DOE/SC-0113

U.S. DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Systems Biology Knowledgebase for a New Era in Biology

A Genomics:GTL Report

from the May 2008 Workshop

U.S. DEPARTMENT OF ENERGY | Office of Science

DOE GENOMICS:GTL
SYSTEMS BIOLOGY FOR ENERGY AND ENVIRONMENT

Genomics:GTL Systems Biology Knowledgebase Workshop

May 2008

Convened by

U.S. Department of Energy

Office of Science

Office of Biological and Environmental Research

Organizers

Susan Gregurick
Biological Systems Science Division
susan.gregurick@science.doe.gov, 301.903.7672

James K. Fredrickson
Pacific Northwest National Laboratory

Rick Stevens
Argonne National Laboratory

www.science.doe.gov/Program_Offices/BER.htm

As part of the U.S. Department of Energy's (DOE) Office of Science, the Office of Biological and Environmental Research (OBER) supports fundamental research and technology development aimed at achieving predictive, systems-level understanding of complex biological and environmental systems to advance DOE missions in energy, climate, and environment. Specific OBER mission priorities include the development of biofuels as secure national energy resources, understanding relationships between climate change and Earth's ecosystems, and investigating the fate and transport of subsurface contaminants.

To promote development of a data and information management system, OBER hosted the Genomics:GTL Systems Biology Knowledgebase Workshop in May 2008. Experts from scientific disciplines relevant to DOE missions and from the enabling technologies (e.g., bioinformatics, computer science, database development, and systems architecture) met to determine the opportunities and requirements for establishing and managing the proposed GTL Knowledgebase (GKB). Participants envision the GKB as an informatics resource that will support both systems biology research and DOE science-application areas, particularly biofuel development, carbon biosequestration, and environmental remediation. Also discussed were requirements for developing the GKB data capabilities needed to advance such research. To define such capabilities, participants were organized into working groups based on four GKB themes: data, metadata, and information; data integration; database architecture and infrastructure; and community and user issues. This report outlines the workshop's findings and highlights key opportunities for establishing the GTL Knowledgebase.

Suggested citation for this report: U.S. DOE. 2009. *U.S. Department of Energy Office of Science Systems Biology Knowledgebase for a New Era in Biology: A Genomics:GTL Report from the May 2008 Workshop*, DOE/SC-113, U.S. Department of Energy Office of Science (http://genomicsgtl.energy.gov/compbio/).

DOE/SC-113

U.S. Department of Energy Office of Science

Systems Biology Knowledgebase for a New Era in Biology

A Genomics:GTL Report from the May 2008 Workshop

U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research

March 2009

Genomics:GTL Knowledgebase Workshop

Table of Contents

Executive Summary

Biology has entered a systems-science era with the goal to establish a predictive understanding of the mechanisms of cellular function and the interactions of biological systems with their environment and with each other. Vast amounts of data on the composition, physiology, and function of complex biological systems and their natural environments are emerging from new analytical technologies. Effectively exploiting these data requires developing a new generation of capabilities for analyzing and managing the information. By revealing the core principles and processes conserved in collective genomes across all biology and by enabling insights into the interplay between an organism's genotype and its environment, systems biology will allow scientific breakthroughs in our ability to project behaviors of natural systems and to manipulate and engineer managed systems. These breakthroughs will benefit Department of Energy (DOE) missions in energy security, climate protection, and environmental remediation.

The Genomics:GTL Systems Biology Knowledgebase Workshop

To promote development of a data and information management system, or *knowledgebase,* DOE's Office of Biological and Environmental Research (OBER) hosted a workshop May 28–30, 2008, in Washington, D.C. Experts from scientific disciplines relevant to DOE missions and from the enabling technologies (e.g., bioinformatics, computer science, database development, and systems architecture) met to determine the opportunities and requirements for developing and managing this knowledgebase for OBER's Genomics:GTL program (GTL).

Workshop participants defined the proposed GTL Knowledgebase, or GKB, as an informatics resource that would focus on DOE science-application areas yet also be widely and easily applicable to all systems biology research. Also discussed were requirements for effective development of data capabilities for systems biology that could be applied specifically to plants and microbes (i.e., bacteria, archaea, fungi, and protists—unicellular eukaryotes such as microalgae) as well as to three areas of science related to DOE missions: (1) researching and developing biofuels, (2) advancing fundamental understanding of the global carbon cycle, and (3) understanding and using biological systems for environmental remediation. Participants were organized into working groups based on four knowledgebase themes: data, metadata, and information; data integration; database architecture and infrastructure; and community and user issues.

Summary Findings

The workshop highlighted DOE's unique and extensive data-management needs as a foundation of mission-inspired systems biology research. These needs require a principal GTL data resource, the GKB, with critical links to complementary systems supported by other agencies and community organizations worldwide. This knowledgebase would facilitate a new level of scientific inquiry by serving as a central component for the integration of modeling, simulation, experimentation, and bioinformatic approaches. The GKB also would be a primary resource for data sharing and information exchange among the GTL community. Furthermore, not only would the GKB allow scientists

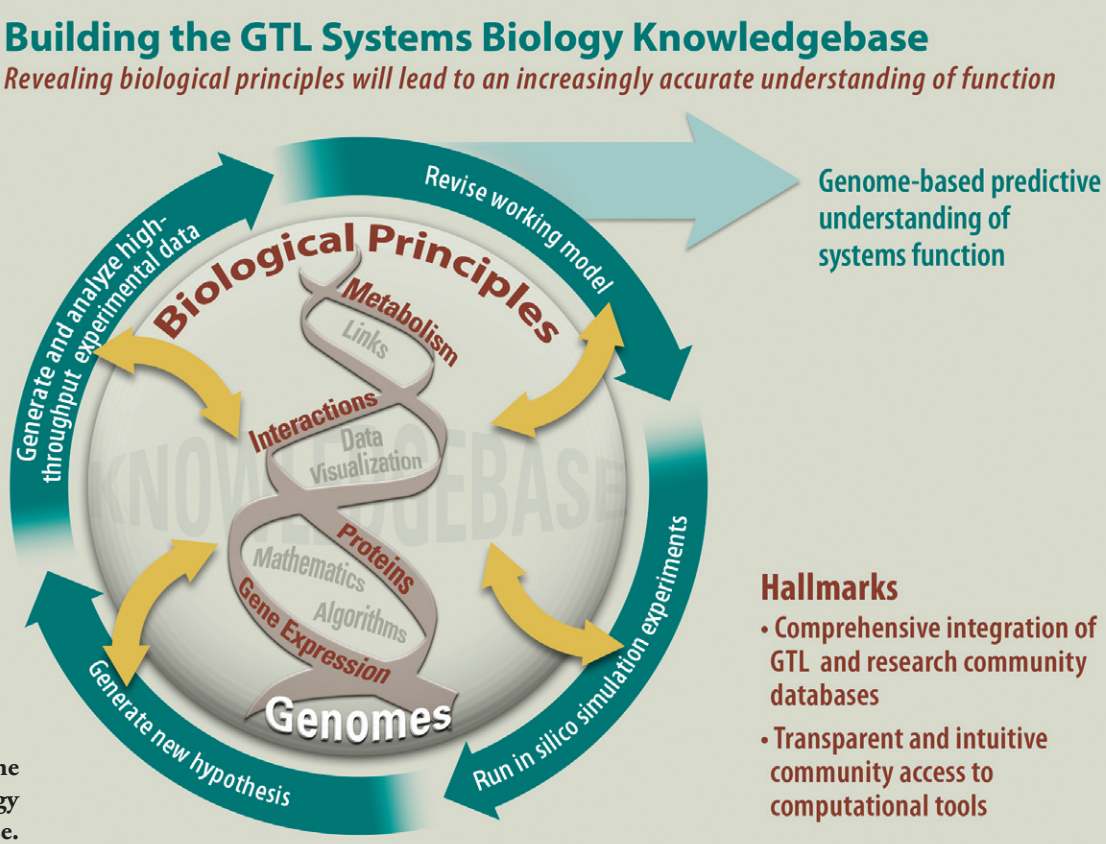


Fig. ES.1. Building the GTL Systems Biology Knowledgebase.

to expand, compute, and integrate data and information program wide, it also would drive two classes of work: experimental design and modeling and simulation. Integrating data derived from computational predictions and modeling, as envisioned in the knowledgebase project, would increase data completeness, fidelity, and accuracy. These advancements in turn would greatly improve modeling and simulation, leading to new experimentation, analyses, and mechanistic insight. Scientists' ever-increasing exploitation of the dynamic linkages among data integration, experimentation, and modeling and simulation—aided by the GKB—will advance efforts to achieve a predictive understanding of the functions of biological systems. The knowledgebase, therefore, must serve multiple roles, including (1) a repository of data and results from high-throughput experiments; (2) a collection of tools to derive new insights through data synthesis, analysis, and comparison; (3) a framework to test scientific understanding; (4) a heuristic capability to improve the value and sophistication of further inquiry; and (5) a foundation for prediction, design, manipulation, and, ultimately, engineering of biological systems to meet national needs in bioenergy, environmental remediation, and carbon cycling (see Fig. ES.1. Building the GTL Systems Biology Knowledgebase, above).

