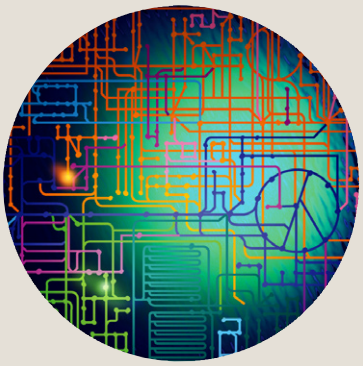
4.0. Creating an Integrated Computational Environment for Biology

To accelerate GTL research in the key mission areas of energy, environment, and climate, the Department of Energy Office of Science has revised its planned facilities from technology centers to vertically integrated centers focused on mission problems. The centers will have comprehensive suites of capabilities designed specifically for the mission areas described in this roadmap (pp. 101-196). The first centers will focus on bioenergy research, to overcome the biological barriers to the industrial production of biofuels from biomass and on other potential energy sources. For more information, see Missions Overview (pp. 22-40) and Appendix A. Energy Security (pp. 198-214) in this roadmap. A more detailed plan is in Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda, DOE/SC-0095, U.S. Department of Energy Office of Science and Office of Energy Efficiency and Renewable Energy (http://genomicsgtl.energy.gov/biofuels/).



Creating an Integrated Computational Environment for Biology

The concepts in this computing roadmap were developed over several years, and much of the material comes from reports of nine workshops held by DOE's Office of Advanced Scientific Computing Research and Office of Biological and Environmental Research since 2001. See Appendix D. GTL Meetings, Workshops, and Participating Institutions, p. 239.

Workshop reports are available at doegenomestolife.org/pubs.shtml.

Every facility description has a roadmap for computational needs. These roadmaps, starting in 5.0. Facilities Overview, p. 101, form a more complete picture of the integrated computational environment.

4.1. An Essential Foundation

Computation is essential to the GTL program goal of achieving a predictive understanding of microbial cell and community systems. Computing and information technologies allow us to surmount the barrier of complexity that separates genome sequence from biological function. The integrated GTL computational environment will link data of unprecedented scale, complexity, and dimensionality with theory, modeling, simulation, and experimentation to derive principles and develop and test biosystems theory. GTL computation will employ data-intensive bioinformatics, compute-intensive molecular modeling, and complexity-dominated cellular systems modeling.

Models and simulations represent an ultimate level of integrated understanding. A key goal for cell modeling is to predict cell phenotype from the cell's genotype and extracellular environmental information. Such predictions, resulting from comparative genomics studies, will include cell ultrastructure, morphology, motility, metabolism, life cycle, and behavior under a wide range of environmental conditions. These models not only will be descriptive and phenomenological but also will be predictive at multiple levels of detail. Although this 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, modeling and simulation, and experimentation will enable us to develop (albeit 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.

A comprehensive knowledgebase will be at the heart of GTL systems microbiology (see 3.2.2.3. Milestone 3: Develop the Knowledgebase, Computational Methods, and Capabilities to Advance Understanding and Prediction of Complex Biological Systems, p. 51). The knowledgebase foundation is the DNA sequence code that will relate the many data sets emanating from microbial systems biology research and discovery. Building over time to an intensely detailed and annotated description of microbial functions, the GTL Knowledgebase will assimilate a vast range of microbial data as it is produced. It will grow

to encompass program and facility data and information, metadata, experimental simulation results, and links to relevant external data and tools. Underlying the knowledgebase will be an array of databases, bioinformatics and analysis tools, modeling programs, and other transparent resources (see Fig. 1. GTL Integrated Computational Environment for Biology, this page).

Some examples of core capabilities required by the GTL program follow.

- **Bioinformatics: Collecting and Analyzing Data on Cellular Components.** The term "bioinformatics" includes a range of computational analyses characterized in part by reliance on data, especially genomics and proteomics data, as the critical investigative feature. Sequence analysis, largely the prediction of genes and gene function by homology, has been a core task. GTL will generate many such data types as measurements of protein complexes, protein expression, and microbial cell and community metabolic capabilities. Vast new data sets must be correlated or annotated to genome data and archived to provide foundational data for computer models of biochemical pathways, entire cells, and, ultimately, microbial ecosystems.
- **Molecular Measurements and Modeling: Revealing Processes Carried Out by Cellular Components.** GTL seeks to understand fully the cell's biological machinery and its relationships with other cells and the environment. To reach this goal, investigators must know and be able to computationally model and test concepts in which cellular components interact directly with each other and with other molecules in a cell. They also must know how proteins dock structurally to form a complex and how the proteins of a complex interact dynamically to accomplish a biological function. For example, detailed characterization

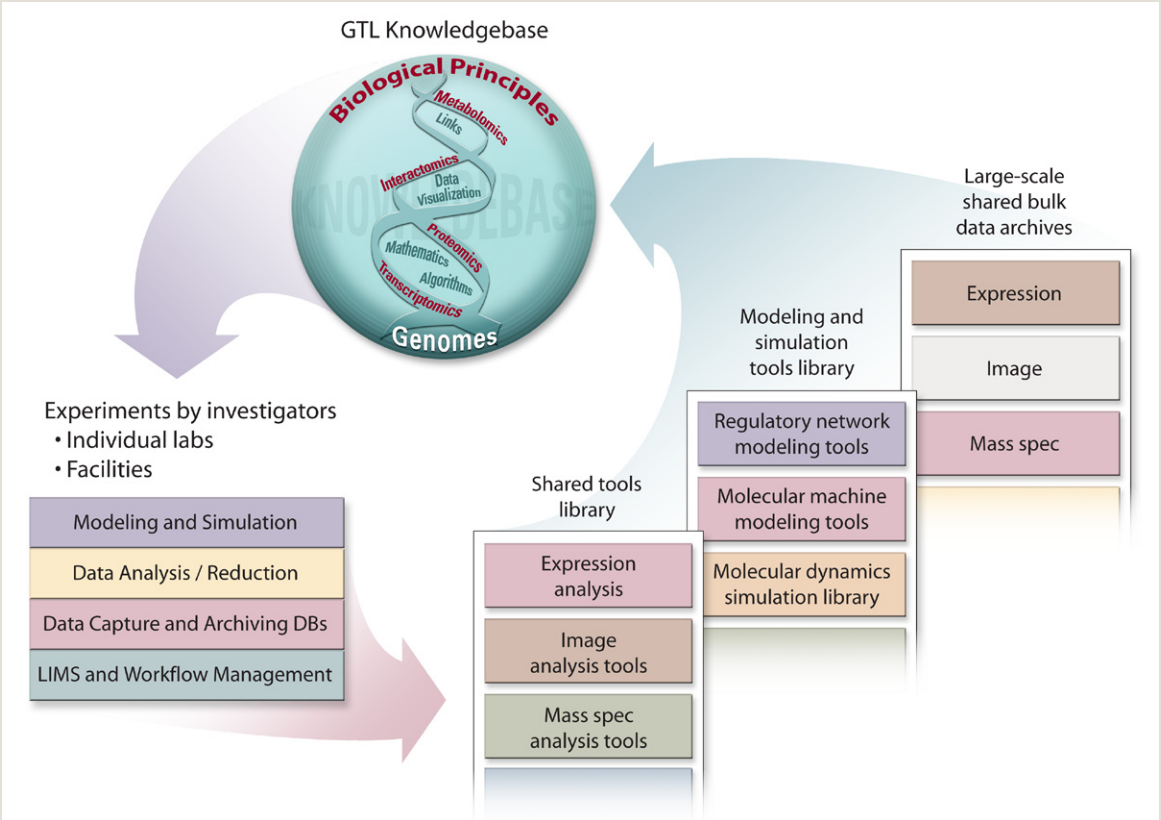


Fig. 1. GTL Integrated Computational Environment for Biology: Using and Experimentally Annotating GTL's Dynamic Knowledgebase. At the heart of this infrastructure is a dynamic, comprehensive knowledgebase with DNA sequence code as its foundation. Offering scientists access to an array of resources, it will assimilate a vast range of microbial data and knowledge as it is produced.

of protein complexes is the prerequisite for understanding the functions of molecules, cells, regulatory complexes, and networks as well as the interactions of cell surface proteins and complexes with the environment.

