental data sets with advanced data management, analysis, and computational simulations to create predictive models of microbial function and of the protein machines and pathways that embody those behaviors. The program's computational component will require developments ranging from more efficient modeling tools to fundamental breakthroughs in mathematics and computer science (CS) to algorithms that efficiently use the fastest available supercomputers. Vast sets of genome sequences, protein structures, interactions, and expression profiles will be generated by this and other biology initiatives. The information must be annotated and archived to provide raw data for computer models of biochemical pathways, entire cells, and, ultimately, microbial ecosystems.

A long-term goal of the computational-modeling section of Genomes to Life is to develop the next generation of methods for simulating cellular behavior and pathways. Other goals are to create molecular-modeling and bioinformatics tools for studying multiprotein complexes, along with new computational methods to explore the functional diversity of microbes. In addition to developing new technologies, Genomes to Life will leverage information and methods from a variety of sources, including cell-systems data from the DOE Microbial Cell Project (forerunner of GTL), protein structures produced in the NIH Protein Structure Initiative, Protein Data Bank, databases of such metabolic processes as MetaCyc and WIT, and a host of available analytical tools in areas such as molecular dynamics, mass spectrometry, and pathway modeling and simulation.

Successful production of advanced tools for computational biology will require the sustained efforts of multidisciplinary teams, teraflop-scale and faster supercomputers, and considerable user expertise. This task for the entire biology community will involve many institutions and federal agencies, particularly the National Institutes of Health and the

National Science Foundation. A central component of Genomes to Life will be the establishment of effective partnerships with these and other agencies to ensure that computational tools and standards are widely adopted and to eliminate redundant efforts.

Challenges in biology bring with them an equally daunting set of requirements in computer and information sciences. Successful systems biology combines an extraordinary array of complex data from experiments, models, and publications. The heterogeneity, complexity, and dynamic nature of this information present CS demands unlike those of any scientific domain before. Likewise, orchestrating the flow of parameters to and from biological models of myriad systems at multiple scales presents new CS challenges in architecture and component design. In many respects, the success of systems biology in general and Genomes to Life in particular rests on a coordinated set of research advances in biology, computer science, and mathematics as well as the establishment of a community computing and information infrastructure necessary for collaboration.

Computer Science for Genomes to Life Workshop

Recognizing the essential contributions of computer and information sciences to advances in the biological sciences, DOE's Genomes to Life program involves a major partnership in the Office of Science between the Office of Biological and Environmental Research and the Office of Advanced Scientific Computing Research. Beginning in August 2001, a series of workshops has been refining plans for this program. Three workshops between January and March 2002 explored GTL requirements for computing advances and their implications for research.

The purpose of the March 6–7, 2002, Computer Science for Genomes to Life workshop was to develop recommendations about how computer science can contribute to major thrusts in GTL. Interest is in both the research needed to provide technologies critical to the program's success and in related concepts that might be applied to biological systems. Consistent with this primary objective, workshop participants also were charged with identifying signature computer science topics in which DOE should establish major research and development. The bias in this second objective is toward critical CS research for Genomes to Life goals that are not being pursued in existing DOE programs.

The workshop brought together visionary researchers in computational and experimental biology, bioinformatics, and computer science. Experts in biology problems mingled with experts in new computing technologies, with the intent of sharing views and developing a consensus on key CS challenges for DOE's GTL Program. The 49 participants were drawn from federal laboratories, DOE, universities, research institutes, and industry. The workshop featured motivational presentations and breakout groups in some of the most challenging areas at the intersection of next-generation biology and computer science, as well as background information on the underlying research programs. Five breakout groups ran concurrently over a day and a half.

- **Genome Annotation:** This is a broad topic, so we set the focus on the implications for the distinguishing pursuits of GTL science (e.g., high throughput, input to models, quickly understanding a new genome, methods to integrate information from multiple annotation methods to predict function, methods to identify multifunction proteins, and new visualization approaches). These connect strongly to a range of CS research issues in data representation, probabilistic integration of evidence, and analysis.
- **Protein Expression and Proteomics:** This arena often involves very large quantities of data, but analysis techniques are still maturing. High throughput is essential, and new techniques for visualization and analysis are much needed. A significant challenge is to provide input to systems models.
- **Technical Text Mining for Biological Data:** This area focused on literature mining (e.g., automatic or semiautomatic extraction of information and knowledge from existing archives of biology literature). Computer science issues include text-data and

natural-language mining techniques and specific issues relating to biology literature (e.g., lack of well-defined ontologies).We were interested in issues specific to both the literature and GTL goals.

- **Simulation Tools for Cell Networks:** Although work has been going on for some time, new classes of models are emerging that are likely to have very different characteristics. Scalability will be important, as will a model architecture that can accommodate many different submodels. Important work needs to be done in automated model development, simplification, and analysis as well as compartmentalization. Work is needed in analysis techniques and tools for understanding mathematical (analytical) and numerical network properties. This area also includes CS ideas about databases of pathways; algorithms for pathway comparison, characterization, and analysis; and techniques to visualize and represent pathways and networks.
- **Facilitating Interoperability:** Enabling integrated and extensible software will be crucial for problem-solving environments in biology. Good interoperability depends upon a number of interrelated design elements, many of which are ongoing CS research topics (e.g., database development, data standards, component technologies, and problem-solving environment middleware). This area also addresses database issues of using information across experiments, models, scales, and disciplines.

Each breakout group was provided with a starting list of questions particular to its area and was encouraged to revise and augment the list. As output, each group prepared a list of findings and recommendations, which were presented to all workshop participants, discussed in open session, and then summarized for this report. The following sections are the reports of the five workshop breakouts.

Genome Annotation

Group members: Breakout Lead, Peter D. Karp, SRI International; Ian Paulsen, The Institute for Genomic Research; Lei Liu, University of Illinois at Urbana-Champaign; Andrey Gorin, Oak Ridge National Laboratory; and Evgeni Selkov, Argonne National Laboratory

Massive amounts of genome data have been generated through projects funded by DOE and other agencies, and even larger amounts will be forthcoming under the Genomes to Life project. The data's usefulness for biologists and bioinformaticists is directly proportional to the quality and accessibility of its annotation. In particular, the quality of genome annotation will directly impact the success of the GTL program. Contrary to some perceptions, genome annotation is not a "solved problem," and a variety of important challenges remain. In the context of GTL, genome annotation encompasses the identification of all genes and functional prediction of their gene products, characterization of other genome features such as operons, analysis of genome structure and evolution, prediction of the cell's biochemical and genetic networks, and representation and visualization of the annotated genome.

This section describes the capabilities of current genome-annotation systems and considers problems and limitations that must be overcome to yield a next-generation system to drive the processing of the increasing deluge of genome data.

State of the Art in Genome Annotation

Current genome-annotation systems (GAS) typically consist of a high-throughput computational-analysis pipeline that runs gene-finding software to identify genes, applies search programs (e.g., BLAST and HMMer) to identify sequence similarity to proteins or protein families, and executes programs that identify other genomic features such as tRNAs, operons, and terminators. GAS may also make use of comparative genomic approaches that compute gene synteny and paralogous protein groups. Integration of these multiple data provides functional predictions for many of the identified proteins. Typically, program outputs are stored either in a relational database or as flat files, and human annotators perform manual synthesis and refinement before release to the scientific community. Additional stand-alone systems have been used to predict an organism's metabolic networks based on genome data. Genome annotation typically is made available to the scientific community through GenBank, second-generation comparative genomic databases such as the TIGR Comprehensive Microbial Resource and WIT, and organism-specific databases such as EcoCyc and *Saccharomyces* Genome Database.

The current generation of genome-annotation systems combines automated prediction of protein function with manual review, correction, and refinement.

- Current GAS require about 2 days to perform automated processing for a microbial genome of 4000 genes.
- Manual review, correction, and refinement require 20 to 40 person-weeks.