Summary Recommendations

The Department of Energy should establish the GTL Knowledgebase as a growing and extensible system of open and integrated biological, ecological, and environmental databases uniquely suited to DOE missions and distinct from, but linked to, other biological databases. To guide users and developers, the knowledgebase must be framed by a governance model with a set of user and programmatic interfaces;

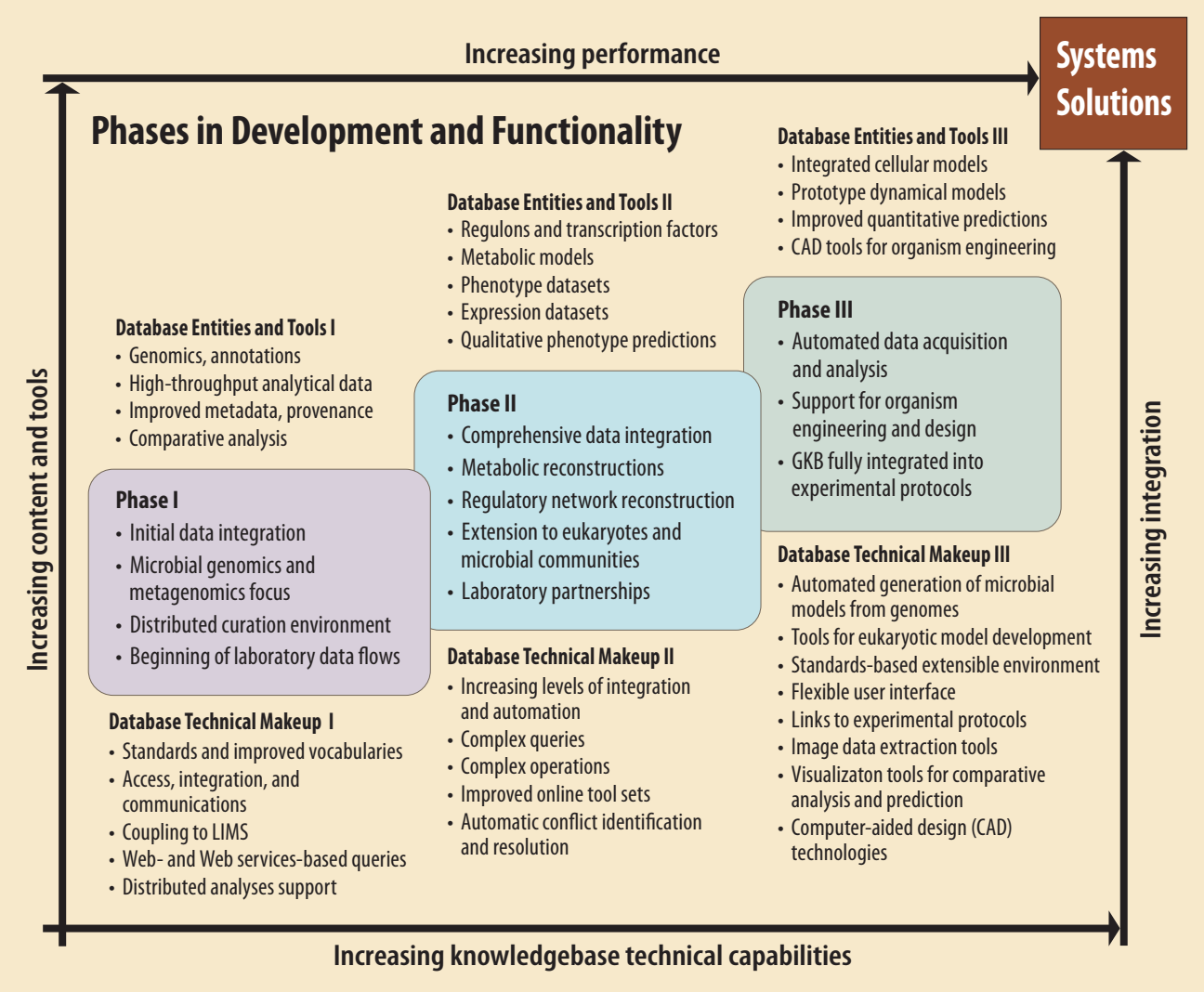


Fig. ES.2. Phases in DOE GTL Knowledgebase Development and Functionality. Phase I is centered around pulling together the components and developing functional elements. In Phase II, the components become more integrated, coupled, and automatic. In the final, mature phase, the knowledgebase becomes fully integrated, automatic, and transparent.

data-quality practices and standards; policies for data submission and data access; and a supporting communications, computing, and informatics infrastructure. Robust knowledgebase use among members of the scientific community would require a consonant suite of algorithms, tools, and services for data analysis, visualization, annotation, curation, extraction, and mining of datasets. Providing these resources would involve capturing a rapidly growing flow of data, correcting errors, and enlisting the expertise of researchers skilled in data integration, analysis, and extraction. Moreover, to support the ultimate goals of systems biology and DOE missions, the GTL Knowledgebase should be the focal point for a set of capabilities to reconstruct, model, and simulate biological and ecological systems. Workshop participants prioritized development of these integrated capabilities and outlined a strategy to implement each in phases to span a 5-year period (see Fig. ES.2. Phases in DOE GTL Knowledgebase Development and Functionality, above).

Implementation: Levels of Capabilities, Content, and Functionalities

The first and arguably most straightforward phase to implement involves the establishment of capabilities to gather new data for high-quality genomic annotation. This important process identifies and assigns biologically meaningful descriptions to DNA sequences, for example, by identifying genes, developing metabolic reconstructions, and creating estimates of regulons and regulatory circuitry. Additional challenges are associated with annotating the genome sequences in eukaryotes (e.g., protists, plants, and fungi) and metagenomic samples relevant to DOE missions. To meet these annotation challenges, the GTL Knowledgebase would require substantial automation and greater data and information depth (e.g., increased accuracy, consistency, and coverage) and breadth (e.g., expansion from hundreds to thousands of genomes). In this paradigm, the concept of annotation must be extended to encompass the growing suite of datasets and objectives of high-throughput experimental systems biology.

Building on the initial data integration and annotation of Phase I, the next developmental phase involves coupling knowledgebase components to bridge the spatial and process scales in an organism. This second phase includes draft reconstructions to progressively encompass greater complexity that would characterize individual proteins; their interactions to form molecular machines; and, ultimately, metabolic and regulatory pathways and networks in functioning cells. A synergistic, two-way, and iterative workflow is envisioned for this phase, in which annotations provide the foundation for reconstruction, and reconstruction imposes consistency on annotations. The knowledgebase must comprehensively integrate all relevant data and information for Phase II to be viable.

These integrated genome-scale (whole system) reconstructions will describe progressively more complex cellular networks, contributing to predictive modeling of physiological properties, behavior, and responses at the organismal level. The two mission-relevant and readily tractable layers of reconstruction to be developed at this stage would be metabolic and transcriptional regulatory networks in bacteria, archaea, and unicellular eukaryotes. Such reconstructions would lay the foundation for applying similar techniques to more complex systems—including plants and microbial communities—and have the potential to capture temporal and spatial aspects of systems behavior. They also would provide a natural framework for integration of various types of genomic and postgenomic data (e.g., proteomic).

The third phase involves predicting and manipulating the functions of biological systems. Accomplishing these objectives requires integrating different layers of reconstruction (e.g., metabolic and transcriptional regulatory networks) to generate more realistic, predictive models of the "stable states" of organisms. Enhanced modeling of these states would allow prediction of organism behavior in response to environment, support a new generation of hypotheses, and reveal novel insights for systems design and engineering. Furthermore, the dynamic modeling of transitions between stable states—resolved for space and time—would contribute to multiscale exploration and prediction of the behavior of systems. Critical to such research is the modeling of microbial communities (i.e., prokaryotic or eukaryotic organisms such as protozoans, bacteria, archaea, algae, and fungi) and ecosystems, which includes representing associations among biota, such as plants and microbes, and their interactions with the environment. Together, the modeling of stable states, communities, and ecosystems

will enable system investigation spanning all scales—from molecular to global. To achieve advanced modeling and predictive capabilities, Phase III of the knowledgebase must include acquisition of the experimental data needed to validate physiological and functional predictions.

In summary, the long-range goals of the GTL Knowledgebase are twofold: (1) enabling and providing support for progressively more inclusive, predictive modeling of various cellular processes, organisms, and communities and (2) facilitating the use of these capabilities to inform ecosystem-level models and engineering applications. Attaining these goals would require a knowledgebase framework that precisely and comprehensively integrates data and information critical to DOE missions.

Introduction

The GTL Knowledgebase as a Foundation for Mission-Inspired Systems Biology Research

Fundamental Research Foundation

The Department of Energy's (DOE) Office of Science historically has pursued scientific frontiers to ensure a secure energy future for the United States. Today, the Office of Science focuses on simultaneously providing the scientific foundations for achieving energy growth and security, understanding climate change, and protecting the environment. Modern biology has great potential to inform sound decisions regarding U.S. energy strategy and to provide science-based solutions for a wide range of challenges. Under the auspices of the Office of Biological and Environmental Research, within the Office of Science, the Genomics:GTL program (GTL) supports fundamental science that will form the foundation for solving critical problems in biofuel development, climate stabilization, and environmental cleanup (see Fig. 1.1. GTL Science for DOE Missions, below).

Finding 1: The emergence of systems biology as a research paradigm and approach to DOE missions is founded on the dramatic increase in the volume of data from a new generation of genomics-based technologies. Data management and analysis are critical to the viability of this approach.

Systems biology is broadly defined as the study of interactions among the components of a biological system and the mechanisms by which these interactions influence system function and behavior (see Fig. 1.2. Multiscale Explorations for Systems Understanding, p. 3). A systems approach typically includes an iterative cycle of theory, computational modeling, and experimentation to quantitatively describe cells, cellular processes, or interactions. The genomics revolution—with its vast data and associated technologies—has enabled the emergence of systems biology, which offers promise for tractably addressing the complexities of DOE missions. Such an approach seeks to predict a system's collective phenotype from its collective genotype in the context of its environment. The power of the systems approach is rooted in the fact that—at the molecular level—all life is based on similar sets of fundamental processes and principles. Knowledge gained about one biological system therefore can advance the understanding of other systems when information is readily available in an integrated and transparent format.

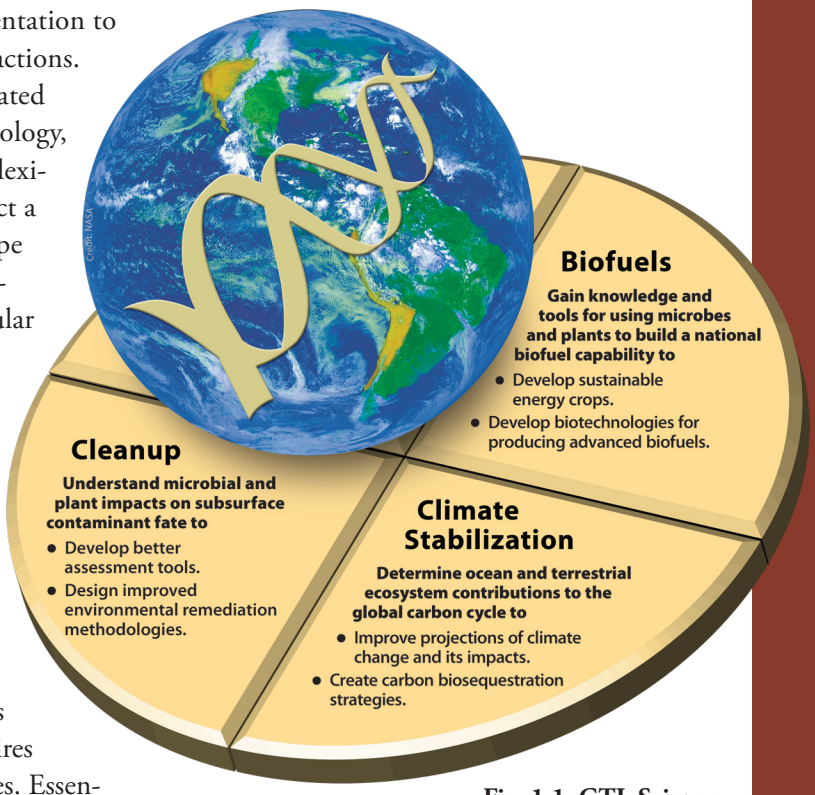


Fig. 1.1. GTL Science for DOE Missions.