- **Cell and Community Modeling: Coalescing the Cell's Components into a Whole-Systems Predictive Understanding**. Biosystem models encapsulate our understanding of biology, and simulation is becoming a key tool for furthering understanding at the systems level. Through computational analysis of predictive mathematical models, we will understand how microbial organisms and communities may be manipulated to solve problems, how microbes regulate the expression of genes involved in environmental interactions, and how protein complexes are assembled to carry out important processes. Predictive models also will prove most useful in integrating and summarizing the vast amounts of data to be generated by the GTL program.

The computational biology environment will provide the networking "nervous system" to connect experimental and computational facilities with the large, geographically dispersed community of biology researchers, advancing collaboration and education. The environment will make tractable the project's inherent science diversity and its expected scale and duration. Computing will be tailored to meet the needs of biological research with transparent available tools linked to high-quality and interoperable databases.

Two offices in DOE's Office of Science—BER with its experience in biology and genomics and OASCR with its leadership and experience in computing (see sidebar, Scientific Discovery Through Advanced Computing, this page)—have teamed to achieve the goals of systems biology, the next generation of life sciences research. DOE's experience and capabilities in harnessing computing for science goals already have led to such breakthroughs in biology as annotation and sequence-assembly tools. This trend will be continued as described in this chapter. GTL also will leverage biocomputing developments in other agencies and institutions to contribute to the creation of sophisticated concepts and tools for advancing systems biology worldwide.

This chapter describes the attributes and uses of the community-accessible GTL computational environment, presenting the strategy and roadmaps for establishing essential capabilities that tie together GTL scientists and research facilities. As described in the supporting roadmaps, establishing these capabilities will be part of a rigorous development process involving the scientific community, other federal agencies, and industry.

Scientific Discovery Through Advanced Computing

SciDAC, launched in 2001, is a DOE Office of Science (SC) program to develop the infrastructure needed to take full advantage of DOE's commitment to the next generation of scientific computing. The next generation includes terascale machines capable of performing at 1000 times the speed of those available to the U.S. scientific community today, connected by high-speed networks with the most advanced middleware.

The SciDAC program is designed to bridge the gap between advanced applied mathematics and computer- and computational-science research in the physical, chemical, biological, and environmental sciences. This same kind of integrative and cross-disciplinary research is envisioned for GTL. The SciDAC model has proven especially effective in driving advances in large-scale simulation by tackling problems that are too large, too expensive, hazardous, or otherwise impossible to be solved through traditional theoretical and experimental approaches. Results have provided levels of detail and accuracy never before possible.

Two of the largest and most-successful SciDAC projects are

- **Terascale Supernova Initiative: A multidisciplinary collaboration of one national laboratory and eight universities to develop models for core collapse supernovae and enabling technologies (www.tsi-scidac.org).**
- **Accelerated Climate Prediction Initiative: A multi-institutional collaboration to develop, validate, document, and optimize the performance of the Community Climate System Model (www.ucar.edu/communications/CCSM/overview.html).**

Credits: National Energy Research Scientific Computing Center: 2003 Annual Report (www.nersc.gov/news/annual_reports/annrep03/annrep03.pdf); SciDAC web site (www.osti.gov/scidac/)

4.2. Capabilities for an Integrated Computational Environment

To support the achievement of its science and mission goals, GTL must establish a number of essential elements in building the program's computational environment. These components include a seamless set of foundational capabilities to support the pinnacle capability of theory, modeling, and simulation. They include a rigorous and transparent system for tracking, capturing, and analyzing data with a computing and information infrastructure accessible to the scientific community.

1. **Theory, Modeling, and Simulation Coupled to Experimentation of Complex Biological Systems:** Build concepts and models of microbial cells and communities that capture and extend our knowledge, based on a combination of experimental data types. Test and validate component models and use integrated models to understand mechanisms and explore new hypotheses or conditions to design new experimental campaigns.
2. **Sample and Experimental Tracking and Documentation – Laboratory Information Systems (LIMS) and Workflow Management:** Provide systems for experiment design, sample specification, sample tracking and metadata recording, workflow management, process optimization and documentation, QA, and sharing of such data across facilities or projects.
3. **Data Capture and Archiving:** Capture bulk data from many different measurements and instruments in large-scale data archives.
4. **Data Analysis and Reduction:** Provide analysis capabilities for systems biology data to enable insights, input, and parameters for systems models and simulations.
5. **Computing and Information Infrastructure:** Furnish hardware and software environments to support analysis, data storage, and modeling and simulation at the scales required in GTL.
6. **Community Access to Data and Resources:** Provide community access to data, models, simulations, and protocols for GTL. Allow users to query and visualize data, use models, run simulations, update and annotate community data, and combine community data and models with their local databases and models.

These capabilities are described more fully in the following text.

4.2.1. Theory, Modeling, and Simulation Coupled to Experimentation of Complex Biological Systems

Theory and Modeling Objective: Build concepts and models of microbial cells and communities that capture and extend our knowledge, based on a combination of experimental data types.

Simulation Objective: Test and validate concepts and use integrated models to understand mechanisms, explore new hypotheses or conditions, and drive new experimentation.

The only conceivable methodology for success in achieving GTL goals is a coherent and tight integration of theory, modeling, and simulation (TMS) with experimentation (E) and resultant data. Theory refers to the hypothetical concept that underlies properties and phenomenological behavior. Modeling is the translation of that theoretical concept into mathematical terms so calculations can be carried out. Simulation combines multiple models into a meaningful representation of the whole system, encompassing physicochemical and other variables that together evolve computationally to identify "emergent" behaviors.

Computationally driven TMSE provides an interface between the researcher and huge resultant data sets from complex systems, involving (1) at a mechanistic level, multiple strongly interacting processes and elements; (2) at a functional level, multiple strongly coupled phenomena; and (3) behaviors that are unforeseen and not intuitively accessible. Rapid and inexpensive in silico experiments via simulations can be used to

gain first insights, form hypotheses, and conceive and carry out meaningful tests. Utilizing simulations for understanding critical parameters, investigators can technically and statistically design physical experiments for maximum efficacy. Resultant data from all experiments will be compared against simulations in various ways to test assumptions and hypotheses, identify new phenomena, and spark new theories. Computational simulations are a time machine, microscope, and telescope, allowing complex systems to be analyzed from any conceivable organizational, temporal, spatial, process, or phenomenological perspective.

To address DOE's mission interests, we will need to go beyond understanding how cells work in known environments. We must predict how organisms will respond to new sets of conditions, how selected collections of components might be put to work in vitro (another set of conditions), and how we can tune the biochemical processes to do different things. In other words, a chief goal in making models and simulations will be to apply them to circumstances different from the situations for which we have data.

When dealing with complex systems, TMS and high-performance computing have emerged as the universal methodology to drive experimentation, which can be prohibitively expensive, difficult, and time consuming. The use of TMS has become the foundational capability for every aspect of science and engineering, from chemical engineering to aerospace designs, and has fostered a dramatic change in research and technology-development cycles. The GTL Knowledgebase will serve as a discovery-driven resource for developing new modeling capabilities, conducting experiments to verify models quantitatively, and simulating how interacting phenomena affect each other. Insights gained will be readily accessible in the GTL Knowledgebase for new hypotheses and statistically designed experiments, bringing to bear cumulative, cross-referenced data on building new models and simulations.