Few scientific studies have been done regarding the accuracy of genome annotation, despite the importance of annotation to extracting biological meaning from a genome.

- Most microbial genome annotations contain predicted functions for 50 to 60% of the genes in the genome.
- Accuracy of final function predictions for these genes is unknown.
- Increase in accuracy by manual refinement over automated processing is unknown; one estimate is that automated processing is 70 to 90% as accurate as the final function predictions.
- The relative accuracy of different GAS is unknown.
- There is no agreed-upon procedure for assessing GAS accuracy.

Final structure of genome annotations:

- Protein-function predictions in genome annotations are encoded as free-text strings, not in a controlled vocabulary.
- The emerging standard controlled vocabulary for genomes—Gene Ontology (GO)—is still immature and needs improvement.

- Different genome-annotation groups use different evidence thresholds to determine when they are confident enough to infer a protein function.
- Genome annotations do not include confidence values on function predictions or other annotation features.
- Genome annotations do not differentiate between computationally and experimentally determined function predictions.

Current GAS are based exclusively on recognizing already known biochemical functions in new genomes.

Although improved computational-annotation methods and improved sequence databases are emerging continuously, there is little reannotation of already annotated genomes to revise incorrect predictions or generate predictions for previously unidentified proteins.

Other annotations in addition to protein function:

- The majority of current GAS do not annotate other genome features such as operons, promoters, and transcription-factor binding sites.
- Network annotation to predict metabolic, signaling, and genetic pathways is not a standard part of GAS.
- Separate programs for annotation of metabolic networks typically require under an hour for computational analysis and about a week for subsequent manual review and refinement.
- Computational annotations of metabolic networks do not produce confidence values.
- Accuracy of computational annotations of metabolic networks is unknown.

Requirements for a Next-Generation Genome Annotation System

The GTL program requires a next generation of GAS that overcomes a number of current limitations and shortcomings and produces genome annotations, including protein-function predictions, whose accuracy is on a par with or exceeds that of the best human experts. Automated predictions will

- Achieve accurate high-throughput annotation by decreasing the current manual-refinement bottleneck.
- Decrease the subjectivity and uneven quality of manual revisions and refinements made by different annotation groups with varying levels of expertise.

GTL should develop methodologies for assessing the accuracy of GAS and apply them to current and next-generation GAS:

- Measure accuracy in a quantitative fashion through
  - Experimental validation of function predictions and
  - Benchmark annotation sets

- Measure the relative accuracies of different GAS to identify which performance most needs improvement.

To achieve or surpass the performance of current human annotation experts, next-generation GAS probably will need to use several approaches:

- An expert-systems methodology for encoding the expertise of human annotation experts.
- Fusion of evidence generated from systematic application of multiple analysis tools, including
  - Multiple types of heuristic sequence-similarity searches (each method achieves speedups through heuristics that limit its sensitivity; combining multiple methods increases overall sensitivity).
  - Various types of sequence-motif searches.

    Phylogenetic analyses such as the evolutionary trace.
  - Protein-structure prediction methods such as threading and ab initio modeling.
  - Position-specific structural alignment incorporating knowledge of enzymatic function.
  - Proper processing of multifunctional and multidomain proteins that identifies the protein region with which a predicted function is associated.

Final form of genome annotations:

- Protein-function predictions in genome annotations must be expressed in a controlled vocabulary such as GO.
- GTL should support development of controlled genome-annotation vocabularies such as GO.
- Genome annotations must include confidence values on function predictions, expressed in an agreed-upon confidence scale.
- Genome annotations must differentiate computational function predictions from experimentally determined predictions.
- Genome annotations should include an explanation of the reasoning by which a function prediction was made.
- Function assignments and associated levels of confidence should be explicitly recorded at multiple levels of granularity; for example, a given annotation might state with 95% confidence that a protein is a kinase but with only 10% confidence that the protein is a pyruvate kinase.
- A standard ontology and data-exchange format must be developed for encoding the more complex types of annotation that will be produced by next-generation GAS.

Genome-annotation accuracy is dependent on the quality of public sequence databases. Long-term support for databases is essential.

Other issues:

- Next-generation GAS should include computational methods for postulating unknown functions that have not previously been observed and known functions for which no sequences have previously been observed.
- Next-generation GAS should include methods for systematic reannotation of outdated genome annotations.
- Next-generation GAS should provide tight linkages and feedback among different levels of the genome-annotation process including gene finding, prediction of protein function, operon prediction, and network reconstruction.
- Some amount of manual review and refinement will be required for some time to come. Improved user interfaces for manual annotation should be developed.
- Comparative analyses across genomes play a critical role in genome annotation. New tools are required for comparing tens of very closely related genomes and hundreds of diverse genomes.

Next-generation GAS should produce many types of annotations in addition to protein function, including

- Operons, promoters, transcription-factor binding sites.
- Single nucleotide polymorphisms (SNPs).
- Protein complexes.
- Network reconstruction for metabolic, signaling, and genetic networks.
- Predictions made by all the preceding computational methods should include confidence values, and all should be validated to assess their accuracy and drive improvements.

Protein Expression and Proteomics

Group Members: Breakout Lead, William Cannon, Pacific Northwest National Laboratory; Chris Ding, Lawrence Berkeley National Laboratory; Jim Glimm, Brookhaven National Laboratory; Betty Mansfield, Oak Ridge National Laboratory; Reinhold Mann, PNNL; and Daniel Drell, DOE

Findings

The current state of proteomic analysis reflects the complexities and immaturity of the field. Many experimental approaches and measurements fall under the heading of proteomics, but no one approach or technology has dominated the field. The possible exception is mass spectrometry (MS), which is the main detection technique to determine protein expression and interactions. Current mass spectral approaches can be cast into either tandem or standard MS.

Roughly speaking, tandem MS deals with peptide identification, while standard MS deals with protein identification and quantitation. Both approaches use existing sequence databases extensively in data analysis. Reliance on sequence databases has led to increased efficiencies in data analysis but at the same time has introduced inflexibility into the analyses. For instance, proteins can exist in many combinations of states including not only post-translational modifications but also multiple states resulting from alternative splicing, frame shifts, existence of leader sequences, and existence of pre- and pro- forms. DNA sequence information from databases, however, provides no clues as to when or which of these alternate forms will be present. Due to the sophistication of protein biology, the central dogma of gene-to-mRNA-to-protein has been recast into a complex interacting web.

Built on the widespread use of DNA microarrays for gene-expression analysis, protein chips are another viable proteomics technology that is seeing rapid development. The main pattern-recognition methods include supervised and unsupervised learning. Supervised learning predicts the target patterns' class based on known structures in training samples. Unsupervised methods attempt to automatically discover new data phenotypes, which are important in light of the expected high volume of data. Although many algorithms have been developed, including cluster algorithms, dimension reduction, and feature selection, many critical issues remain. Some examples are method robustness, selection of cluster number, and heavy dependence on initial starting configurations. The challenge is to further develop methods to build mature, robust, automatic analysis tools for proteomics. An active research area, unsupervised learning may also be useful in gene annotation, biological text–information processing, and many aspects of scientific-data analysis.

Due to the complexity of the proteome, data integration will be a significant issue. The proteome is more complex than the genome simply by the number of players involved. For example, if there are 30,000 genes, there are at least 300,000 proteins. The sheer numbers and incomplete nature of experimental techniques in this area mean that clean, precise data will be the exception for the near future. Reflecting this complexity, current analysis techniques focus on particular technologies and are not general. When attempts are made to integrate data from multiple sources, data generally will be nonconforming and not easily amenable to assigning confidence measures to computational models that arise.

With regard to protein-protein and macromolecular complexes observed in the cell through the use of MS or other means, methods are needed to determine how the complexes serve the cell. Are specific complexes involved in metabolic processing, signal transduction, DNA regulation, and other cellular processes? The protein "complexome" should be integrated with gene expression, protein expression, and metabolomics.