Progressing from descriptive to predictive science through the use of systems biology is a goal of the GTL program. Achieving this goal depends on the ability to integrate and manage vast, diverse data. Moreover, the complex mission-inspired research that GTL pursues spans all temporal and spatial scales of biology and requires the collective expertise of scientists from many disciplines. Essential to effective research across these scales and domains are coordinated

application and integration of various technologies and experimental approaches, including genome sequencing; gene expression profiling; proteomics; metabolomics; imaging; and a wide range of physiological, functional, and even environmental data. To advance biological research, this wealth of data must be integrated, analyzed, and incorporated into modeling frameworks. The costs of associated technologies and data acquisition, the breadth and complexity of the data, and the value in relating insights across disciplines compel the open sharing of data and resulting information within the GTL program and throughout the scientific community.

Finding 2: The GTL program has several large and highly focused research efforts in, for example, systems biology, bioenergy, and genomics. Each area is investing in and dependent on rapidly growing capabilities for data resources and management, making the associated needs of each an ideal initial focus for GTL Knowledgebase development.

To facilitate communication and collaboration, GTL has broadened its research model beyond individual principal investigators to a team approach focusing on specific DOE mission areas and central challenges in biology. This approach—founded on the viability of researchers jointly using large quantities of data—requires well-coordinated efforts among scientists not necessarily co-located. Such coordination is particularly evident in the three DOE Bioenergy Research Centers, whose diverse portfolios address the challenges of this mission area and the concomitant challenges of data sharing and integration on a scale far greater than any effort to date. Similar teaming approaches have developed across GTL, including DOE's Joint Genome Institute (JGI), consortia such as the *Shewanella* Federation and the Virtual Institute for Microbial Stress and Survival (VIMSS), and smaller integrated projects of principal investigators.

Finding 3: Development and use of the GTL Knowledgebase require a comprehensive, flexible policy and supporting programs that will meet GTL's current and emerging research needs.

The long-term success of the GTL program and systems biology in general depends on establishing the capability for high-level integration and sharing of data and information. To expedite scientific and systems understanding, DOE should make such information more readily accessible to the global scientific community. Failing to do so will result in lost opportunities, barriers to scientific innovation and collaboration, and the problem of unknowing repetition of similar work. In contrast, open access to highly integrated data will enhance scientists' ability to establish links between and across disciplines. This in turn will lead to new insights into the functions of systems and these functions' potential shifts in response to perturbations. GTL is committed to open access to data and information as outlined in the program's Information and Data Sharing Policy (see Appendix 1, p. 59), which requires public accessibility to all publishable information. Ongoing development of this policy will help define standards and guidelines for establishing the GTL Knowledgebase (see http://genomicsgtl.energy.gov/datasharing/GTLDataPolicy.pdf).

DOE has a long history of successful research programs to develop data and information systems. Seeking to build on this foundation for its knowledgebase, the GTL program maintains a robust partnership with DOE's Office of Advanced Scientific Computing Research. This partnership includes continued GTL investments in both the Innovative and Novel Computational Impact on Theory and Experiment (INCITE, http://www.sc.doe.gov/ascr/INCITE) program and the Scientific Discovery through Advanced Computing (SciDAC, http://www.scidac.gov/) program. Under sponsorship

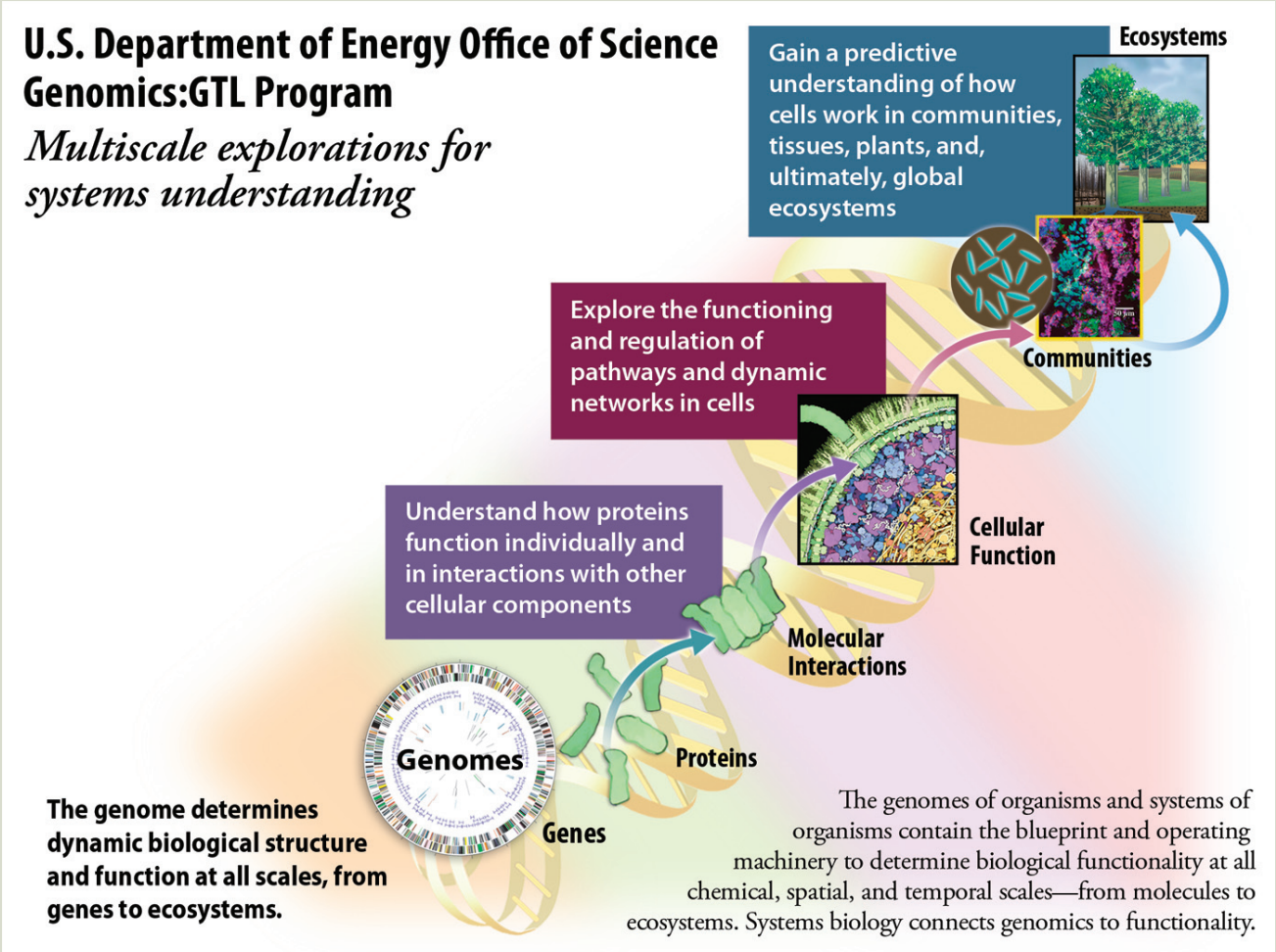


Fig. 1.2. Multiscale Explorations for Systems Understanding.

of the Office of Science, SciDAC is implementing a new integrated knowledgebase for climate research—Earth System Grid II (ESG). Overcoming the challenges associated with analyzing and deriving knowledge from global Earth System Models is the primary goal of ESG, which will include data on the global carbon cycle as described in more depth later in this chapter.

Defining and Developing the GTL Knowledgebase

The GTL Knowledgebase (GKB) is envisioned as a system for data management and information retrieval and analysis for life science investigators and computational scientists. Both groups would benefit from the availability of well-maintained, quality-controlled, and highly integrated datasets. The key objective of the knowledgebase project is to provide the computational environment needed to effectively support systems biology. This would involve integration of a rapidly growing body of relevant data, development of tools to extract and analyze the integrated data, and a commitment to ease of use and data exchange.

Dramatic progress in understanding biological systems in the past has required the use of a combination of theory, modeling, and experimentation, often in an iterative manner. Equally important today, this combination—in conjunction with GKB capabilities—would enable biologists to integrate new and existing knowledge and

information and generate novel theories to guide research. The knowledgebase framework would allow bioinformatics and computing researchers to use both existing and new algorithms for developing and testing new hypotheses. Furthermore, the GKB would be a platform for presenting, describing, accessing, and evaluating research results, new theory, and underlying data in real time. It also would provide tools for rapid data queries and design of next-generation experiments resulting in yet more insight and knowledge. As one of its primary design features, the GKB would be capable of quickly incorporating new types of data and information.

Finding 4: Researchers require the integration of a wide range of high-volume data and a computational environment designed to support modeling, derivation of predictions, and exchange of data.

The GTL Knowledgebase should explicitly support the exportation and reintegration of data and software programs from groups of computational scientists. Such a capability would enable new groups to develop software either compatible with or easily integrated into the GKB through relevant abstractions in a convenient and compatible format. This integration would lead to new software and data products flowing into and from the knowledgebase. Several existing smaller-scale data projects incorporate these features, as discussed in Chapter 2, Data, Metadata, and Information, p. 19.

Continually assessing the data needs of the systems biology community and using these assessments to define the appropriate scope of the GKB are critical to the success of the GTL Knowledgebase. These data requirements must be balanced among all stakeholders, and the information needed to support the most substantial GTL research projects should be explicitly identified at the outset of knowledgebase planning.

Overview of the Science Enabled by the GKB and the Resultant Data and Analysis Requirements

What Are the Data Requirements of Systems Biology?