Ultimately, scientists will be able to create in silico models of a microbe by comparing its genomic sequence with highly annotated gene and functional information in the knowledgebase. The goal is to create increasingly accurate mathematical models of life processes to enable prediction of cell and community behavior and develop new or modified systems tailored for mission applications (see box, Examples of Biological Understanding and Possible Applications Enabled by TMS, this page).

Examples of Biological Understanding and Possible Applications Enabled by TMS

Advancing Understanding of Biological Subsystems via Modeling and Simulation

- Regulatory networks: Observations of protein expression and other biomolecules correlated with environmental cues
- Protein interaction networks: Interaction data of several types
- Organization and function of protein and other multimolecular complexes: Homology, interaction, structure, and image data
- Cells: Regulation, metabolism, and biomolecular interactions

Simulation Examples

- Molecular dynamics to visualize the workings of a molecular machine
- Expression of a protein set in a condition that can be tested experimentally to validate a regulatory network
- Combined metabolic pathway, regulatory network, and protein interaction network to explore cell response to environmental changes

Possible Application and Engineering Scenarios Enabled by Advanced Understanding of Microbial Systems

- Elucidate intercellular communication pathways in bacterial communities to understand microbial contributions to ecosystem function, including carbon and nutrient cycling in terrestrial ecosystems
- Understand the roles of cyanobacteria, diatoms, and other microbes in carbon cycling and sequestration

4.2.1.1. Microbial Behavior: Modeling at the Molecular Level

The starting point for GTL analysis is to decipher microbial processes at the molecular level. The centerpiece of GTL is the ability to analyze, reconstruct, and model the networks of molecular interactions at the core of 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, cellular molecular networks and pathways are at the center of cell modeling and cellular behavior and, ultimately, of microbial-community modeling and behavior. Such models would predict how a cell's genome and environmental factors combine to yield its phenotype. Models will be powerful tools for scientific discovery as we explore the enormous complexity of microbes and their communities.

4.2.1.2. Computer Science and Mathematics Challenges

Achieving predictive capabilities will require overcoming many technical challenges. For example, cell modeling eventually may involve a more complex collection of components and materials than do existing models of climate or mechanical systems. Many needed developments involve research in computer science and mathematics. New mathematical methods are needed for analysis of raw biological data to include in models and the subsequent statistical design of experiments to validate those models. Additionally, major research challenges relate to database query and design in support of modeling, as well as the development of effective databases to capture modeling output and the models themselves.

Modeling complex biological systems will require new methods to treat the vastly disparate length and time scales of individual molecules, molecular complexes, metabolic and signaling pathways, functional subsystems, individual cells, and, ultimately, interacting organisms and ecosystems. Such systems act on time scales ranging from microseconds to thousands of years. These enormously complex and heterogeneous full-scale simulations will require not only petaflop capabilities but also a computational infrastructure that permits model integration. Simultaneously, it must couple to huge databases created by an ever-increasing number of high-throughput experiments. Challenges include determining the right calculus to describe regulation, metabolism, protein interaction networks, and signaling in a way that allows quantitative prediction. Possible solutions include use of differential equations, stochastic or deterministic methods, control theory or ad hoc mathematical network solutions, binary or discrete value networks, Chaos theory, and emerging and future new abstractions.

4.2.1.3. Fundamental Questions and Issues

As this systems-level approach to understanding microbial cells and their communities develops, several questions must be addressed:

- What are the biological design variables?
- Can biological systems be modeled to the same degree as physical and chemical systems?
- How do physical and chemical principles and approximations developed for modeling nonliving systems apply to the simulation of living systems?
- Are numerical values for parameters such as enzyme-catalyzed reaction rates known, or even knowable, since such properties change with time and environmental conditions and from cell to cell?
- How can we quantify the levels of uncertainty in our understanding and predictions and the sensitivity of our models to variations in input parameters and structure?
- How do we address important issues of model and knowledge representation and formalism?

4.2.1.4. Chemistry Challenges

Chemistry is essential to our understanding and exploitation of cellular processes. The functions of a cell are increasingly being understood through explanation of the underlying chemistry. Structural imaging

technologies enable construction of models of protein machines as they carry out many cell functions, including processes that relate directly to DOE missions that are the focus of GTL. As we learn more about cells, we will want to broaden the range of operating conditions to which these protein machines can be exposed and the range of environmental substrates that they can convert to other substances. We also will modify proteins to make them active with nonnative substrates of interest to DOE, resulting in levels of specificity different from the native system. For example, a machine that enables a microbe to convert an environmental contaminant such as carbon tetrachloride to benign products might be modified to enable the microbe to destroy the related contaminant trichloroethylene. Understanding the detailed chemical mechanisms taking place as the protein machine processes a substrate will be critical to planning its use intelligently and engineering it to meet our needs.

4.2.1.5. Structure, Interactions, and Function

Reliable high-throughput determination of protein and protein-complex structures and functions will require computational methods capable of integrating several sources of experimental data; examples include mass spectrometry (MS), X-ray crystallography and scattering, protein arrays, numerous imaging modalities, cross-linking, yeast two-hybrid, and nuclear magnetic resonance (NMR). High-throughput MS experiments involving complexes and cross-linkers pose significant informatics and computational challenges.

These data sets will enable molecular-level simulations and prediction, thus populating the GTL Knowledgebase with functional annotations at three levels: (1) Computationally driven high-throughput protein-structure prediction, (2) integrated experimental and computational approaches to structures and function for hard-to-isolate proteins and complexes, and (3) advanced molecular simulations of biochemical activity.

An important driver for high-performance computing systems will be modeling and simulation to predict the behavior of complexes of specific sets of proteins chosen from network analyses and other experiments. Computational requirements for such simulations are the best characterized among all areas of computational biology; moreover, many of these simulation methods already are implemented on teraflop-scale computers. Pure computing power is the major limitation on size and accuracy of many biochemical simulations, which will involve data and models of protein-protein interactions, ligand-protein interactions, electron-transfer interactions, and membrane characteristics. Molecular dynamics and quantum mechanics–based molecular modeling will spur high-end computing and require development of more effective scalable algorithms. The GTL program will push the envelope for biophysical modeling, in particular, to develop the ability to predict the actual behavior of proteins and protein complexes for a set of biological processes chosen for their importance to GTL goals.

4.2.1.6. Microbial Behavior: Metabolic Network and Kinetic Models of Biochemical Pathways

4.2.1.6.1. Current State of Cell-Network Modeling: Moving from Experiment (Real Life) to Simulation (Abstract Systems Model)

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., from annotated genomes, proteomics databases, in vitro experiments, expression, and the historical literature). Models are used to facilitate a general understanding of cellular networks and for simulations that attempt to reproduce or predict a particular experimental result. When attempting to develop a systems understanding of complex biology, investigators will use simulation and modeling as one of a few ways to derive insight from complicated interactions involving numbers of variables and details that cannot be grasped intuitively.