Network inference tools are necessary to discover the regulatory network based on protein-expression data. Bayesian networks can, in principle, detect causal relations among different genes or proteins. Time-series analysis can detect all overall patterns of relations among proteins. These analyses help to provide initial knowledge on protein-protein interactions. The computational challenges are the effective adaptation of cutting-edge machine-learning methods. A particular challenge is how to effectively incorporate human knowledge into the machine-learning framework.

Docking algorithms can be used as tools to determine the binding sites' location and the structure of the complexes. Accurate protein structures are the basis for determining binding energies in energy-minimization calculations, which may take advantage of statistically derived potentials. Reaction rates can be determined through Brownian dynamics simulations. Applications of these methods are fundamental to modeling complexes at the structural biochemical level, including such specific DOE interests as analysis of rate-limiting reactions in the microbial modification of uranium oxidation state and carbon-sequestration processes.

These analyses also must provide information to the biologist in a transparent manner. Visualization technologies are currently lacking but will be crucial in this area, particularly for examining computational results from cluster analysis, network inference, mapping gene-expression data to metabolic pathways, and mapping proteins onto the various abstractions that biologists use. As reflected in current gene ontologies, these abstractions include but are not limited to molecular function, biological process, and the cellular role of a particular protein.

Recommendations

- Develop and refine tools to map hierarchies of information from peptides and their post-translational modifications to proteins and then to open reading frames. These may be supervised- or unsupervised-learning techniques, algorithms derived from artificial intelligence research, and other approaches. The tools must statistically integrate data from diverse experimental sources and propagate measures of confidence from individual analyses to an overall model.
- Tools also will be needed to analyze diverse MS data to quantitate peptide and protein-expression levels so the final measures can be compared to DNA microarray data. These tools must statistically integrate data from diverse experimental sources and propagate measures of confidence from individual analyses to an overall model. This includes the development of advanced methods for pattern discovery using unsupervised-learning methods, with emphasis on robustness, maturity, and completeness. Additionally, future extraction of patterns in MS data sets will require appropriate tools.
- Investigate the effectiveness of network inference methods in protein- and gene-expression modeling to provide initial causal patterns between genes and proteins for guiding further experimental and computational analysis.
- Develop and refine tools for proteomic analyses that can computationally determine from experimental data how observed molecular complexes serve the cell (whether the complex is involved in gene regulation, protein regulation, metabolic processing,

or other biological processes).

- Complex membership depends on binding constants, which span the range from irreversible to nonspecific binding. Tools to define what it means to be part of a complex may include protein-docking algorithms to assign a confidence or coherence score to complexes or complex classification. In addition, development and enhancement of protein-protein docking and binding energy determinations will help in understanding protein molecular machines on the macrostructure level, especially when only low-resolution structures may be available. Similarly, algorithms are needed to allow a broad understanding of biomolecular reaction rates and kinetic pathways from a simulation or theoretical point of view.
- Computational tools must help define the nature of protein interaction networks by examining experimental data to determine the meaning of edges in an interaction network.
- Visualization tools to display both networks and clusters should be able to
  - Drill down to raw data.
  - Link nodes and proteins to structured knowledge contained in a database.
  - Be easily expandable (i.e., visual layout should not change when a node or protein is added).
  - Zoom in and out with dynamic and intelligent scaling of node sizes so protein names and annotation can be read.
  - Display structured knowledge in nodes and annotate node clusters with respect to biological function.

Technical Text Mining for Biological Data

Group Members: Breakout Lead, Lynette Hirschman, MITRE; David Israel, SRI International; Hinrich Schuetze, Novation Biosciences; Jim Sluka, InPharmix; and Sylvia Spengler, National Science Foundation

Findings

Biology literature is the central repository of our knowledge of biology. Biologists rely on literature access to identify who else is working on a given problem, learn what has been discovered in previous experiments, and build on this work in their own experiments. Although we are seeing increasing requirements for submission of sequence data to specific databases, these are of limited use without extensive annotations as to source, experimental conditions, form, and function. Where possible, the annotations are expressed in terms of a controlled vocabulary or nomenclature, preferably linked to an ontology. These annotations, however, are still expressed in natural language, particularly when the relations and processes discussed are complex. This means that a huge amount of extraordinarily valuable information is stored in natural-language form in the literature and, increasingly, in annotations.

MEDLINE contains over 11 million abstracts (mid-2001), and its rate of growth is increasing. But MEDLINE represents only a fraction of the literature that biologists need to access; it contains only abstracts, not the full text of articles, and its primary focus is on human biology and medical issues. Much critically important literature on microbes is not covered by MEDLINE.

Modern search-engine technology is providing impressive access to distributed information on the Web. The biological literature, however, is not well covered by these general search engines. And where there are important repositories such as MEDLINE, search and indexing tools rely on Boolean keyword search with only limited lists of synonyms available. The result is that a search on a given gene is likely to return a large number of irrelevant or redundant hits and to find only about 30 to 40% of the articles that discuss that gene.

Finding 1: GTL research requires access to the literature, particularly the microbial literature and annotated sequences, in ways that are not well supported by current infrastructure.

At the same time, computer scientists have made significant progress in developing techniques to provide better access to the information in the literature. These techniques include the following.

- **Information retrieval or document search.** The user creates a query, and the system returns a ranked list of documents (or passages) in decreasing order of relevance to the query.
- **Information extraction.** From running text, the system can extract lists of entities and relations among them for improved indexing or database creation.
- **Text-data mining.** Terms appearing more than once can be used to create clusters of related entities (e.g., all pairs of gene-product terms that occur in the same sentence might be hypothesized to be involved in protein-protein or gene-protein interactions).
- **Question answering.** Users can query in natural language such as “What are all the organisms that can live at 150°C?” and receive an answer, including pointers to sources that contributed to the answer.

We know that these techniques work, at least for domains such as general news reporting. Information-access and -retrieval techniques have proved effective in selecting documents relevant to a specific topic or in providing answers to questions based on information located in document collections. The best search engines can provide 70 to 80% accuracy for the first 5 to 10 documents retrieved [Text REtrieval Conferences (TREC-9) results]. For question answering, the leading systems can provide correct answers to simple factual queries at 75 to 85% accuracy (cf. TREC-9 results).

Results from other evaluations show that information-extraction systems can identify and classify entities such as person name, organization, and location at an accuracy of greater than 90%. Commercial systems now do this for multiple languages, and these are increasingly being incorporated into search engines and other systems that require indexes or summaries. Information-extraction systems also can successfully extract binary relations among entities such as *ORGANIZATION located_at LOCATION* or *PERSON works_at ORGANIZATION* at over 80% accuracy [cf. Message Understanding Conference (MUC-7) results]. To date, extraction of more complex relations or events has proved more difficult; the best systems have reached an accuracy of only about 60%. In general, extraction systems are most accurate when their models rely on local information (adjacent words). Entity extraction is easier because clues to an entity's type are generally within a few words. The more complex the relation, the more it is necessary to look beyond the immediate phrase or sentence.

There has been increased interest in applying these techniques to biology, but to date the results are scattered and impossible to compare because there have been no standard data sets or challenge evaluations. Thus one researcher may report a precision of 92% and recall of 21% in determining subcellular localization of a gene product from the literature, while another may report a precision of 90% and recall of 57% on extraction of gene-inhibition relations (i.e., relations that explicitly use the word *inhibit* or a morphological variation such as *inhibits*, *inhibition*, or *inhibiting*). When tasks are so different, measurements cannot be compared. As a result, we do not know how well or poorly these techniques will work for biological applications.

We do not even know whether biology will be easier or harder than newswire. On the one hand, it may be easier because of existing nomenclatures and ontologies (at least for some areas). On the other hand, the language of biology is changing daily; for example, 166 "terminology events" occurred in one week for the mouse genome as new names were added or existing names removed. New facts (and therefore, new relations) also are being discovered on a daily basis. In addition, many terms are ambiguous; for example, a single term may designate a gene or a gene product, and most terms have many alternate forms (synonyms) including the full term and multiple abbreviations. Because of the systematic ambiguity and rapid language changes that characterize biology, any natural-language processing system must be able to adapt to—or even "learn"—new terms and relations. A system cannot succeed by memorizing all known terms.