Biological systems are more than a series of finite components working together to produce an effect. Although such systems are composed of finite elements, the multifunctionality of these elements produces a functionally diverse organism or community that may express only a small portion of its genetic repertoire at a particular time or in a given environment. As an example, consider the ever-changing nature of oceanic microbial communities as they respond to variations in environmental conditions. Historically, scientists have parsed biological systems into smaller, finite components to better understand their individual functions. However, research seeking to integrate these components ultimately will yield collective—and thus more realistic and relevant—insight into the overall function of a biological system. This integrative research approach, a relatively new concept in biology, is the basis for systems biology (see Fig. 1.2, p. 3; an elaboration of this diagram for global carbon research is in Box 1.1, beginning on p. 10).

To collect quantitative data for model construction and validation, various high-throughput methodologies are used, including genome sequencing, gene expression profiling, proteomics, metabolomics, new molecular-specific imaging techniques, and cutting-edge approaches for gathering environmental data. The analysis and modeling framework incorporating the resultant information and data then constructs, in a functional hierarchy, the molecular machines, pathways, networks, and

cellular systems and communities carrying out biological function, allowing further levels of inquiry.

In short, systems biology is a living science. As such, experimental data should be process oriented, integrative, explanative, and incorporable into a reliable modeling framework that will support the predictive capabilities needed for this approach to succeed. Moreover, two enabling requirements for effective systems research are the sharing and integration of heterogeneous data and information. Currently, such data often are stored in numerous locations and databases having inadequate annotation and contextual information, inconsistent data standards, and little or no connections to or compatibility with other information systems.

Finding 5: Systems biology is contingent on the ability to integrate and utilize a wide variety of types of data and computational technologies to systematically address a progression of problems leading to effective modeling of organisms.

Major priorities for the GTL program, therefore, are developing and implementing a GTL Knowledgebase to overcome these deficiencies in data and information management. Long-range objectives include enabling and providing support for progressively more precise and comprehensive predictive modeling of various cellular processes, organisms, and communities and facilitating the use of knowledgebase capabilities to inform system models (e.g., from populations in bioreactors to ecosystems). To accomplish these goals, the GTL Knowledgebase would provide seamless access to all layers of content—from underlying data, tools, and algorithms to high-level conjectures. This access would be available to all types of users, from scientists developing new computational techniques to those pursuing focused applications, and would encompass data at all levels of biological hierarchy, from individual genes and pathways to entire organisms and environments (see Table 1.1. Hierarchy of GTL Knowledgebase Applications, p. 7). Figure 1.3. Modeling Marine Ecosystems: Genomes to Biogeochemical Cycles, p. 12, illustrates how these features might be employed across the biological scales shown in Box 1.1, Global Carbon Cycling Research, beginning on p. 10.

A major challenge facing environmental scientists is using genome-based data to gain insight into metabolic processes occurring at the molecular and microscopic scales and then scaling these activities to inform biogeochemical processes and rates at macroscopic levels in the field. These processes and their rates are essential for predicting the fate and transport of radionuclide contaminants in complex subsurface environments such as those at DOE's Hanford site. This biogeochemical information also is critical for understanding carbon transformations in terrestrial ecosystems that ultimately must interface with global models to predict climate feedbacks. The GTL Knowledgebase would support these objectives by providing an essential foundation for connecting genome-based data to environmental properties and developing metabolic models with predictive capacities.

Finding 6: The GTL Knowledgebase should lead to the creation of abstract models that demonstrate increasing correspondence with the underlying physical reality. These models would play increasingly important roles in addressing major applications of interest to DOE.

Although the functional hierarchy for the GTL Knowledgebase described in Table 1.1, p. 7, implies stages of implementation, the use and functionality of the knowledgebase are not confined to a linear progression of phases. After development stages are complete, the GKB will have a wide range of uses; at maturity, following the development phases described in Fig. ES.2. Phases in

DOE GTL Knowledgebase Development and Functionality, p. vii, many of the functional levels will be pursued individually and concurrently. The knowledgebase components of capability and data content will be assembled in the initial phase of development. In ensuing stages, components will be coupled and integrated to yield higher and more complex functionality. At full development, GKB capabilities, content, and functionality will be fully integrated, automated, and transparent to users.

When properly designed and positioned, the GTL Knowledgebase would assume a new role for data management systems—from one traditionally perceived as bioinformatics support of mainstream experimental research to one that actually guides such research by providing conjectures for experimental testing and by revealing the most efficient strategies for data acquisition.

Finding 7: Other agencies and groups, most notably the National Institutes of Health, have developed integrated databases for studying organisms related to human diseases. These community-driven efforts have dramatically impacted biomedical research. A similar effort in systems biology for bioenergy, carbon cycling and biosequestration, and environmental remediation will significantly aid these DOE missions.

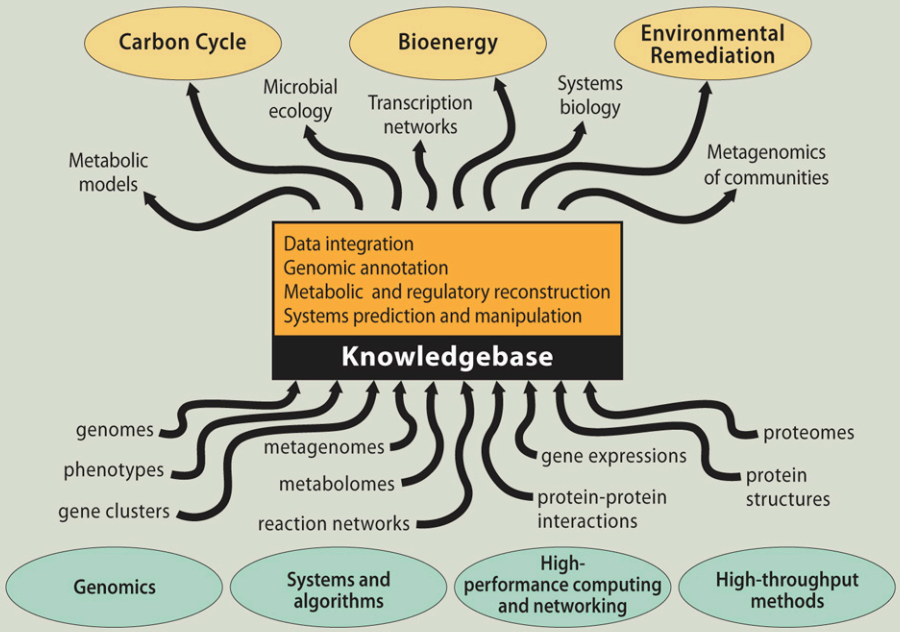
Capabilities for integrating and synthesizing various classes of existing data and data that will be acquired by current and developing technologies are recognized as major unmet needs and thus impediments to the advancement of systems biology (American Academy of Microbiology 2004). Some capabilities for the integration and comparative analysis of sequence and expression data have been developed and effectively implemented by several ongoing efforts, including successful prototypes. These prototypes have proven valuable and routinely reveal the need to integrate information at even more breadth and depth (e.g., phenotype, structure, and phylogenetic and population distribution). Key development challenges for such integrations mostly involve the scaling and automation required to support inference-generating workflows (e.g., annotations, reconstructions, and modeling). Even greater challenges are associated with integration and comparative analysis of new classes of information, including data and measurements not considered "omic" (i.e., high throughput and mapping to the genome) yet are critical for understanding and modeling communities and key environmental processes and systems.

Finding 8: DOE's national laboratory enterprise, collective and individually, has developed much of the necessary infrastructure to rapidly deploy components of the GTL Knowledgebase. A concerted effort would be needed to integrate these elements.

Focused science-application areas associated with DOE missions will drive development of the GTL Knowledgebase—a strategy distinguishing it from currently available informatics resources (see Box 1.1, Global Carbon Cycling Research, beginning on p. 10, and Box 1.2, DOE Bioenergy Research Centers—Strategies at a Glance, p. 13). In addition, the GTL Knowledgebase would substantially employ existing DOE-wide computational infrastructures that satisfy requirements and specifications for capability and capacity, operations, security, scalability, and long-term curation (see Box 1.3, Factors in Designing, Developing, and Using the GTL Knowledgebase, p. 14).

Knowledgebase planning discussions of specific mission-inspired applications in areas such as bioenergy, carbon cycling and biosequestration, and contaminant fate and transport (see Appendices 2–4, beginning on p. 65) have revealed that many scientists studying these challenges will tap into the same suite of fundamental technologies and use core systems biology data types. Classes of GTL Knowledgebase applications would range from interpretation and modeling of organisms and communities

Table 1.1. Hierarchy of GTL Knowledgebase Applications



1. Genomic Annotation

Burgeoning capabilities to gather new data present a rapidly growing challenge in genomic annotation. To meet the requirements for GTL Knowledgebase applications in (1) annotation, (2) metabolic and regulatory reconstruction, and (3) systems prediction and manipulation, the knowledgebase would require substantial automation and greater data and information depth (e.g., increased accuracy, consistency, and coverage) and breadth (e.g., expansion from hundreds to thousands of genomes). Specific challenges are associated with annotation of eukaryotes (e.g., protists, plants, and fungi) and metagenomic samples relevant to DOE missions. Within the GTL Knowledgebase framework, the concept of annotation must be expanded to encompass the growing suite of datasets and objectives of high-throughput systems biology.

2. Metabolic and Regulatory Reconstruction

Draft Reconstruction of Pathways and Networks. This application involves characterizing individual proteins and their interactions to form molecular machines and, ultimately, metabolic and regulatory pathways and networks within the compartmentalized interior in functioning cells. To achieve this, a synergistic, two-way, and iterative workflow is needed in which annotations provide the foundation for reconstruction, and reconstruction imposes consistency on annotations. The GTL Knowledgebase must comprehensively integrate all relevant information for this development phase to be viable.

Integrated Genome-Scale Reconstruction. Such reconstructions would describe progressively more complex cellular networks, leading to predictive modeling of physiological properties, behavior, and responses at the organismal level. The two mission-relevant and tractable layers of reconstruction to be developed at this stage are metabolic and transcriptional regulatory networks in bacteria, archaea, and unicellular eukaryotes. These reconstructions are quantitative and scalable to more complex systems, having the potential to capture temporal and spatial aspects both within and among cells. They also would provide a natural framework for integration of various types of genomic and postgenomic data.