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. Although

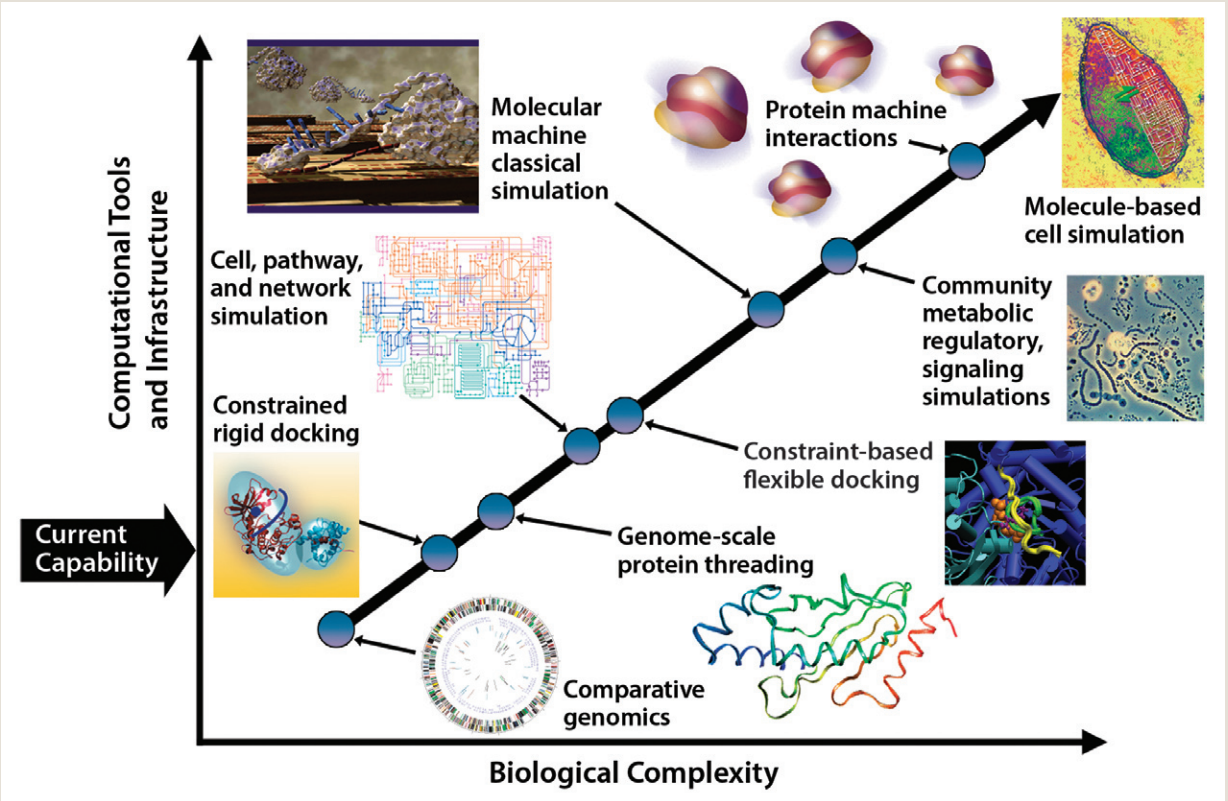
more-general qualitative predictions can be made for larger, more complete networks, the current lack of kinetic constants for most enzymes and of concentration data for intermediate metabolites limits the ability to simulate quantitative results for entire networks including cells and communities. Figure 2 illustrates current capabilities on the path from genome data to full cell simulation. Modeling also is 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.

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, ChemCell, and BioSpice). Several different formalisms (e.g., rule based, ordinary differential equation, logical, and qualitative) represent and simulate cell-network models. Current cell-network simulations typically are running on serial computers (PCs and workstations) and are used mostly to simulate processes in individual cells or simple cellular interactions.

4.2.1.6.2. Advanced Modeling Capabilities

No dominant formalism, however, has emerged that can satisfactorily represent both the 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 multiple levels of spatial and

Fig. 2. From Genome Data to Full Cell Simulation. This concept diagram schematically illustrates a path from basic genome data to a more detailed understanding of complex molecular and cellular systems and of the need to develop new computational analysis, modeling, and simulation capabilities to meet this goal. The points on the plot are very approximate, depending on the specifics of problem abstraction and computational representation. Research is under way to create mathematics, algorithms, and computer architectures for understanding each level of biological complexity.



temporal scales in cellular systems and incorporate them into models. Most models of cellular networks are one dimensional (1D) (e.g., box models that assume a completely mixed environment). To make progress toward 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 different classes of problems, each with distinct temporal and spatial scales and each with potentially different computational complexity. All classes of problems have 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.

Compartmentalized models will become increasingly important for depicting distinct types and phases of metabolism in organisms such as cyanobacteria, which have both oxygenic and anoxic pathways separated either spatially or temporally. Compartmentalized models will be needed to fully describe life cycles of prokaryotes, which include mechanisms such as sporulation, heterocyst formation, and differentiation. 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 DOE National Nuclear Security Agency's Accelerated Strategic Computing and the climate modeling community. Compartmentalization and coupling also will 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 hierarchies or functional subsystem modules are not necessarily obvious, yet exploiting modularity or lumping the system may be essential for efficient modeling and simulation.

4.2.1.6.3. Crosscutting Research and Development Needs

A major challenge is the need to integrate heterogeneous data types into cell models for molecular interactions, metabolism, and regulation. Types include data generated by different imaging modalities, structure determinations, MS, coexpression analyses, and an array of binding and other constraints.

Mathematical models ultimately must be developed from fundamental biological principles. Mathematics and computer science research will aid in understanding the following:

- Organization of principles or theories that could lead to successful models, even with incomplete knowledge, missing data, and errors.
- Determination of strengths and weaknesses of different types of simulation methods for different systems biology problems (e.g., stochastic, differential equation, mathematical networks).
- Use of high-performance computing to provide the compute power to run long time-scale simulations (e.g., in milliseconds and longer time frames for ab initio or directed and constrained molecular dynamics for simulating machines).

Computationally, no single architecture is appropriate for all aspects of predictive cell modeling. Because computational requirements are so diverse, coupling informatics with modeling and simulation establishes the need for a fully general-purpose computing infrastructure. Hardware needs for such a challenge range from commodity clusters to tightly coupled, massively parallel architectures with greater investment in interprocessor communication. Implications for operating systems are equally disparate, requiring in some cases extremely high rates of parallel input-output to move data among processors and memory, as well as efficient management of single-application codes distributed over hundreds or thousands of processors (see Theory, Modeling, and Simulation Roadmap, p. 91).

4.2.2. Sample and Experimental Tracking and Documentation: Laboratory Information Management System (LIMS) and Workflow Management

Objective: Provide systems for experiment design, sample specification, sample tracking and metadata recording, workflow management, process optimization and documentation, QA, and sharing of such data across facilities or projects.