Finding 2: Relevant technology from text-data mining and natural-language processing can be applied to biology, but we do not know yet how well these techniques will perform on biological applications, and, most important, whether they will work well enough to be useful to biologists. Much of the progress in natural-language processing and text-mining domains has come about because of systematic common evaluations conducted in natural-language processing and information retrieval at the two conference series, MUC and TREC. There also is the annual Knowledge Discovery and Data Mining (KDD) Challenge Cup to evaluate data-mining techniques (this year focused in part on text-data mining for biology). These challenges are similar to the very successful Community Wide Experiments on the Critical Assessment of Techniques for Protein Structure Prediction (called CASPs). They have provided a baseline and a measure of collective progress over time, created a research community, and provided for rapid dissemination of research approaches.

Finding 3: A challenge evaluation for text-data mining in biology could push progress in this field and help to match maturing technologies with relevant GTL problems, providing improved access to biology literature.

Needs and Requirements

Text mining and information extraction can contribute to solving many of the key problems identified in the Genomes to Life program.

- **Genome Annotation.** Entity detection and mapping into a standardized nomenclature can greatly improve indexing and speed human annotators in database curation. Information-extraction techniques can identify biological entities and relations among them, such as subcellular locations and key gene or gene-product relations.
- **Protein Arrays.** Text data–mining programs already have been able to identify clusters of related proteins, based on their descriptions in the literature. Such techniques will become increasingly important as the volume of available array data increases.
- **Simulation Tools.** Text-data mining can contribute to linking relations together across articles, leading to hypothesis generation and pathway discovery.
- **Interoperability.** Information-extraction techniques can link mentions of genes, gene products, and locations with their canonical representations in an ontology. This not only greatly improves indexing, searching, and linkages between literature mentions and sequence information but also provides a major step towards true semantic interoperability across multiple "nodes" in a biological computing grid.

Below, we sketch out a roadmap to adapt and scale current technologies to the challenges of GTL problems. The roadmap has five stages, moving from very near term problems that could be tackled within a year to longer-range problems requiring 5 or more years to solve.

- **Nomenclature Mapping.** This can be accomplished by assembling large lists of terms (e.g., gene names or terms for subcellular locations) and their synonyms from a variety of sources, including nomenclatures and existing database synonym lists; extensions to existing ontologies would be required to accommodate a wide range of terms. Nomenclature mapping provides a comprehensive list of synonyms that can be used immediately to improve indexing and search over the literature.
- **Entity Extraction.** Using nomenclature mapping to quickly identify names of entities in text would be relatively straightforward, providing training data for statistical and machine-learning approaches to entity extraction. Such extraction systems could identify new names in running text and map them to the canonical name, thus im-

proving search capabilities and allowing new text-mining applications to be developed through the use of standardized terminology.

- **Extracting Entities in Relations.** Once entities can be reliably extracted, extracting entity-entity relations should be possible. The easiest would be explicit relations between entities occurring in a single sentence, which could then be extended to those spanning several sentences. Relation extraction would provide aids to database curation and support to automated gene annotation based on information from the literature.
- **Extracting Complex Relations.** Relation extraction could be extended to complex or second-order relations. For example, *protein bind_to gene* is a simple relation that can, in turn, be an argument in a complex relation such as *protein bind_to gene stimulate transcription*. Complex relations are needed to represent the complexities of real biological pathways.
- **Hypothesis Generation.** This represents a long-term goal. The tools described above should make it possible to automatically extract, correlate, and combine complex information with other data sources to generate hypotheses with confidence measures.

The problems above pose major research challenges for computer scientists. In general, we believe solutions to these problems are algorithm bound, not compute bound. However, this may be because many machine-learning and pattern-recognition algorithms are starved for data. If we were able to generate massive quantities of training data, training algorithms might become more compute bound. Some key computer science challenges follow.

- **Automatic methods for extraction of complex relations from literature.** Although current systems can extract entities reasonably well, extraction of multiargument complex relations is still an open issue. In addition, the biological literature is potentially more complex because novel entities and relations are constantly being discovered. Thus any system would have to identify novel occurrences of names and relations and "learn" them—that is, insert them into an ontology and store them for validation and future use. The ultimate solution would require a system that can learn by reading text.
- **Semantic interoperability and ontology learning.** Semantic interoperability implies the ability to exchange semantic information across data structures and programs, requiring that we solve the nomenclature-mapping problem and standardize the naming of functional relations and processes and their arguments. Key problems in computational linguistics involve issues in lexicography, semantics, syntax, and recovery of information implicit in context. Tools would make it possible to support communication across "nodes" in a biological computing grid, build or extend ontologies (semi-) automatically, and maintain links from databases to the literature.
- **Integration of text with other data types (sequences, numerical data).** It is critical that biological applications combine evidence of different types from various sources. The literature is one source, but it must be linked to others such as sequence, pathway, and numerical experimental data. The challenge is to explore ways of providing confidence estimates for information mined from the literature, as well as methods of evidence

combination. The database community must ensure that databases support this kind of linkage among disparate types of data.

- **Automatic generation of semantic metadata from literature abstracts.** Given the ability to normalize names and extract relations from the literature, it should be possible to create a set of metadata linked to the online literature (or at least the abstracts). This would provide an enormous resource to biologists, creating a comprehensive index (using standardized terminology) over the entire literature. With sufficient computing resources, keeping up with the rapidly growing literature should be possible, even reindexing as needed to support vastly improved searching, text mining, and question answering.

Recommendations

Based on our findings and an analysis of the needs and requirements of the GTL program, we put forth the following recommendations.

- **Assemble a large repository for the microbial and ecological literature relevant to GTL.** Many of these resources are not available through MEDLINE. Such a collection would provide a valuable tool to GTL researchers. Furthermore, it would provide a test bed for computer scientists that would allow them to experiment with state-of-the-art search engines, full natural-language query interfaces (for question answering), and improved indexing through generation of semantic metadata. These resources could be extended over time to link to other types of data such as sequences and experimental data.
- **Develop tools and resources for semantic interoperability, including a name database, nomenclature mapping, entity extraction, and relation extraction.** The plan would be to start with the relatively easy tools (name database, entity extraction) and build and release tools incrementally over the life of the GTL project. In the long term, it should be possible to tackle harder problems such as extraction of complex relations by building on previously developed tools.
- **Create a big relational database derived by automatic generation of semantic metadata from literature.** Building the database would require the tools developed under the second recommendation above and could be applied to the repository described in the first. It would provide vastly improved access to the underlying literature and would support experiments in database query, literature search, and data mining.
- **Create a series of challenge evaluations to track the state of the art in text processing and data mining as applied to biology.** Such a series should foster communication across groups and help to marry maturing tools to important biological applications. The text-mining evaluation might be associated with a similar one for genome annotation, ensuring that text mining is closely linked to a critical biology application.

Simulation Tools for Cell Networks

Group Members: Breakout Lead, Rick Stevens, Argonne National Laboratory; Co-Lead, Nagiza Samatova, Oak Ridge National Laboratory; Phil LoCascio, ORNL; Mary Anne Scott, DOE; John Ambrosiano, Lawrence Livermore National Laboratory: Buff Miner, DOE; John van Rosendale, DOE; Patrick Lincoln, SRI International; Jean-Loup Faulon, Sandia National Laboratories; John Houghton, DOE; and Evgeni Selkov, Argonne National Laboratory.

Overview

Great predictive value lies in developing the ability to reconstruct and model the networks of molecular interactions at the core of all life processes. Cell networks arise from the series or chains of molecular interactions during metabolism, protein synthesis and degradation, regulation of genetic processes such as transcription and replication, and cell signaling and sensing. In short, cell networks and pathways are at the center of much of what we think about when we discuss cell modeling and cellular behavior. One of the major goals of systems biology is the ability to comprehensively model the complete set of a cell's molecular interactions, an essential requirement for addressing DOE's environmental, energy, and health-protection missions.