3. Systems Prediction and Manipulation

Integrating different layers of reconstruction (e.g., metabolic and transcriptional regulatory networks), in the context of environment, would generate more realistic, predictive models of organisms' "stable states." Improvements in the modeling of these states would enable predictions of organism phenotype and behavior, support a new generation of hypotheses, and reveal novel insights for systems design and engineering. Spanning all scales of investigation—from molecular to global—requires both dynamic modeling (resolved for space and time) of transitions between stable states and the modeling of microbial and mixed communities (such as plant-microbe) and ecosystems. To achieve greater modeling and predictive capabilities, this phase of knowledgebase development must contain comprehensive spatial and temporal information encompassing all physiological and functional dimensions.

(including the ability to select organisms for a task and predict and control their behavior) to synthetic biology to improve rational systems engineering (see Table 1.1. Hierarchy of GTL Knowledgebase Applications, p. 7). The following use case scenarios are distilled from Appendix 2, p. 65.

Use Case Scenarios of Systems Biology Investigations Using the GKB

The GTL Knowledgebase will support a series of high-priority objectives based on systems biology challenges and the research needs inspired by DOE missions. Described below are research examples based on these objectives, along with an indication of their relevance to mission challenges (for details, see Appendix 2, p. 65, concerning systems biology investigations).

Use Case Scenario 1

- Support a capability to rapidly assess the metabolic potential and regulatory features of any cultured, sequenced prokaryote that is of primary importance for DOE mission areas.
  - Map parts (e.g., genes) and modules (e.g., pathways, subsystems, and regulons) comprising essential life processes across thousands of diverse species (see Table 1.2, item 1, Parts and Modules, beginning on p. 16).

Mission Relevance of Use Case Scenario 1

Bioenergy

- Identify improved pathways, enzymes, and strategies for degradation and conversion of biomass by screening large, integrated datasets from natural environments.
- Within metagenomic and microbial libraries, conduct comparative analyses of component processes to pinpoint new organisms and properties that can be manipulated for enhanced biomass production.

Biogeochemistry and Environmental Remediation

- Identify critical geochemically driven metabolic pathways through comparative analyses of environmental microbial and community (metagenomic and metaproteomic) datasets.

Carbon Cycling and Biosequestration

- Understand the component metabolic and regulatory pathways determining the efficiency of photosynthesis in marine phytoplankton by analyzing metagenomic and individual microbial datasets.

Use Case Scenario 2

- Support a capability to predict and simulate microbial behavior and response to changing environmental or process-related conditions.

Mission Relevance of Use Case Scenario 2

Bioenergy

- As part of microbial manipulation efforts, use key insights—such as discovery of new bioenergy traits—to predict behaviors significant to biofuel research (e.g., the ability to degrade cellulose or ferment its component sugars to fuels). From these predictions, estimate the metabolic potential for improving the behaviors of interest. For example, evaluate whether a microbe can be altered to yield high levels of ethanol or whether it can use multiple sugars. This will include assessing the production capability for traits that scientists cannot yet manipulate (e.g., a microbial cell wall that might be tolerant to very high levels of ethanol).

Biogeochemistry and Environmental Remediation

- Develop a coupled metabolic-regulatory model of biofilm-forming heterotrophic bacteria to predict biofilm phenotype and metabolic responses to changes in nutrient and energy fluxes or to environmental perturbations. Such models can be built and validated by analyzing global physiology and expression (e.g., transcriptomic and proteomic) datasets.

Carbon Cycling and Biosequestration

- Understand the fundamental regulation of the light-harvesting and photo-protection apparatus in individual cells of cyanobacteria (prokaryotes) in response to changing ocean environments, including light conditions.

Use Case Scenario 3

- Expand Use Case Scenarios 1 and 2 to encompass key application-related aspects of more complex target organisms such as unicellular and multicellular eukaryotes, including fungi, microalgae, and plants.

Mission Relevance of Use Case Scenario 3

Bioenergy

- Use integrated analyses of plant genomic and physiological data to design improved biomass feedstocks based on insight into the thousands of genes involved in the chemical and regulatory aspects of plant cell-wall and lignocellulose formation.
- Understand the genes and processes regulating the life cycle of perennials to improve the sustainability of biofuel production.

Carbon Cycling and Biosequestration

- Derive the underlying molecular mechanistic basis of and environmental influences on plant productivity, partitioning, respiration, and carbon sequestration. This can be achieved by comparing observations, experimentation, and modeling studies of natural and model systems.

(text continues on p. 15)

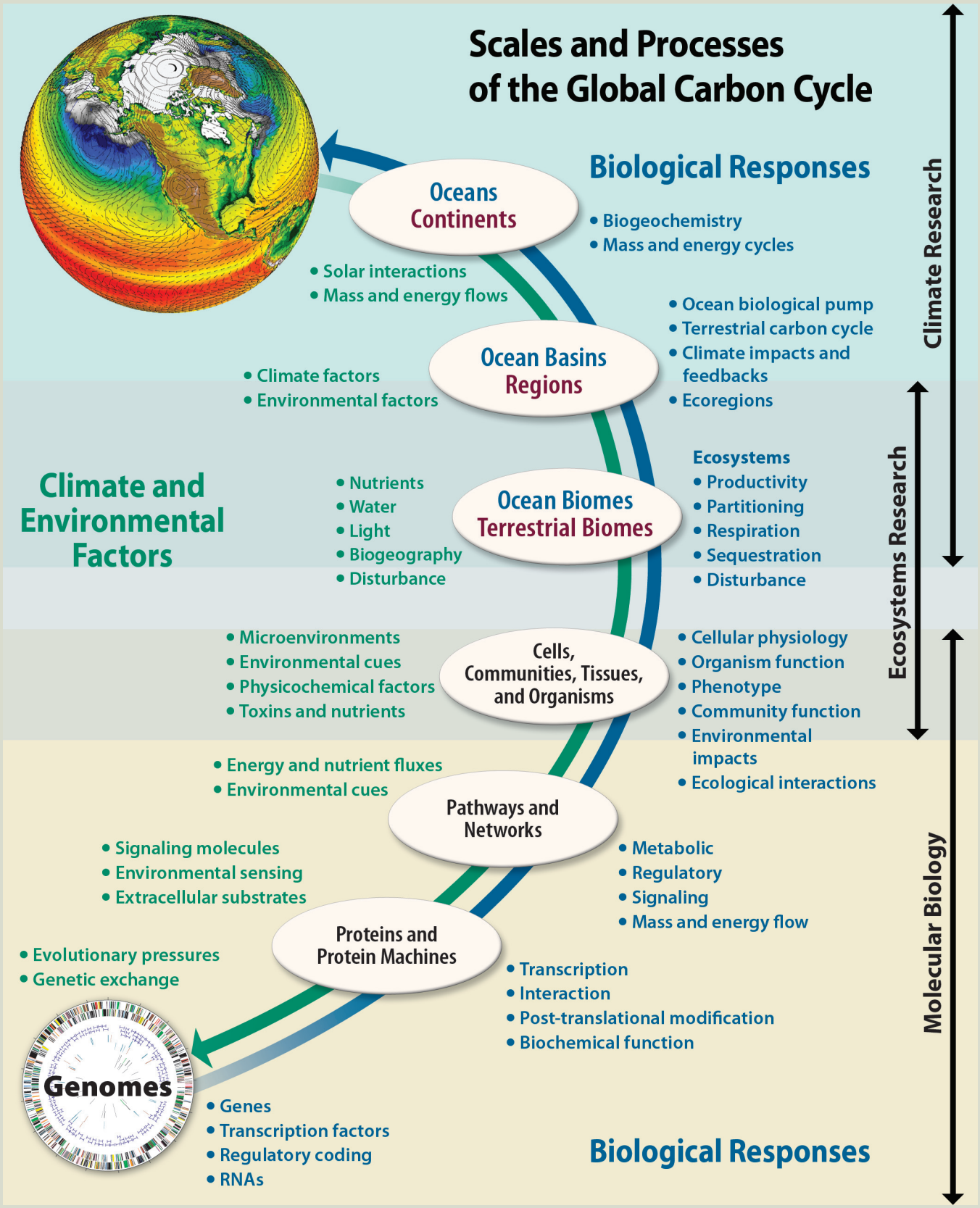
Box 1.1

Global Carbon Cycling Research

To reach a consensus on projections for future climate scenarios, the scientific community needs a better understanding of the fundamental mechanisms controlling carbon sources and sinks. Biological processes play central roles in global carbon cycling, and a mechanistic, systems-level understanding of complex biogeochemical processes at multiple scales will be essential for predicting climate-ecosystem feedbacks. Key topics in carbon cycling research include (1) photosynthetic productivity; (2) partitioning of photosynthate into energy or biomass pathways; (3) respiration mechanisms; (4) paths to recalcitrant carbon compounds and structures with long environmental residence times; and (5) the effects of environmental variables, nutrients, and water in the context of climate change. Biological communities significant to the global carbon cycle are microbes responsible for primary photosynthetic production and decomposition in oceans and symbionts and decomposers of plant-derived photosynthate in terrestrial systems. A broad understanding of carbon cycling will help define options for biosequestration in managed ecosystems as strategic elements for mitigating atmospheric CO_2 increases that result from human activity. Research details from DOE's Carbon Cycling and Biosequestration Workshop can be found in the report, *Carbon Cycling and Biosequestration: Integrating Biology and Climate Through Systems Science* (U.S. DOE 2008, http://genomicsgtl.energy.gov/carboncycle/).

Global ecosystems display tremendous complexity—with plants, microbes, and other biota working in multifaceted webs and associations. This complexity challenges carbon cycling research with the classic problem of scaling—connecting spatial and temporal levels of molecular processes to the macroscales of ecosystems and beyond (see Box 1.2, DOE Bioenergy Research Centers—Strategies at a Glance, p. 13, and figure at right, Scales and Processes of the Global Carbon Cycle). Understanding carbon cycling processes at all scales and coupling them across these levels will require all the capabilities and features envisioned for the GTL Knowledgebase.

Scales and Processes of the Global Carbon Cycle. The global carbon cycle is determined by the interactions of climate, the environment, and Earth's living systems at many levels, from molecular to global. Relating processes, phenomena, and properties across spatial and temporal scales is critical for deriving a predictive mechanistic understanding of the global carbon cycle to support more precise projections of climate change and its impacts. Each domain of climate, ecosystem, and molecular biology research has a limited reach in scales, constrained by the complexity of these systems and limitations in empirical and modeling capabilities. While comprehensive linkage of genomes to global phenomena is intractable, many insightful connections at intermediate scales are viable with integrated application of new systems biology approaches and powerful analytical and modeling techniques at the physiological and ecosystem levels. Biological responses (blue) are to the right of the systems ovals, and climate and environmental factors (green) are to the left of the systems ovals. [Globe portion of figure courtesy of Gary Strand, National Center for Atmospheric Research, with funding from the National Science Foundation and the Department of Energy.]