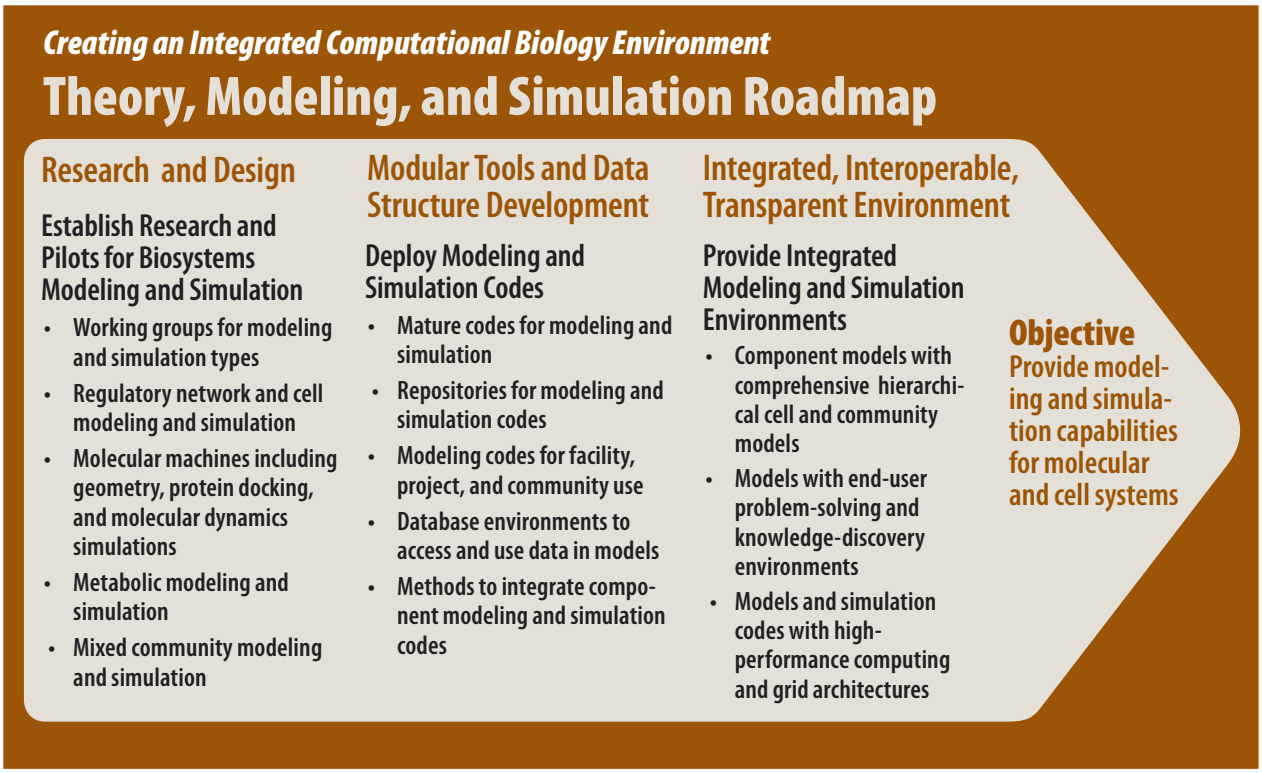
4.2.2.1. LIMS Impact

The goal of creating—from genome sequence—a knowledgebase for efficiently understanding the functions of microbes and communities requires many iterations of modeling, experimentation, and simulation. LIMS ensures the rigor of experimental data by linking it with associated QA/QC factors, characterizations, protocols, and related experiments and data.

LIMS maintains a detailed pedigree for each sample by capturing processing parameters, protocols, stocks, tests, and analytical results for the sample's complete life cycle. Project and study data also are maintained to define each sample in the context of research tasks it supports. LIMS will be required for each analytical pipeline to track all aspects of sample handling.

4.2.2.2. LIMS Requirements for GTL

Scientists funded by the GTL program and users of GTL facilities will conduct many thousands of experiments, each with hundreds to thousands of individual samples upon which several analytical measurements will be made. Although a number of LIMS are sold by commercial vendors, no single LIMS will be able to meet the large-scale, varied needs of all GTL facilities and projects. The broad range of experimental protocols used in the facilities and in the laboratories of GTL investigators will require LIMS customizations flexible enough to meet constantly changing requirements (e.g., new experimentation, protocols, parameters, and data formats).



Throughput is vital to the GTL facilities, so care must be exercised in the design of systems critical to the facility's uptime. The core LIMS at each facility is just such a system. When it is not operating, data cannot be processed and the facility cannot run. LIMS must be very robust, highly available, and secured in ways similar to an institution's critical information technology systems. An external data query or database operation must not impact LIMS or operations. Databases assimilating a facility's data must be inaccessible to hackers, and the system and databases for recording data should be separate from those for sharing data.

A working group will be established to examine existing and future needs of GTL grantees and the four facilities. The group will assess and analyze the existing LIMS as a prelude to adopting or creating a flexible and interoperable LIMS across a number of laboratory and facility environments (see LIMS and Workflow Management Roadmap, this page).

4.2.3. Data Capture and Archiving

Objective: Capture bulk data from many different measurements and instruments in large-scale data archives.

Perhaps the greatest challenge to GTL is the explosion of biological data. Massive and very complex, the body of data comes in different types and formats determined by experiments or simulations. It spans many levels of scale and dimensionality, including genome sequences, protein structures, protein-protein interactions, metabolic and regulatory networks, multimodal molecular and cellular imagery, and community properties.

The challenge is less about storage and retrieval, however, and more about fundamental support for new ways of doing science. Research groups must interact with these data sources in new ways. The GTL infrastructure will provide users with cutting-edge data-management and -mining software tuned to biology's needs. This capability is beyond the reach of any single research institution. This is a key area for GTL interaction with other agencies that would have great impact on the biology community as a whole.

Multiterabyte biological data sets and multipetabyte data archives will be generated by high-throughput technologies and petascale computing systems. Among the issues are types of GTL-generated data; mechanisms for data capture, filtering, and storage; ways of disseminating data (publicly accessible, central vs dispersed repositories, federations); and integration with existing databases. Given the hierarchical nature of

Creating an Integrated Computational Biology Environment

LIMS and Workflow Management Roadmap

Research and Design

Establish LIMS/Workflow Research and Pilots

- LIMS/facility workflow working group
- LIMS pilots at GTL facilities
- Research pilots on workflow systems
- GTL experimental-design working group
- Research pilots in QA/QC
- Research and pilots for facility process design
- Shared LIMS and workflow technologies

Modular Tools and Data Structure Development

Deploy LIMS and Workflow Systems

- Mature LIMS and workflow systems
- Production dataflow at each facility
- LIMS linked to bulk dataflow archives
- Intermediate process-management environment
- Plan for LIMS and workflow integration

Integrated, Interoperable, Transparent Environment

Integrate LIMS and Workflow

- Dataflow integrated across facilities and projects
- Workflow process integrated across facilities and projects

Objective

Optimize sample tracking, experimentation, workflow management, process documentation, QA, and data sharing

biological data, GTL databases should be organized according to natural hierarchies. Types of data supported by databases should go beyond sequences and strings to include trees and clusters, networks and pathways, time series and sets, 3D models of molecules or other objects, shapes-generator functions, and deep images. Tools are needed for storing, indexing, querying, retrieving, comparing, and transforming those new data types. For example, such database frameworks should be able to index and compare metabolic pathways to retrieve all that are similar. Also, current bioinformatics databases should support descriptions of simulations and large complex hierarchical models.

Data standards, developed in conjunction with other national biological research programs and standards organizations, are required for experimental observations of both biological phenomena and representative counterparts within the data model. Standards must be supported by statistical methods to design meaningful experiments and analyze resultant data. A framework of controlled vocabularies, common ontological definitions of basic GTL objects, and low-level data-interchange and -access methods should be developed to permit effective communication. Standardized semantics is a key technical challenge in accomplishing the goal of data standards. Due to the complexity of biological data, its rapidly evolving nature, and problems with synonymy (different names with the same meaning) and polysemy (the same name for different concepts), GTL will use temporary standards and continue their refinement. Data types will be determined by new experiments, analyses, and simulations, so data-storage strategies will evolve over time. Through cooperative development of data models and database schemas, the GTL data-integration enterprise will lay the groundwork for a distributed but integrated suite of research-project and facilities databases. These databases will permit the unique knowledge acquired by each research group to be used by the larger research community, thus allowing users to mine data from the combined sites.

Key features of databases and structures include:

- Probabilities and confidence factors, visualization tools, "query-by-example" capabilities, model parameters and elements for simulation environments, and new data models natural to life science;
- Interfaces to such experimental systems as chips, detectors, microscopes, and mass spectrometers; workflow support and experimental planning; and metadata processing;
- Search infrastructure that enables search services to operate across domains and metadata schemas.

Bioinformatics applications often are trivially parallel. Thus, hardware and operating-system requirements for bioinformatics are less about flop rates and interprocessor communication speeds and more about parallel input and output between processors and memory. For some applications, compute-cycle needs can be predicted; for others, however, the problems call for advancements in methods, so algorithmic and high-performance computing requirements are not yet clear. Successful bioinformatics tools should enable life-science researchers to seamlessly link data (often geographically distributed via the internet) with modeling and simulation results (see Data Capture and Archiving Roadmap, p. 94).

4.2.4. Data Analysis and Reduction

Objective: Provide analysis capabilities for systems biology data to provide insights, input, and parameters to systems models and simulations.

Bioinformatics encompasses a range of computational analyses characterized in part by reliance on data, especially genomics and proteomics data, as the central feature. Sequence analysis, largely the prediction of genes and gene function by homology, has been a core task.