Current State of Cell Network Modeling

The primary goal of cell network modeling is to capture in an abstract mathematical model the structure (topology), kinetics, and dynamics necessary to analyze and simulate the behavior of networks present in a particular organism. Models are constructed from a combination of mathematical principles and experimental data (e.g., annotated genomes, proteomics databases, data from in vitro experiments, expression data, and data from the historical literature). Models are used both to facilitate a general understanding of cellular networks and for simulations that attempt to reproduce or predict a particular experimental result.

Experiment (Real Life) ↔ Simulation (Abstract Systems Model)

Current state-of-the-art models can be used to make specific quantitative predictions for limited regions of well-characterized metabolic pathways or a limited set of specific regulatory or signaling circuits. More general qualitative predictions can be made for larger, more complete networks, but the current lack of kinetics constants for most enzymes and of concentration data for intermediate metabolites limits the ability to simulate quantitative results for entire networks. Modeling is also hampered by the incomplete specification of networks due to lack of functional gene assignments, protein complex and association data, and data for regulatory elements and interactions. Bioinformatics techniques are used upstream of modeling and simulation to extract from experimental data the relationships and functions needed for simulation.

Data ↔ Information ↔ Knowledge

Mathematical-analysis techniques are used to further develop, understand, and improve abstract models and our ability to simulate them. A number of software systems have been developed to model and simulate cell networks (e.g., Gepasi, E-Cell, V-Cell, DBsolve, and BioSpice). Several different formalisms exist for representing and simulating cell network models [e.g., rule-based, ordinary differential equation (ODE), logical, and qualitative]. Current cell network simulations are running typically on serial computers (PCs and workstations) and are used mostly to simulate the processes in individual cells or simple cellular interactions.

Data and Bioinformatics Needed to Support Modeling

Multiple data resources needed to support modeling and simulation span the full range of genomic, molecular biology, and cell biology experimental methods. Powerful bioinformatics tools must be developed to integrate data, information, and knowledge across the multiple biological domains important for cell network modeling:

- Functionally annotated genomes
- Protein-protein interactions
- Protein-expression levels and metabolite production for various conditions
- Gene- and transcript-expression levels for various conditions
- Databases of known metabolic, signaling, and regulatory pathways
- Molecular structures and functions
- Protein complexes
- Assays of metabolites

A number of important issues relating to data must be addressed for more effective modeling and simulation. These issues include the nature of queries required by those setting up simulations and building models. The functionality needed in bioinformatics tools to address the requirements of simulation and modeling are those relating to functional integration (i.e., knowing what is connected to what), dynamics (i.e., what changes over time and by how much), and known data points to experimental or environmental conditions (i.e., establishing boundary conditions and forcing). An important related issue is that we need to provide mechanisms in the bioinformatics infrastructure to record and archive the results of simulations (and the models themselves) so researchers can share and leverage the model building and computational experiments of others and resolve important conflicts between experimental data and simulation. Once data from simulations and models become part of the database, curation requirements are more complex.

New capabilities are also required at the query and comparative-biology level. For instance, we wish to query the biological database about properties of known and annotated pathways and compute derived properties of pathways, like limit cycles, and attractors. Visualizing, comparing, and contrasting pathways and associated annotations are important. Models must be represented and stored in a form suitable for archiving in databases, and the ability to derive simulation codes directly from the model database will become important in managing computational experiments and collaborative workflow. New analysis tools and environments are needed that can support comparisons of models with experiments and with each other.

Thoughts About Advanced Modeling Capabilities

One of the ultimate goals for cell modeling is to automatically predict cell phenotype from the cell's genotype and extracellular environment. Such predictions will require automating genome annotation and the prediction of cell ultrastructure, morphology, motility, metabolism, life cycle, and behavior in a wide range of environmental conditions. In this way, models and simulations will represent our ultimate level of integrated understanding. These models not only will be descriptive and phenomological but will also be predictive at multiple levels of detail. Although this ultimate vision is still a distant goal, we can take important steps within our current scope of understanding and create experimental and computational capabilities that will have dramatic near-term impact. Even simple models can be used to help guide experiments, and the results of iterating among theory, simulation, and experiment will enable us to develop (perhaps slowly at first) an integrated understanding of cellular systems. This understanding undoubtedly will be framed initially in some qualitative form, but over time and with additional experiments and improved analysis methodologies, it will become much more quantitative.

Qualitative Models ↔ Quantitative Models

An important need is to be able to quantify the levels of uncertainty in our understanding and predictions and the sensitivity of our models to variations in input parameters and structure. Progress is also needed on important issues of model and knowledge representation and formalism. Many different formalisms exist for representing and modeling cell networks:

- Boolean Models
- Bayesian Networks
- Generalized Logical Networks
- Petri Nets
- Rule Based
- Fuzzy Logic
- Stochastic ODEs
- Deterministic ODEs

No dominant formalism, however, has emerged that can satisfactorily represent both kinetics and dynamics of metabolic networks and the logical structure of signaling and regulation. Much new work is needed in this area. Another critical topic that must be addressed is how best to represent the multiple levels of spatial and temporal scales in cellular systems and incorporate them into models. Most existing models of cellular networks are one dimensional (e.g., box models that assume completely mixed environment). To make progress towards the ultimate goal of accurate phenotype prediction, future modeling schemes need to incorporate 3D modeling and intracellular compartmentalization.

Multiple modeling and inference techniques can address difference classes of problems, each with distinct temporal and spatial scales and each with potentially different computational complexity. Each class of problems has specific data limitations and a diverse set of data sources, as mentioned above. Limitations on the models themselves depend on the levels of abstraction used and the mathematical treatment of the problem. Common challenges facing modeling include:

- Design of effective numerical experiments
- Multiscale domains (e.g., molecules, clusters, networks, membranes)
- Multiscience (e.g., biophysics, biochemistry)
- Complexity of inferring networks from global experiments is NP hard
- Analysis of large-pathways (e.g., flux analysis) is P-space hard
- Development of combinatorial models

Compartmentalized models will become increasingly important, not only for the treatment of eukaryotes, which have multiple cellular compartments with distinct processes enclosed or isolated by membranes, but also for modeling distinct phases of metabolism in organisms such as cyanobacteria, which have both oxygenic and anoxic pathways separated either spatially or in time. Compartmentalized models will be needed to fully model life cycles of prokaryotes, which include systems like sporulation in *B. subtilis*, heterocyst formation in *Anabaena*, and differentiation in myxobacteria. Models with multiple compartments will have to address coupling of compartments (e.g., data and flux representations and stability and fidelity) in a scalable fashion. Much may be learned from the experiences of the ASCI and climate-modeling-coupling groups. Compartmentalization and coupling will also become an issue in multicellular systems (e.g., bacterial communities and multicellular organisms).

A major modeling challenge is the choice and effective exploitation of mathematical abstractions. Biological systems differ from those produced by human engineering in that the hierarchies or functional subsystem modules are not necessarily obvious, yet exploiting modularity or lumping the system may be essential for efficient modeling and simulation.

Strategies for Addressing Complexity

It may be important to leverage techniques from combinatorial mathematics to map the cell network problem to some auxiliary concerns via a combinatorial transformation that may admit a less complex solution. Techniques of parameter estimation from control theory may be used to improve the tractability of the parameter problem for large networks where we have only a few rate constants. Techniques from computer science, including constraint-based modeling, may also find application in reducing the size of the solution space. An important approach to investigate is the feasibility of integrating the environment for mathematical analysis (e.g., bifurcation and complexity) with the simulation environment so that both systems can use the same model representation. Another major challenge is to scale up the modeling of cell networks to include large-scale metabolism and regulation. Large-scale models will need many rate constants; some may be obtained from the literature or databases on enzyme kinetics, others may require new experiments, and some may be estimated or eliminated by parameter-estimation techniques or model transformations that have proven useful in large-scale systems science. Ultimately, large-scale metabolic models must be coupled to comprehensive regulatory models. Hence, new techniques may be necessary to synthesize comprehensive systems models.