Scales and Processes of the Global Carbon Cycle
Biological Responses
Oceans Continents
• Biogeochemistry
• Mass and energy cycles
• Solar interactions
• Mass and energy flows
• Ocean biological pump
• Terrestrial carbon cycle
• Climate impacts and feedbacks
• Ecoregions
Ocean Basins Regions
• Climate factors
• Environmental factors
Climate and Environmental Factors
• Nutrients
• Water
• Light
• Biogeography
• Disturbance
Ocean Biomes Terrestrial Biomes
Ecosystems
• Productivity
• Partitioning
• Respiration
• Sequestration
• Disturbance
• Microenvironments
• Environmental cues
• Physicochemical factors
• Toxins and nutrients
Cells, Communities, Tissues, and Organisms
• Cellular physiology
• Organism function
• Phenotype
• Community function
• Environmental impacts
• Ecological interactions
• Energy and nutrient fluxes
• Environmental cues
Pathways and Networks
• Signaling molecules
• Environmental sensing
• Extracellular substrates
• Metabolic
• Regulatory
• Signaling
• Mass and energy flow
Proteins and Protein Machines
• Evolutionary pressures
• Genetic exchange
• Transcription
• Interaction
• Post-translational modification
• Biochemical function
Genomes
• Genes
• Transcription factors
• Regulatory coding
• RNAs
Biological Responses
Climate Research
Ecosystems Research
Molecular Biology

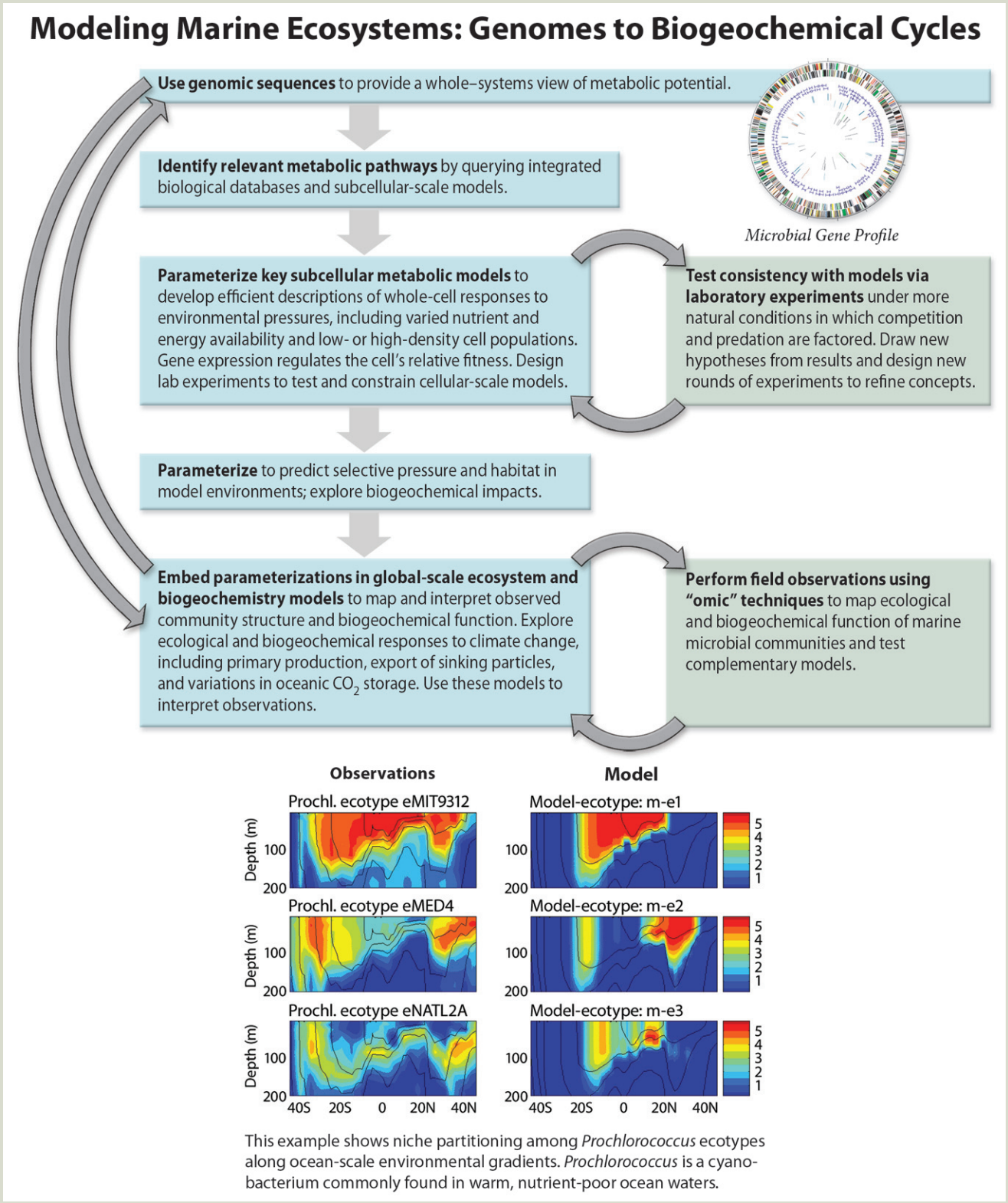


Fig. 1.3. Modeling Marine Ecosystems: Genomes to Biogeochemical Cycles. Depicted are observed and modeled distributions of ecotypes of *Prochlorococcus* [log (cells ml^{-1})]—one of the important marine phytoplankton groups—along a meridional transect in the Atlantic Ocean. Black lines indicate isotherms. Observations are from Johnson, Zinser, et al. (2006). Model ecotypes that qualitatively reflected real-world counterparts in terms of *Prochlorococcus* geographic habitat, ranking of abundance, and physiological specialism were emergent in the self-assembling model of global phytoplankton communities (Follows et al. 2007). [Source: Observations and Model graphs used with permission from *Science* and AAAS.]

Box 1.2

DOE Bioenergy Research Centers—Strategies at a Glance

Achieving industrial-scale bioenergy production requires overcoming three biological grand challenges:

- Development of next-generation bioenergy crops for easier conversion and more sustainable production.
- Discovery and design of enzymes and microbes with novel biomass-degrading capabilities.
- Discovery and design of microbes that transform fuel production from biomass.

The complexity of these challenges demands numerous coordinated research approaches to ensure timely success. The DOE Bioenergy Research Centers* represent a portfolio of diverse and complementary scientific strategies that will address the three grand challenges on a scale far greater than any effort to date. All these strategies (some of which are listed briefly below) rely on the use of data from high-throughput genomic analyses and other technologies and from screening for complex phenotypes of many natural or modified microbes and plants. One such complex phenotype is plant biomass resistance to degradation (or its recalcitrance). To effectively use and mine the data amassed from these methods, the Bioenergy Research Centers require viable development, maintenance, and operation of the GTL Knowledgebase (GKB), which would encompass relevant bioenergy domains and links to broader knowledge. Each center would use multiple GKB capabilities—including complex assemblages of metabolic and regulatory networks—described in Table 1.1. Hierarchy of GTL Knowledgebase Applications, p. 7. Scientists do not fully understand the functions of the thousands of genes and pathways involved in lignocellulose formation in plant cell walls nor those of the hundreds of genes influential in microbial hydrolysis and fermentation into fuels [see Appendix 3, Systems Biology for Bioenergy Solutions, p. 79, and *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda* (U.S. DOE 2006), http://genomicsgtl.energy.gov/biofuels/]. The data and analytical capabilities of the GTL Knowledgebase hold promise for facilitating improved understanding of these functions.

Listed below are DOE's three biological grand challenges for bioenergy production and brief descriptions of the strategies each Bioenergy Research Center is pursuing to address them.

Challenge: Development of Next-Generation Bioenergy Crops	
Center Strategies	• **BESC** – Decrease or eliminate harsh chemical pretreatments by engineering plant cell walls in poplar and switchgrass to be less recalcitrant; simultaneously increase total biomass produced per acre. • **GLBRC** – Engineer "model" plants and potential energy crops to produce new forms of lignin and more starches and oils, which are more easily processed into fuels. • **JBEI** – Enhance lignin degradation in "model" plants by changing cross-links between lignin and other cell-wall components; translate genetic developments to switchgrass.
Challenge: Discovery and Design of Enzymes and Microbes with Novel Biomass-Degrading Capabilities	
Center Strategies	• **BESC** – Screen natural thermal springs to identify enzymes and microbes that effectively break down biomass at high temperatures; understand and engineer cellulosomes (multifunctional enzyme complexes for degrading cellulose). • **GLBRC** – Identify combinations of enzymes and pretreatment needed to digest specific biomass types; express biomass-degrading enzymes in the stems and leaves of corn and other plants. • **JBEI** – Improve performance and stability of enzymes harvested from the rainforest floor and other environments; engineer, through directed evolution, highly efficient cellulase enzymes.
Challenge: Discovery and Design of Microbes That Transform Fuel Production from Biomass	
Center Strategies	• **BESC** – Reduce the number of cellulosic ethanol production steps by engineering a cellulose-degrading microbe to produce ethanol more efficiently. • **GLBRC** – Reduce the number of cellulosic ethanol production steps by engineering an efficient ethanol-producing microbe to degrade cellulose. • **JBEI** – Connect diverse biological parts and pathways to create new organisms that produce fuels other than ethanol; engineer organisms to produce and withstand high concentrations of biofuels; derive useful chemical products from lignin degradation.