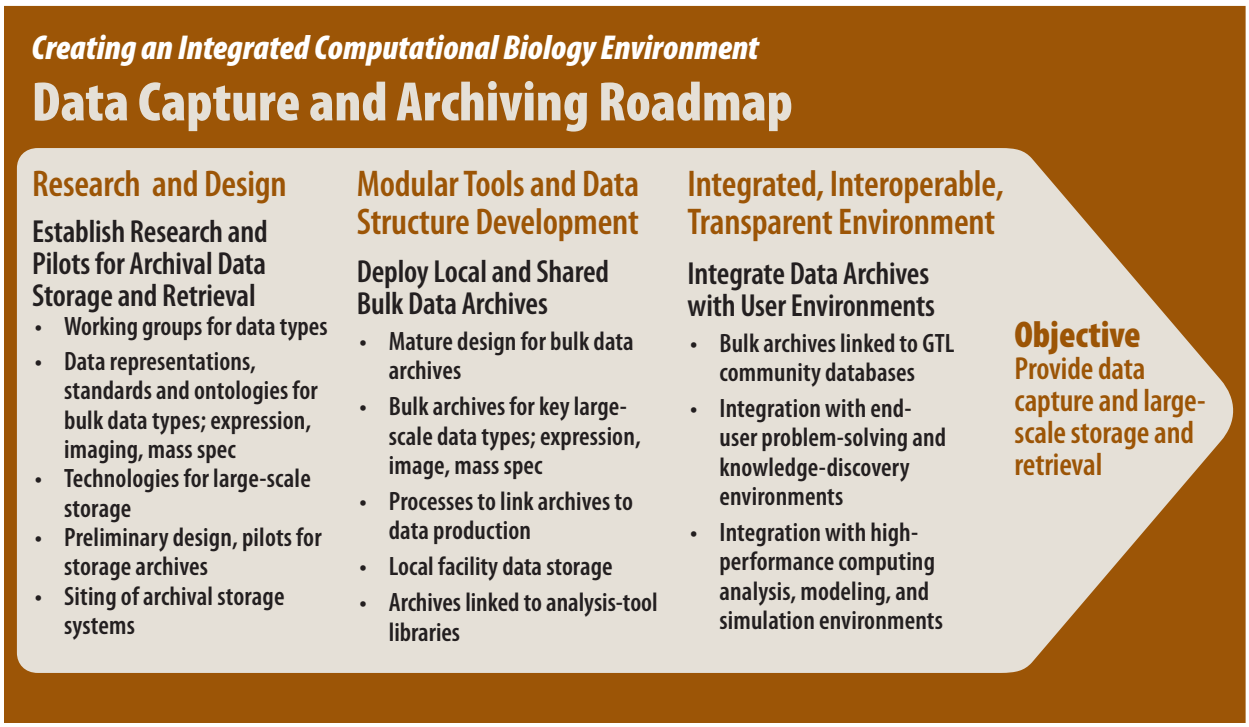
But in GTL, bioinformatics describes a broader set of investigations that will consider a wide variety of data types and sources—genome sequences, proteomics, metabolomics, expression, pathways, and simulation data. Many challenges are emerging as the amount and complexity of data are increasing exponentially and the types of analyses across multivariate data sets also become more complex. Many of these analyses can no longer be supported by local computing capabilities (see Data Analysis and Reduction Roadmap, p. 95).

4.2.4.1. Infrastructure

Data-analysis infrastructure will support an environment for creating and managing sophisticated, distributed data-mining processes. The unprecedented amount and complexity of biological data require that computational analysis is a key component of GTL (and systems biology in general). By developing the necessary tools and tool frameworks, GTL will allow biologists to derive inferences from massive amounts of heterogeneous and distributed biological data. Using intuitive visual interfaces, developers and data analysts will be able to program new data-mining applications or open existing application templates that easily can be customized to a given problem's unique requirements. Such processes will have both application and web-based streamlined interfaces. An infrastructure should encompass a large repository of analysis modules including sequence analysis, gene expression, phylogenetic tree, and mass spectrometry.

An objective of GTL is to provide high-throughput experimental data that can be used for rapid functional annotation of genomes. Understanding functions of microbes and microbial communities depends critically on the ability to develop and validate models and drive simulations based on experimental data. Massive data sets must be incorporated into systems simulations and models to infer function of genes and proteins. Such analyses will require advances in mathematical methods and algorithms capable of incorporating experimental data produced by a variety of techniques, including NMR, MS, X ray, neutron scattering, various microscopies, biofunctional assays, and many more. GTL will develop the methodology necessary for seamless integration of distributed computational and data resources, linking both experiment and simulation and taking steps to ensure that high-quality, complete data sets are linked to the validation of models of metabolic pathways, regulatory networks, and whole-cell functions.

Sequence annotation and comparative analyses across multiple genomes are recurring computational tasks that require a high-performance computing infrastructure to ensure that regular information updates are part of the most current annotation and to facilitate interactive exploratory genome analyses. Finding regulatory elements, an unsolved research problem in even the simplest genomes, is expected to involve significant computational and mathematical challenges. Some analysis of regulatory regions can be accomplished by large-scale genome comparisons. There remain significant research challenges in high-level annotation, including assignment of functions to every gene found in whole-genome sequences. This is particularly difficult because