Major Findings

1. Current modeling efforts are extremely constrained by a lack of available data, including reaction-rate constants, gene and gene-product functional assignments, and the verification of fluxes and small-molecule inventories.
2. Current computational approaches to eliminate or estimate parameters are mathematically complex and computationally intensive; they may require computational resources at least as large as or larger than the simulation itself.
3. Existing component and modeling frameworks provide a starting point for future cell network simulations, but much investigation will be needed to ensure that they are appropriate for biological problems and cell modeling in particular.
4. Current modeling and simulation environments are decoupled from mathematical-analysis environments. Coupling these environments may dramatically improve productivity of model development.
5. Simulation and modeling should be integrated with experimental data infrastructure. In particular, an integrated database system should be developed that provides a long-term repository of simulation data and modeling runs and supports coupling to bioinformatics databases containing experimental data.

A progression of modeling capabilities (detailed below) will be needed to accomplish GTL goals.

Paths to GTL Simulations

The panel generally agreed on the approach for accelerating GTL simulations. We discussed four barriers to progress: algorithms (A), coupling (C), processing power (P), and software integration (I). Below are listed two parallel development tracks focused respectively on increasing the sophistication of modeling individual cells and multicellular systems. We indicate after each line the factors the panel felt are limiting progress.

Single-Cell Model Progression

1. Unregulated metabolic model
2. Allosteric regulation (binding-induced changes conformation) (A)
3. Gene-Regulated + Metabolic Model (A, C)
4. Heterogeneous/Compartmentalized/Diffusion (A, C, P)

5. Active Regulation + Transport (A, C, P, I)
6. Complete Integrated Cell (geometry) (A, C, P, I)

Multicellular-Model Progression

1. Multicellular models (homogeneous) (P)
2. Multicellular (homo) with complex communication (P)
3. Multicellular (hetero) mixed population (P, I)
4. Multicellular differentiation and motility (A, C, P, I)
5. Multicellular structures with complex geometry (A, C, P, I)

Computer Science Needs

The breakout group identified several needs that will speed up research work in this area:

1. Standard software and mathematical framework for functional composability, including support for multiple modules, time scales, and space scales; and empirical, semiempirical, phenomological, and data-driven models.
2. Software tools for the interpretation of output of complex models, particularly those that exploit advanced scientific visualization or methods of automated interpretation.
3. High-performance scalable algorithms for parameter estimation, graph theory, combinatories, and algorithm analysis.
4. Appropriate software frameworks and systems-level architectures for developing scalable models with support for control and synchronization of multicomponent models.

Recommendations for Computer Science Research

The cell-modeling breakout group recommends the formation of a comprehensive research program that would address the following CS research problems:

1. Development of biologically relevant module-encapsulation methodologies.
2. Techniques and software for linking knowledge and databases with simulation environments.
3. Methods for integrating dissimilar mathematical models into complex integrated overall models (e.g., techniques for addressing stability and efficiency).
4. Fundamental algorithm research and development for network analysis (regulatory and signaling networks in particular).
5. Research into visualization-based approaches for the interpretation and understanding of complex streams of data, in particular, methods to compare high-throughput experimental data with simulation output.
6. Development of robust model and simulation-validation techniques.

7. Development of a scalable simulation information-management system.
8. Investigation of new strategies for experimental validation of models.
9. Algorithms and tools for understanding and improving complex models (e.g., bifurcation analysis, parameter estimation, optimization, and uncertainty analysis).
10. Performance analysis and improved scalability of multicomponent, multiscience models.

Facilitating Interoperability

Group Members: Breakout Lead, Deborah Gracio, Pacific Northwest National Laboratory; Christopher Johnson, University of Utah; Daniel Drell, DOE; George Seweryniak, DOE; Fred Johnson, DOE; Esmond Ng, Lawrence Berkeley National Laboratory; Ann Chervenak, University of Southern California; Terence Critchlow, Lawrence Livermore National Laboratory; Susan Davidson, University of Pennsylvania; Forbes Dewey, Massachusetts Institute of Technology; David Benton, GlaxoSmithKline.

Scope and Purpose

To solve the next generation of complex life science problems will require a systems approach that integrates data across various scales and provides interoperability among computational methods and resources. The development of enabling software tools to support the research processes of systems biologists will be essential and will require leveraging work across research groups, taking advantage of research others have done, and collaborating in a more efficient and productive manner.

Life sciences research takes place in a distributed, heterogeneous, experimental, and computational environment. This heterogeneity extends to the data and resources used to represent the research including data types, metadata, file formats, computing systems, programming languages, concepts, and organizations. Information-systems development and operations productivity can be significantly increased by improving the integration, interoperability, and reuse of computational resources and artifacts. Recognizing that research will always be carried out using diverse information resources and software systems, we must extend our capabilities to tie traditional tools and resources together in innovative ways to support fundamental scientific progress. Interoperation is the mechanism for exploiting the required diversity while providing for integration and flexibility.

The Facilitating Integration and Interoperability breakout session focused on necessary design elements and requirements for developing software tools that are interoperable and provide integration of multimodal data and information across time and scale. We specifically addressed barriers to facilitating interoperability and integration, the current state of software technologies and methodologies for interacting with data, and recommendations to support leveraging work across programmatic efforts and collaborations among the scientific community.

Findings

Computing has significantly changed the way that we practice science. Over the past two decades, researchers have begun to exploit advances in computer hardware and software and in new mathematical and theoretical approaches. There has been a significant shift towards the partnering of theory, experiment, and simulation through the evolution of high-performance computing with enough power to solve complex scientific problems. In this partnering, significant interoperability issues have arisen through differences in semantic understanding, representation of data and information in electronic form, and integration of science across multiple scales and disciplines.

Traditionally, there have been many barriers in the integration and interoperability of software tools, computational methods, and data resources. These barriers include the following:

- Inability to abstract data and information to a level that maintains semantic understanding across disciplines.
- Lack of tools to support discovery across distributed, heterogeneous resources.
- Inability to track information across multiple experiments and associate data with resulting analyses.
- Complexity of scaling data across time and space, ranging from individual molecules to microbial communities.
- Proliferation of various standards efforts that do not reflect the complete biological system.
- Gap between software development of low-level tools and domain-specific applications.
- Lack of software-engineering standards to support an integrated and interoperable environment for a systems biology approach.
- Sociological barriers associated with collaborations.

The overall goal should focus towards distributing usable, accessible software to support scientists in their research so they may leverage their work across research groups, collaborating in a more efficient and productive way. Research will be facilitated with the ability to use collections of codes in a loosely coupled fashion exercised through a problem-solving environment. This will support integration of the experimental data required to determine simulation parameters and to validate results. Only when researchers can leverage the complete scope of resources will they be able to transform this critical information into knowledge.

Interoperation will facilitate both the construction of component-based information systems and the integration of data and information from multiple heterogeneous sources. Component-based software construction can produce higher-quality systems at a lower total cost. Specific benefits of this approach include more efficient operations, increased development productivity, and, therefore, better return on investment with reduced risk. Software development, support, and maintenance costs are reduced because applications can be assembled from preexisting components. Applications can be more robust when components have been rigorously tested and validated before being used. Applications can be more portable among platforms if they make use of platform-independent component frameworks. In addition, if such frameworks are used, each component can be developed and run on platforms that provide the best support for its functionality, independent of platforms for other components. Increased flexibility and more rapid application development supported by component-based approaches can result in improved ability to address critical enterprise-wide issues and faster response to changing research needs. Finally, these approaches facilitate upgrade and exchange of system components, for example, when improved implementations or new algorithmic approaches are developed.