*BESC: BioEnergy Science Center; GLBRC: Great Lakes Bioenergy Research Center; JBEI: Joint BioEnergy Institute. http://genomicsgtl.energy.gov/centers/

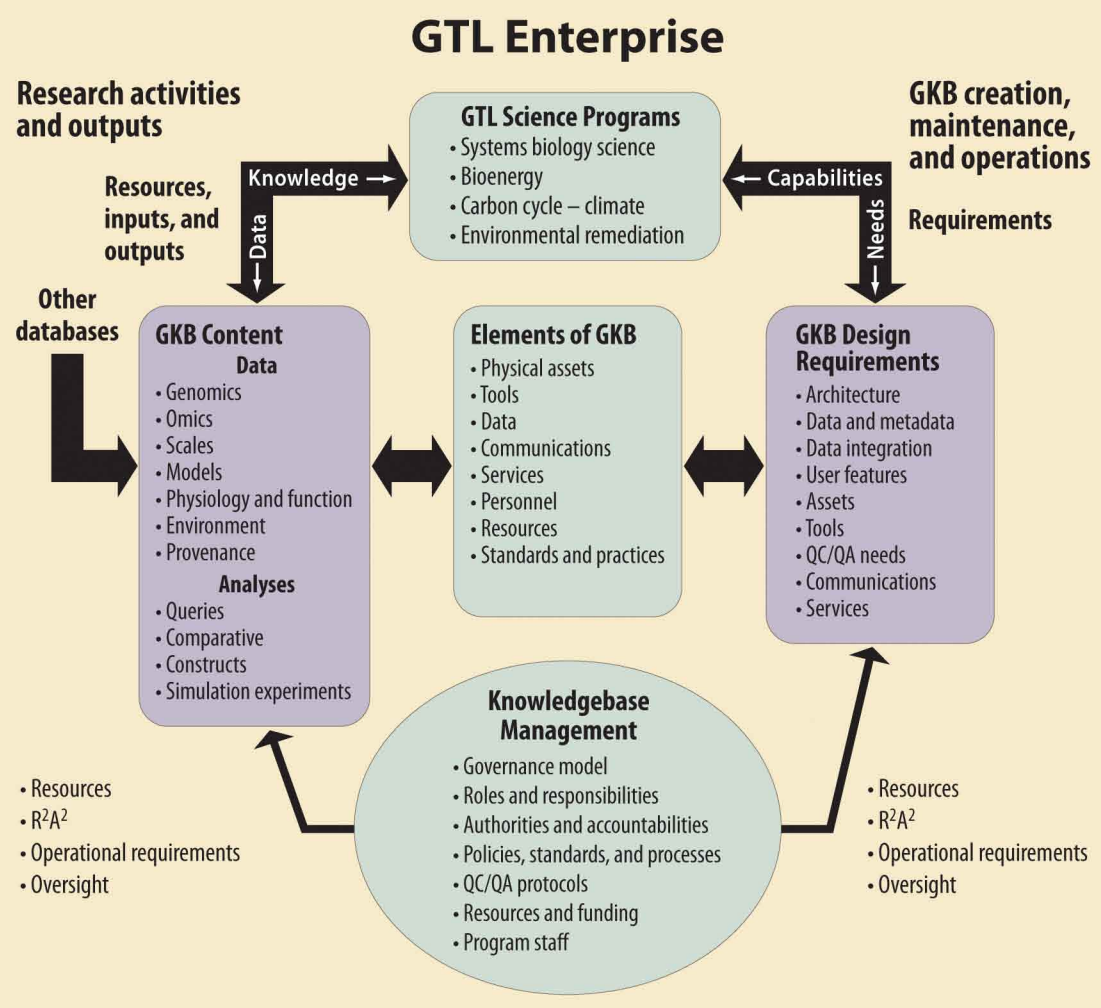
Box 1.3

Factors in Designing, Developing, and Using the GTL Knowledgebase

The GTL Enterprise is the coordinated operation of GTL science programs and the enabling knowledgebase (see figure, GTL Enterprise, below). Two major functions of the science programs are to provide requirements for GTL Knowledgebase (GKB) creation, maintenance, and operation and to establish the needed data and information that the knowledgebase would commensurately supply. GTL science programs also provide the research community with the resources to use and contribute to the knowledgebase. Furthermore, these programs would supply data and information inputs to the GKB and perform analyses resulting in the output of knowledge sought by GTL. Information from other databases also would be incorporated into the knowledgebase as needed.

GTL science programs emphasize systems biology approaches to fundamental scientific challenges in bioenergy, carbon cycling, and contaminant fate and transport. These programs also pursue a variety of other research objectives described in this report and produce diverse data, including those resulting from genomic analyses and accompanying global omic information. Also produced are various types of imaging data; information on the spatial and temporal scales of systems studied; results from modeling experiments; measurements of physiology, function, and the environment; and provenance data for documenting the results of analyses. Analyses conducted by GTL science programs include those that are comparative as well as queries and simulation experiments.

Design features and requirements envisioned for the GTL Knowledgebase (see figure) involve system architecture; provision for heterogeneous data and metadata; data-integration capacity; intuitive user elements; various assets such as computational hardware in multiple locations; tools; quality control/quality assurance (QC/QA) capabilities; communication among data providers, integrators, and users; and other GKB services. The resultant knowledgebase and its infrastructure would be a cooperative endeavor between the biological research community and computational and information scientists who would establish physical GKB assets, required tools, data repositories, appropriate communications capabilities, services, expert personnel, appropriate resources for users, and standards and practices for data providers and users. Knowledgebase developers will create a governance model outlining oversight; operational requirements; and the roles, responsibilities, authorities, and accountabilities for users and those maintaining and operating the GKB (see Box 5.1, Elements of the GKB Management Plan, p. 57). Accompanying these components of knowledgebase management (e.g., standards and processes, QC/QA protocols, program staff, and resources and funding), the GTL program will provide GKB operational requirements, oversight, and resources for research programs and will define the roles, responsibilities, authorities, and accountabilities (R^2A^2) of the GKB community.



Use Case Scenario 4

- Expand Use Case Scenarios 1 and 2 to include progressively more complex communities and ecosystems with multiple temporal and spatial scales.

Mission Relevance of Use Case Scenario 4

Bioenergy

- Gain a better understanding of the biological influences on plant acquisition of nutrients to advance biofuel crop sustainability and productivity. For bioenergy feedstock production, improving nutrient uptake (e.g., to decrease fertilizer use) is a central part of the debate on biofuel energy balances and sustainability. Nutrient uptake is linked to interactions within plant-microbe communities in the soil. Improved knowledge of how these communities function to help plants receive nutrients and water will enable strategies to increase both biofuel productivity and sustainability.

Carbon Cycling and Biosequestration

- Conduct comparative studies of marine phytoplankton communities to better understand oceanic carbon cycling and biosequestration. These studies will examine the strategies and low-level regulation of the acquisition of nitrogen, phosphorus, iron, and other limiting elements by marine phytoplankton, leading to greater insight into the role of competition in microbial community organization in oceans. The composition of these phytoplankton communities is significant in determining the efficiency of oceanic carbon cycling and storage.

Biogeochemistry and Environmental Remediation

- Use omics-based analyses and biogeochemical models to improve functional predictions of subsurface microbial communities active in contaminant transport. Prediction of contaminant transport at the mesoscopic and field scales requires understanding microbial community responses at multiple locations in heterogeneous subsurface environments and then linking this information to reactive transport models ultimately scaled to the field. Understanding the response of subsurface microbial communities to changes in contaminant and nutrient fluxes at the microscale will require (1) integrated analyses of multiple metagenomes with reference to genomes of cultivated organisms as anchors; (2) metabolic and regulon reconstructions; (3) analyses of in situ expression data (e.g., transcriptomic and proteomic); and (4) development of models of community metabolism and concomitant biogeochemical function.

Table 1.2. Critical Datasets and Data Types, beginning on p. 16, summarizes the data and information needed to support these use case scenarios.

Table 1.2. Critical Datasets and Data Types

Parts and Modules	Metagenomic Data and Microbial Communities
To Accommodate Microbial Diversity • Thousands of complete genomes (from ongoing efforts in DOE, the National Institutes of Health, and other agencies) • Ecologically important taxa for which no representatives have yet been sequenced (e.g., the majority of marine protists) • High-quality annotations (from the GKB and other sources) and inferences • Improvement in gene calling and annotation for organisms for which homology-based approaches currently are failing (e.g., marine protists that are highly divergent from other sequenced eukaryotes) • Experimental support of key inferences, both legacy and those to be systematically generated (from PubMed and DOE) • Taxonomic (i.e., phylogenetic) data (from the National Center for Biotechnology Information and potentially from the GKB) • Protein folds, domains, motifs, features, and cofactors [from the Protein Structure Initiative and public archives such as Pfam (protein families database) and Structural Classification of Proteins (SCOP)] • Metabolites and reactions [from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the GKB] **To Target Selected Organisms** ***Unconditional Data (Annotations)*** • Dozens of closely related but distinct genomes around a target (from DOE) • Highly and iteratively curated parts and modules, including annotations, subsystems, complexes, and regulons (from the GKB) • More detailed models of protein structures (from the Protein Structure Initiative and additional modeling capabilities available through the GKB) ***Condition-Specific Data (Operations)*** • Qualitative phenotypes [e.g., nutrient uptake and use (from DOE)] • Genetic tools and conditional gene essentiality (from DOE) • Gene expression, proteomic, and metabolomic data (from DOE)	• Massive sequencing of strategically selected samples (e.g., ecodiversity and applications) • New types of annotations (e.g., embedded uncertainty, clusters of genes, and neighbors) • New types of inferred modules (e.g., "fuzzy" metabolic potential) • Environmental (nongenomic) data in time (day and night) and space (e.g., geography and depth) • Community composition by 16S and other phylogeny markers for binning and global inferences; also capture of data for eukaryotes and viruses • Application-specific probes and markers (e.g., carbohydrate metabolism arrays) • Expressed genes, abundant proteins, and metabolites • Metadata • Imaging of interactions among cells; spatial patterning (e.g., layers in biofilm); community composition and co-localization of species [e.g., using fluorescence in situ hybridization (FISH)]; and key metabolites and enzymes (e.g., using mass spectrometry (MS) imaging of nitrogen fixation and tracing spatial flows of labeled carbon or nitrogen)

Table 1.2. Critical Datasets and Data Types (continued from p. 16)