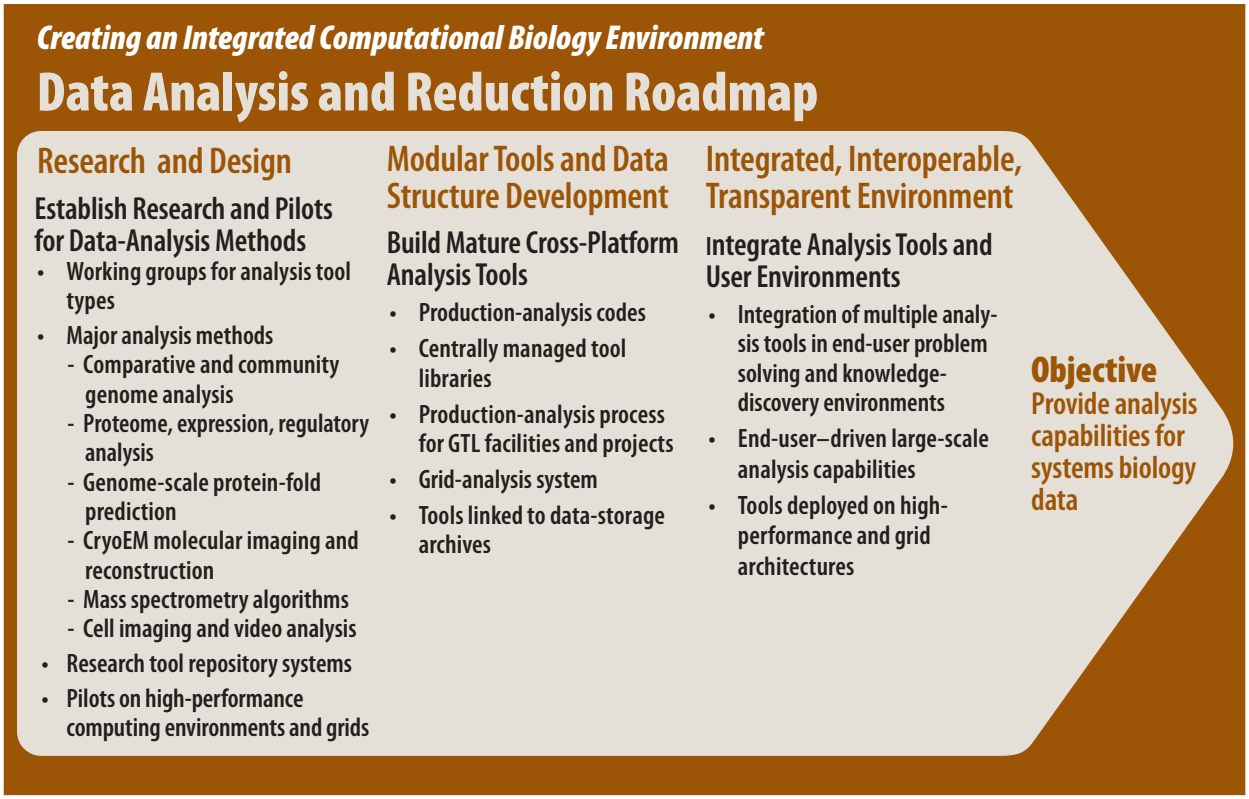


pathway databases are incomplete and microbial genomes encode for metabolic pathways about which very little biochemical data exist. At this time, 40 to 60% of genes found in new genomic sequences do not have assigned functions. Some functions can be inferred by computational structure determination and protein folding, but a wide range of research problems remains to be solved in this area. Computational methods will have a major role in the functional annotations of genomes, a necessary first step in developing higher-level models of cellular behavior. GTL will continue development of automated methods for the structural and functional annotations of whole genomes, including research into new approaches such as evolutionary methods to analyze structure and function relationships.

4.2.4.2. Examples of Analyses and Their R&D Challenges for GTL Science

GTL encompasses many types of data, each with algorithm research and development challenges in analyzing data for a broad range of purposes. Examples of objectives:

- Improve automated genome sequence annotation for microbes and microbial communities
  - New algorithms with improved comparative approaches to annotate organism and community sequences, identifying, for example, promoter and ribosome-binding sites, repressor and activator sites, and operon and regulon sequences
  - Protein-function inference from sequence homology, fold type, protein interactions, and expression
  - Automated linkage of gene, protein, and function catalog to phylogenetic, regulatory, structural, and metabolic relationships
- Identify peptides, proteins, and their post-translational modifications of target proteins in MS data
  - New MS identification algorithms for tandem MS
- Quantitate changes in cluster expression data from arrays or MS
  - New expression data-analysis algorithms



- Automatically identify interacting protein events in fluorescence resonance energy transfer (FRET) confocal microscopy
  - New automated processing of images and video to interpret protein localization in the cell and to achieve high-throughput analysis
- Reconstruct protein machines from 3D cryoelectron microscopy
  - New automated multi-image convolution and reconstruction algorithms
- Compare metabolite levels under different cell conditions
  - Algorithms for metabolite method analysis, both global and with spatial resolution
- Improve general R&D
  - Software engineering principles and practices developed and adopted for GTL software; modular, open source
  - Development of versions of analysis tools suitable for massively parallel processing and large-cluster computing environments

4.2.5. Computing and Information Infrastructure

Objective: Provide hardware and software environments to support analysis, data storage, modeling, and simulation activities required in GTL.

Computational biology has an unprecedented range of computing needs that make a well-planned infrastructure essential to achieving GTL's ambitious goals. The GTL program will require a distributed computing infrastructure that includes the ability to perform informatics analysis on a diverse collection of distributed data sets produced by a variety of experimental methods, run simulations on dedicated supercomputers, and study biological phenomena that no one yet knows how to model. The infrastructure for biology applications must provide high-speed computation for large-scale calculations but also must be compatible with much smaller scale calculations carried out on individual investigators' desktops. This infrastructure must be flexible, adaptable, and responsive to biology's evolving needs. It will consist of special and general-purpose computers and tool libraries linked together and to GTL facilities, research laboratories, and the user community by a national state-of-the-art backbone. The components include:

- GTL experimental facilities and research laboratories that generate large-scale biological data, analyze and manage the data, and make the information available to the community of GTL researchers.
- Data-curation centers where data are collected under strict quality and structure protocols to support modeling and other activities.
- Special and general-purpose computers that focus on such compute-intensive applications as analyzing biological data; modeling protein and molecular-machine structures; and simulating pathways, networks, cells, and communities.
- Tool libraries and modeling repositories that collect, implement, and develop analysis, modeling, and simulation tools related to GTL tasks, making them available to biology users at GTL centers and in the community.
- A national grid (associated with ESNet) with terabit backbone and associated middleware, connecting all the centers and users to provide the scientific community with a major new capability for high-impact biological science.

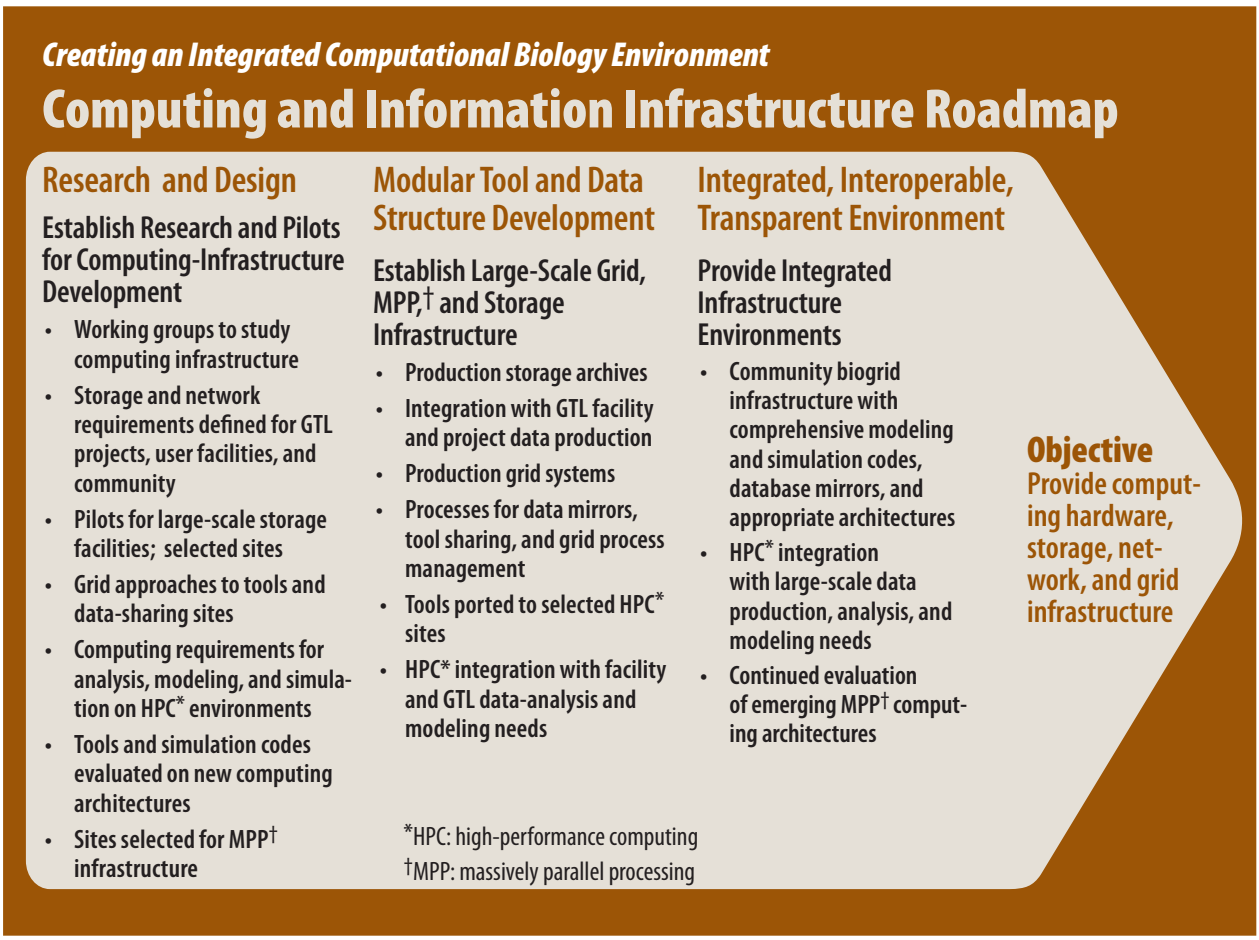
Requirements for this infrastructure must grow to match estimates of data production and data analysis needed in GTL research. It will build on existing computing centers and networking resources and leverage the major DOE user facilities. In general, for at least the first decade of GTL, computing and information technologies will be available in the commercial marketplace to meet the needs of biological research without development of special architectures or technologies (see Computing and Information Infrastructure Roadmap, p. 97).