Component-based architectures facilitate data and information integration. This integration is an absolute requirement for a cross-disciplinary project such as GTL. For developers, component-based interoperability has some specific advantages:

- Reduce infrastructure costs
- Reduce development costs by using standardized component technologies internally
- Reduce cost of supporting other developers' formats
- Focus development on core competencies
- Facilitate productive collaborations with developers in other specialties

For genomics and postgenomics researchers, a corresponding set of advantages includes the reduction in internal software-development costs, the software built around widely supported standards, a simplified process for integrating new methods and approaches, and an efficient method for applying novel analytical techniques.

State of the Art: Architectures

Frameworks and tools to support component-based development have been increasing in numbers over the past few years. CORBA, COM, Enterprise Java, .NET, and XPCOM are all examples of frameworks built to support the architectural move toward component-based software development. The good news is that these architectures are open and many are well documented, but they don't address scaling or latency issues when dealing with parallel computing systems. Parallel computing is critical to solving the grand challenge DOE is facing today. To address this issue, a grass-roots effort among multiple DOE laboratories and academic institutions formed the Common Component Architecture (CCA)[1] Forum with the intent of developing a component framework specification to support scientific-computing applications. A subset of this group formed the Center for Component Technology for Terascale Simulation Software (CCTTSS), formally funded through the Scientific Discovery for Advanced Computing (SciDAC) program of the Mathematical, Information, and Computational Sciences (MICS) division within the DOE Office of Science. The intention of CCTTSS is to research software-component technology for high-performance parallel scientific computing to address problems of complexity, reuse, and interoperability for scientific-simulation software.

In its current form, CCA has defined a draft architecture that scales to 7000 processors. CCA is a specification that provides the information required to support interoperation among applications. CCA furnishes bridges to other architectures such as CORBA, IDL, and databases and separates the user interface from the infrastructure.

State of the Art: Data

Data are stored in a variety of forms within the biology community, most typically as flat files with little descriptive information on how the data were collected. Many researchers use Microsoft Excel as their standard for storing and analyzing data. An open-source movement is beginning to formalize community aspects, but this typically is still only within small subsections and has not addressed the complete biological system. This consortium includes new specifications and tools such as CellML, MAML, BioPerl, BioJava, BioXML, and BLASTWrapper. Schemas being developed are inconsistent and changing quickly, sometimes with each experiment. Few mechanisms cross-reference information among databases, making analysis processing very difficult.

Over the past decade, biology databases have proliferated on the Internet to support genomics and postgenomics research. Each of these systems, however, provides its own structure and communication mechanisms. There has been some movement towards syntactic interoperability using XML as an exchange format, and many ontology groups are addressing mechanisms to deal with semantic understanding. New solutions being developed to support database interoperability include the Object Protocol Model (OPM) used by Lawrence Berkeley National Laboratory,[2] semistructured data models used by Stanford,[3] and the WebDAV open-data architecture model used by Pacific Northwest National Laboratory.[4] To address query issues across a federation of databases, MIT has developed a query engine that intelligently directs a single client query against a distributed set of heterogeneous databases.[5]

With the distributed nature of research and efforts to collaborate with colleagues outside organizational boundaries, there is an ever-increasing need to store and access large amounts of data across a network. The Globus Project's Data Grid group is working to identify, prototype, and evaluate technologies required to access data across a distributed, heterogeneous network specifically targeted at scientific applications.[6] Data Grids provide a basic infrastructure and support a variety of scientific domains and disciplines. Some key features include replica management, secure transfer of data using public key technology, and metadata services.

During the computational work process, tracking of intermediary results is beneficial. Lawrence Livermore National Laboratory has developed SimTracker, a software tool that summarizes results by generating metadata automatically, both textual and image snapshots, throughout the research process.[7] The metadata are presented through thumbnail sketches with hypertext links back to applications and data, so the researcher may continue monitoring the calculation as it is running.

Recommendations

This section outlines our recommendations for CS research focused in the area of facilitating interoperability. Although our recommendations target the GTL program specifically, they also address many needs and issues within the general scientific community. Additionally, we have identified a few cross-cutting issues that, while not focused specifically on integration and interoperability, we believe are important to this program.

Cross-Cutting Issues

With funding opportunities available within DOE, Department of Defense, National Institutes of Health, and private agencies, GTL must encourage interagency coordination to prevent overlapping research and duplicative work. This coordination should include programs such as

- Virtual Human
- Wellcome Trust
- Digital Human
- Physiome
- Computer Science programs addressing time and scale issues (ASCI, SciDAC)
- OpenBIO Consortium
- Object Management Group
- Interoperable Informatics Infrastructure (I3C) Consortium

For successful cross-disciplinary approaches and computational tools to solve complex biological problems, training opportunities must support learning across domains. This includes exposing biologists to various software tools for solving potential immediate problems and giving them an understanding of how CS can help facilitate their research and collaborations. In turn, computer scientists will require education in biology terminology, a base understanding of how biological systems are interconnected and function, and, most important, computational-biology approaches and methods. This educational process, which must be accomplished through funded activities, could be sponsored through various tutorials, workshops, and forums at scientific meetings and conferences such as SuperComputing, BioComputing, International Conference on Systems Biology, and GTL workshops.

Software Engineering

- Good software-engineering practices should be defined and adopted by the GTL program from the beginning.
- Support for full life-cycle development should be required, including testing, documentation, and tool maintenance.
- Software architectures that facilitate interoperability and reusability should be extended and adopted (e.g., Common Component Architecture).

- Existing standards should be leveraged and encouraged, and ongoing standard-definition efforts should be supported by GTL projects.

Data and Model Interoperability

GTL should require fundamental research in the following areas:

- Superimposing biological data
- On-the-fly integration of multimodal data
- Development of coordinate systems-similarity measures and overlay techniques for mapping between data
- Query languages and optimization techniques
- Encourage interoperability at the data sources within GTL and mandate that data centers work together.
- Ensure that data exported or produced adhere to supported standards (e.g., XML).
- Develop and support tools for translation and mapping of data (e.g., wrapper generators).
- Encourage GTL researchers to participate in the development of standards and interface specifications as part of the larger community.
- Use or develop consistent, extensible syntactic and semantic schemas to enable data interchange and to describe data, metadata, and experiments.
- Ensure interoperability of databases and analysis packages:
  - User interfaces should support queries and set-oriented input rather than only point-and-click.
  - Web interfaces should retrieve set-oriented results.
- Develop databases to support versioning:
  - Versioning should be space efficient.
  - Users should be able to access and ask questions about past states.
  - Data should have unique, unchanging identifiers so minimal changes can be described.
  - Users should be able to see minimal edits between versions: push vs pull technologies for being notified of database updates (e.g., SwissProt's support of push notification).
- Support multimodal representations of data:
  - Metadata should be linked to visualization.
  - High-dimensional information should be visualized.

Data Storage and Access

- Fundamental research is required in these areas:
  - Warehousing of biological data
  - Process-oriented replay of updates and corrections (current understanding is limited to certain forms of relational algebraic expressions.)
- Support for data sharing, archiving, warehousing, replication management, and version control is necessary.
- Tools for managing and interacting with metadata, including provenance and derived data, should be developed.

Data Discovery and Analysis

- Discovery and analysis components should be adopted or developed to support GTL applications.
- Components should have "plug-and-play" nature, with common interfaces; both the toolbox and tools need to be created.
- GTL is distinguished by the scale and heterogeneity of data sources and will require multimodal solutions to support data discovery and analysis.

Gap Between Low-Level Tools and Domain-Specific Application Software

- DOE should fund a collaboratory pilot project in biology similar to that funded in SciDAC.
- Significant fundamental research and development activities are required to support the interconnection between application software and lower-level grid toolkits.
- Existing grid activities should be leveraged (including SciDAC Middleware and Global Grid Forum).

Time and Space Scalability

- Develop schemes for sharing data, changing models across different scales (vertical integration).
- Develop scalable computational models (from desktop to teraflop):
  - Scalable algorithms
  - Compatible applications across platforms
- Develop query and analysis methods across scales.