4.2.6. Community Access to Data and Resources

Objective: Provide community access to data, models, simulations, and protocols for GTL. Allow users to query and visualize data, use models, run simulations, update and annotate community data, and combine community data and models with their local databases and models.

Making data and software from one research project accessible and useful to others is a considerable challenge, especially considering the many kinds of information produced by GTL, the variety of computational packages, and the rapidly evolving representations of our understanding of living systems. Not only does the community span a wide variety of interests and expertise, it also is a superset of GTL research. Many GTL researchers draw upon and contribute to other research activities in the life sciences. The usefulness and acceptance of GTL data and resources depends, in part, on how they integrate with other similar activities in the larger life sciences community. Community engagement and support must be provided at all stages of the development of this infrastructure.

Transparent and facile community access to GTL computational resources—specifically the GTL Knowledgebase—for analysis, visualization, modeling, and simulations will require access at several levels. These interfaces must be both user and application friendly and enable a comprehensive integration of GTL databases. Scientists must be able to integrate problem-solving and knowledge-discovery capabilities with custom applications and with other distributed community resources; however, they will not use these capabilities unless they can understand them and have confidence that they will be available and reliable far into the future. Therefore, comprehensive training must be readily accessible to potential users, and software tools and interfaces must be well maintained and supported.



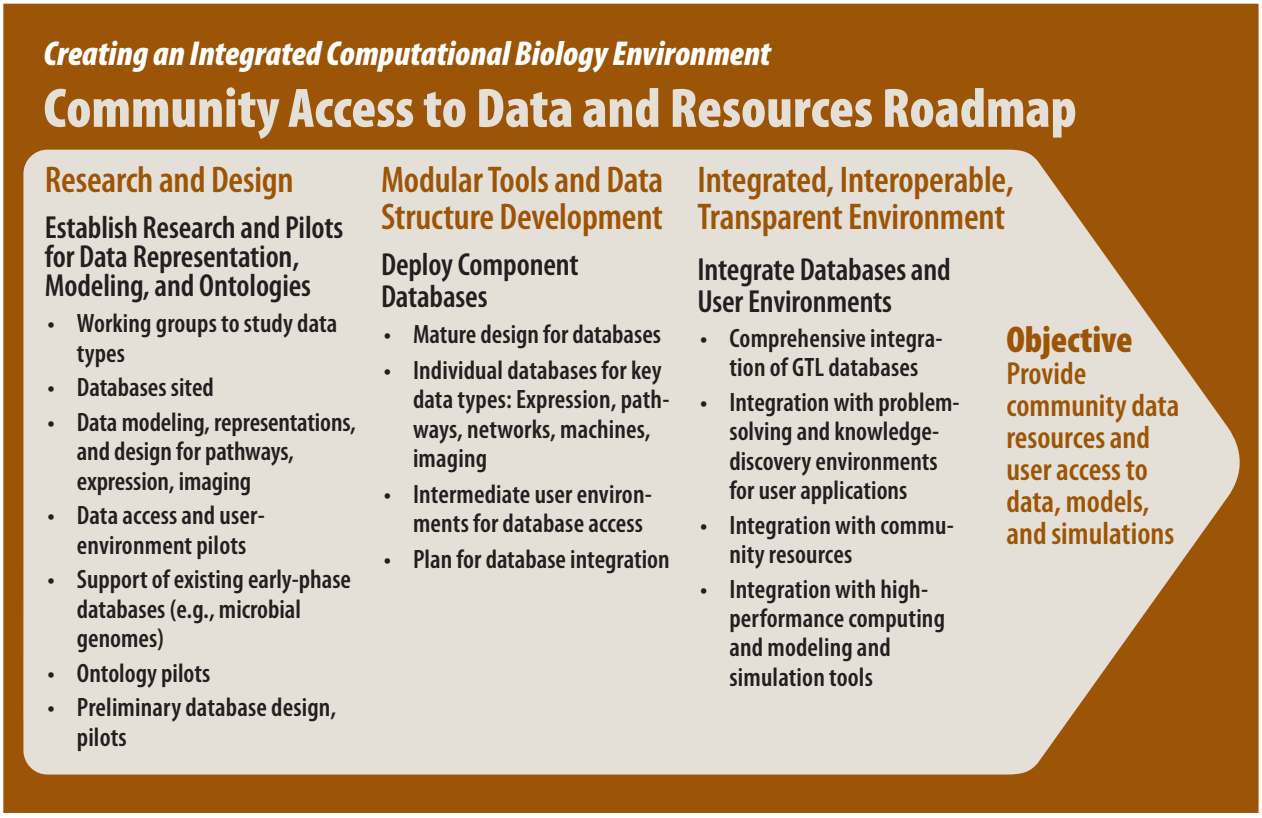
4.2.6.1. Capabilities Needed

To achieve this sixth objective, a range of technical capabilities is required. Some associated research and development challenges are listed below (see 4.2.6.2):

- Community resources for multiple types of data (machines, interactions, process models, expression, genome annotation, metabolism, and regulation).
  - Multiple levels of data—raw data, processed results, dynamic models
  - Data from other community sources
  - Protocols and methods
- Multiple interfaces to the GTL Knowledgebase to enable many kinds of queries
  - Query and update from web portals
  - Interface via web services and database languages
  - Adapters and translators to and from external community databases
  - Integration with community workflow tools
  - Integration with grid services
  - Posting of data directly into computations
- Technologies and tools for access to integrated biology view
  - Ability to cross-annotate genome, proteome, and image databases with other information (e.g., genomes with expression data, images with molecular analyses)
  - Support for automated and on-demand updates of models built on parameters from evolving GTL Knowledgebase
- Broad control over data propagation and collaboration
  - Creation of a local copy of all or part of a data set and ability to reintegrate changes later
  - Publishing of data to a limited set of colleagues and private sharing of notes with them
  - Creation and import of dictionaries and restricted naming rules
  - Propagation of data-analysis code to peers and continuous update of algorithms
- Complete documentation, training, and support services
  - Online documentation of database schema, interfaces, and access protocols with worked examples
  - Documented open-source analysis and modeling and simulation applications, with files for common systems and sample input and output
  - Periodic tutorials on database and application use at several levels
  - Help-desk support for problems and queries
  - Disaster-recovery plans for major databases

4.2.6.2. Some R&D Challenges

- Efficient management of queries that span many widely distributed databases, perhaps having varying internal organizations
- Reliable propagation of updates to replica databases and databases with information derived from central sources
- Intuitive user interfaces for browsing, querying, visualizing, and running analyses or simulations



- Design and integration of major databases, accommodating huge data volumes, large transaction rates, great schema complexity, and continually evolving content (e.g., new types of database hardware and software)
- Data standards and representation for very complex objects (e.g., object-definition languages)

See Community Access to Data and Resources Roadmap, this page.

4.2.7. Development Requirements

The integrated computational environment for biology is the critical technical core of the GTL program and facilities. It must be robust—secure and hardened against failures and down time. For all developments in research programs and technology, an accompanying computing and data-management suite will be needed to integrate all components. Coordinated development of the computing environment has a number of elements that have long lead times, are global in their impact, and crosscut facilities or program elements. These are discussed in 6.4. Computing, Communications, and Information Drivers and Issues, p. 194.

COMPUTING