Workflow and Data Tracking

- Fundamental research is needed:
  - Identification and definition of data provenance
  - Paradigms for automated capture of provenance information
  - Techniques for reasoning about provenance through process (e.g., queries over data and inference and analysis packages that produce new data)
- Databases should be developed to support data provenance.
- Methods and tools are required to describe workflow tracing across independent experiments.
- Methods for tracking and archiving provenance should be developed:
  - Archiving of code
  - Archiving of data, pre- and postsimulation
  - Input parameters, experimental environment
- Support query and analysis of provenance information (e.g., for validation).

Sociology

- A reward structure should be required to support those who work on standards, interoperability, and multidisciplinary projects in a collaborative environment.
- Funding grants and renewals should be prioritized based on successful pairings of biology and CS.
- DOE should require teams of biologists and computer scientists to work together from the beginning through the grant-solicitation process.
- Innovative CS research is required to produce innovative biological research.

Strategies and Goals

The working group ended discussions by identifying specific criteria for success in facilitating integration and interoperability:

- Biological applications are constructed much more quickly, efficiently, and at lower cost if these issues are considered up front.
  - Usability.
  - Researcher access to a more efficient computational pipeline using CS tools. Researchers can spend more time doing science, less time doing reformatting, data movement, sequence of processing steps.

- Researchers will be evaluated on their impact on the biological community beyond DOE.
- Return on investment across GTL:
  - Software developed supports usability, quality, portability, and reusability and is used by multiple groups.
  - Return may take 5 to 10 years (not a typical 3-year grant cycle).

[1] www.cca-forum.org

[2] http://gizmo.lbl.gov/DM_TOOLS/OPM/opm.html

[3] D. Quass et al., "Querying Semistructured Heterogeneous Information" in *International Conference on Deductive and Object-Oriented Databases*, 1995.

[4] K. L. Schuchardt, J. D. Myers, and E. G. Stephan, "Open Data Management Solutions for Problem-Solving Environments: Application of Distributed Authoring and Versioning to the Extensible Computational Chemistry Environment," *Proceedings HPDC-10*.

[5] McCormick, 1998.

[6] www.globus.org/datagrid/

[7] www.llnl.gov/ia/ia.html

Appendix A: Workshop Attendees, March 2002

Workshop Organizers

Ray Bair	Pacific Northwest National Laboratory
Gary Johnson	U.S. Department of Energy
John C. Houghton	U.S. Department of Energy
Peter D. Karp	SRI International
Rick Stevens and Bill Gropp	Argonne National Laboratory

Invited Speakers

Jehoshua Bruck	California Institute of Technology
Susan B. Davidson	University of Pennsylvania
Lynette Hirschman	The MITRE Corporation
Fred Johnson	U.S. Department of Energy
Gary Johnson	U.S. Department of Energy
Peter D. Karp	SRI International
Aristides Patrinos	U.S. Department of Energy
Emanuel F. Petricoin III	Center for Biologics Evaluation & Research/FDA
Walter Polansky	U.S. Department of Energy
Mary Anne Scott	U.S. Department of Energy

Breakout Leaders

William R. Cannon	Pacific Northwest National Laboratory
Deborah K. Gracio	Pacific Northwest National Laboratory
Lynette Hirschman	The MITRE Corporation
Peter D. Karp	SRI International
Rick Stevens	Argonne National Laboratory

Other Participants

John Ambrosiano	Los Alamos National Laboratory
Paul E. Bayer	U.S. Department of Energy
David Benton	GlaxoSmithKline
Ann L. Chervenak	University of Southern California Information Sciences Institute
Mike Colvin	Lawrence Livermore National Laboratory
Terence Critchlow	Lawrence Livermore National Laboratory
C. Forbes Dewey, Jr.	Massachusetts Institute of Technology
Chris Ding	Lawrence Berkeley National Laboratory
Dan Drell	U.S. Department of Energy
Jean-Loup M. Faulon	Sandia National Laboratories
Marvin Frazier	U.S. Department of Energy
Jim Glimm	Brookhaven National Laboratory
Andrey A. Gorin	Oak Ridge National Laboratory
David J. Israel	SRI International

Christopher R. Johnson	SCI Institute/University of Utah
Patrick Lincoln	SRI International
Lei Liu	NCSA Keck Genome Center
Phil Locascio	Oak Ridge National Laboratory
Reinhold C. Mann	Pacific Northwest National Laboratory
William H. Miner, Jr.	U.S. Department of Energy
Esmond Ng	Lawrence Berkeley National Laboratory
Carl Edward Oliver	U.S. Department of Energy
Ian Paulsen	The Institute for Genomic Research
Isidore Rigoutsos	IBM Thomas J. Watson Research Center
Andrey Rzhetsky	Columbia University
Nagiza F. Samatova	Oak Ridge National Laboratory
Hinrich Schuetze	Novation Biosciences
Evgeni Selkov	Argonne National Laboratory
James P. Sluka	InPharmix Incorporated
Sylvia Spengler	National Science Foundation
George Seweryniak	U.S. Department of Energy
David Thomassen	U.S. Department of Energy
John van Rosendale	U.S. Department of Energy
John Wooley	University of California at San Diego

Meeting Logistics

Jane Hiegel	U.S. Department of Energy
Craig Stacey	Argonne National Laboratory
Cheryl Zidel	Argonne National Laboratory

Appendix B: Final Agenda, Computer Science Workshop for theGenomes to Life Program

March 6–7, 2002

Gaithersburg Hilton, Gaithersburg, Maryland

March 6

8:15–8:45 a.m.	Welcome Ari Patrinos, Associate Director, BER
8:45–9:15 a.m.	Genomes to Life Program Vision: Walt Polansky, MICS Director; Gary Johnson
9:15–11:00 a.m.	Provocative Vision Statements 9:15 a.m. – Peter Karp, SRI 9:45 a.m. – Emanuel Petricoin, FDA
10:15 a.m.	Break
10:30 a.m.	Lynette Hirschman, MITRE
11:00–11:15 a.m.	Charge to Breakout Groups
11:15 a.m.–Noon	Breakouts Begin (review and revise questions before lunch) Genome Annotation: Peter Karp, lead Protein Expression and Proteomics: William Cannon, lead Technical Text Mining for Biological Data: Lynette Hirschman, lead Simulation Tools for Cell Networks: Rick Stevens, lead Facilitating Interoperability: Deborah Gracio, lead
Noon–1:15 p.m.	Working Lunch A Vision for the DOE Computer Science Research Program Fred Johnson, DOE
1:15–3:15 p.m.	Breakouts Continue
3:15–3:30 p.m.	Break
3:30–4:45 p.m.	Breakout Status Reports (15 min. each) and Discussion Genome Annotation Protein Expression and Proteomics Technical Text Mining for Biological Data Simulation Tools for Cell Networks Facilitating Interoperability
4:45–5:30 p.m.	Open Discussion (cross-cutting issues)

March 7

8:15–8:30 a.m.	Daily logistics information
8:30–9:30 a.m.	Provocative Vision Statements 8:30 a.m. – Jehoshua Bruck, Caltech 9:00 a.m. – Susan Davidson, University of Pennsylvania
9:30–9:45 a.m.	Guidance to Breakout Groups
9:45–10:00 a.m.	Break
10:00 a.m.–Noon	Breakouts Continue
Noon–1:00 p.m.	Working Lunch A Vision for the DOE National Collaboratories Program Mary Anne Scott, DOE
1:00–3:30 p.m.	Summary Presentations (15 min.) and Discussion (15 min.) Genome Annotation Protein Expression and Proteomics Technical Text Mining for Biological Data Simulation Tools for Cell Networks Facilitating Interoperability
3:30–4:00 p.m.	Open Discussion (next steps)
4:00 p.m.	End of workshop
4:00–7:00 p.m.	Writing Team Drafts Report Sections