Reconstruction of Metabolic Function	Complex Genomes (Limited Set of Model and Target Organisms)	Reconstruction of Transcriptional Regulatory Networks, Predictive Modeling, and Integrating Biology and Applications
• All the above parts and modules plus more condition- and application-specific data • Biomass composition • Quantitative phenotypes (i.e., physiological data) • Media, nutritional, and other requirements for robust growth or desired property • Mutant phenotypes (e.g., conditional gene essentiality and synthetic lethals) • Quantitative assessment of metabolites and fluxes • Kinetic measurements of selected enzymes (first steps toward dynamic modeling)	• More genome sequences (driven by application areas) • cDNA and other data to assist in gene calling (e.g., splicing) • Draft reconstruction package (e.g., annotations of parts and modules) • Variations [single nucleotide polymorphisms (SNPs)] versus traits • Limited "omics" package (as in the previous three items) • Subcellular localization, organelles, and -somes [using imaging techniques such as electron microscopy (EM) and MS]	• All the listed parts and modules as well as omic data related to gene expression and function • Transcription start sites (TSSs) to define promoters (including alternate TSSs) • Changes in gene expression (mRNA, ncRNA, tRNA, and rRNA) • Protein levels [using isotope-coded affinity tag (ICAT), isobaric tag for relative and absolute quantitation (ITRAQ), stable isotope labeling with amino acids in cell culture (SILAC), and peptide counts] • Protein associations (functional relationships and genome context) • Protein-protein interactions [using MS, yeast two-hybrid (Y2H) experiments, co-immunoprecipitation (Co-IP), and crosslinking] • Protein-DNA interactions [using electrophoretic mobility shift assay (EMSA) and chromatin immunoprecipitation (ChIP) methods, including ChIP-chip (combined with microarray technology) and ChIP-Sequencing] • Protein localization (using imaging techniques) • Cellular substructures (using EM and structural reconstructions) • Post-translational modifications (proteomics) • Meta-information describing environmental context in which these data were collected

Technical Components of the GTL Knowledgebase

Data, Metadata, and Information

Modern systems biology is inherently dependent on a variety of data to inform statistical inferences, mathematical modeling, and theoretical work. The GTL Knowledgebase (GKB) thus should provide the appropriate data types, metadata structures, visualization capabilities, and analysis and inference tools to enable critical synergy between computational sciences and more traditional experimental approaches.

The GKB should focus on the acquisition, integration, and accessibility of a rich body of data. Effective use of the knowledgebase will require evolving standards to support emerging research themes. The GKB must incorporate processes to receive, transmit, and update information; it also should contain protocols for documenting and assessing the state and quality of the system and its contents.

In addition to hardware, software, and network capabilities, a broader view of the GTL Knowledgebase clearly reveals the need for sustained support of core personnel with scientific and information technology expertise.

To better understand GKB requirements relating to data and metadata, several critical issues must be addressed, including (1) data and their generation by experimentation or simulation and modeling, (2) the use of metadata for setting the context of data to enable their interpretation, (3) data handling (e.g., archiving, annotation, and maintenance), and (4) quality control and assurance (see Fig. 2.1. GTL Knowledgebase, below, and Box 2.1, Data Stewardship and Availability, p. 20).

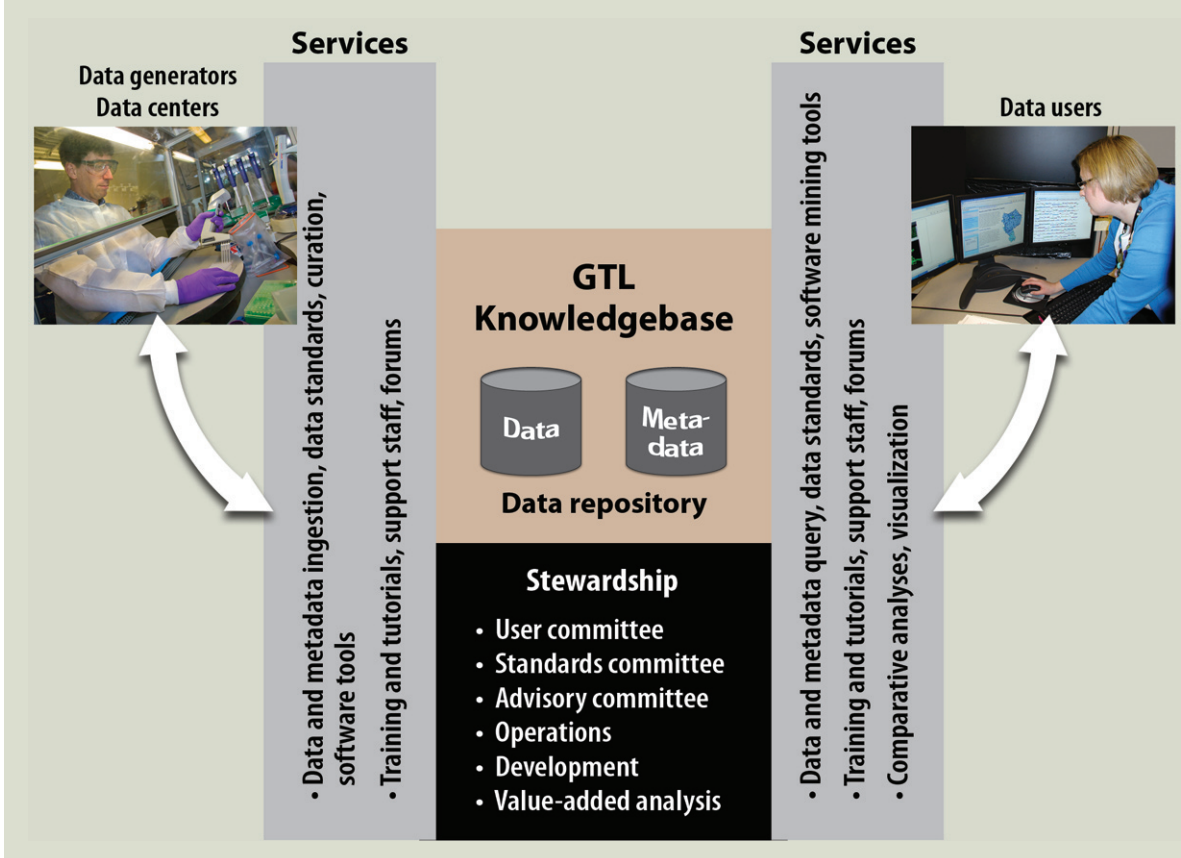


Fig. 2.1. GTL Knowledgebase (GKB). More than a repository of data and metadata, the GKB includes services and tools to generate knowledge from research data.

Box 2.1

Data Stewardship and Availability

Proper stewardship of GTL-generated data will maximize the scientific impact of the program's research investments and will support additional investigations using data-mining activities provided by the GTL Knowledgebase.

- Data submitted to the GKB become public and available to anyone desiring access.
- Regarding data embargoes, the GTL Knowledgebase should be available to the user community for prepublication analysis of experimental or computational data and information. Providing this service would require devising data embargo guidelines that will append the current GTL Information and Data Sharing Policy (see Appendix 1, p. 59, and http://genomicsgtl.energy.gov/datasharing/). In this circumstance, the GKB would serve two functions: integrating publicly accessible data and information and facilitating the analysis of data and information for additional research conclusions.
- The knowledgebase community should develop a reasonable, clear, and extensible embargo policy that can evolve to accommodate the increasing use of nongenomic datasets (e.g., images and simulation outputs).

Data Sources

Findings

The GTL Knowledgebase should support a wide variety of highly complex data from many sources. These data must be comprehensively integrated and structured for analyses and discovery.

- The GKB should gather or link to data from public repositories so users can perform complex queries across information in public systems and across GTL-derived data in the knowledgebase. Public data systems of interest include the Universal Protein Resource (UniProt), Kyoto Encyclopedia of Genes and Genomes (KEGG), National Center for Biotechnology Information (NCBI), and more topically oriented databases (see Appendix 10, List of Web Addresses, p. 139). Effective integration with these and other external data sources is centrally important as institutions make rapid advances relating to many types of relevant data.
- As the primary data repository for GTL-funded projects, the GKB must effectively relate data from such projects to the growing wealth of external data and analytical tools.
- Inferred data, which are the products of modeling activities, comparative analyses, and simulations, are expected to become increasingly important components of the GKB as its use among the scientific community grows.

Recommendations

- **Formal benchmarking.** Although no singularly integrative systems biology knowledgebase currently exists, there are excellent best-in-class particular databases from which to draw examples. The GKB should benchmark data and information standards and, in certain cases, systems interoperability against best-in-class relevant data repositories.
- **Realistic scope and expectations.** The GTL Knowledgebase is an ambitious endeavor, requiring active participation by scientists. For example, using knowledgebase data and services for scientific investigations and then feeding resultant data and knowledge back into the GKB, when coupled with existing data management resources, could constitute 10% to 20% of researchers' efforts. Because of

its scale, the GKB should be developed in phases with consideration to existing, established data management systems. The initial phase should support critical mission-relevant research and foundational science with well-defined needs and should provide resources to facilitate data access and ingestion. Implementation of these features would provide immediate value to the scientific community and would serve as a prototypic template for knowledgebase expansion.

- **Development of a database of critical information.** An extensive list of data entities has been compiled and itemized for capture in the GTL Knowledgebase (see Table 1.2. Critical Datasets and Data Types, beginning on p. 16). Selecting and prioritizing data types for GKB inclusion should be critical first steps in defining system requirements.
  - As part of its early activities, the GKB project should further develop a database of the identified entities and include data types, data volumes, and current format standards. Database development could be facilitated by surveying GTL principal investigators and establishing a website to collect survey data. The database should be reviewed regularly by the scientific community and perhaps be discussed and evaluated during the annual Genomics:GTL Contractor Grantee Workshop.
  - Once database development is under way, data entities should be prioritized in terms of importance to the GTL community and the challenges associated with incorporating the entities and establishing standards for each. The GKB project likely would have a practical limit determined by available funding, which thus will help define the scope of knowledgebase data.

Metadata

The term "metadata" refers to information about data, such as how an experiment was performed, which organism was studied, and what methods were used for data analysis. Because metadata allow scientists to reproduce results, capturing metadata is vital for meaningful knowledgebase use among the scientific community. In many cases, metadata will follow existing community guidelines of minimum standards and ontologies (i.e., structured, controlled vocabularies) set forth by community-driven efforts.

Findings

Metadata management is a core capability that will allow integration of data generated from different technologies. Critical to the success of the GTL Knowledgebase are the following key elements:

- Effective metadata management with common descriptions of data elements across multiple laboratories and investigators.
- Common descriptive language for integrating data from multiple investigators (currently a limiting factor).
- Metadata management tools that allow data generators to easily annotate and describe their data products and to extend metadata ontologies.

For GKB users to make comparisons among data and experimental results, each dataset from an environmental sample must be accompanied by metadata that provide contextual information. Such information would include, for example, the environment from which the sample was collected, methods used in collection and sample processing, types of analyses conducted on a given or nearby sample, and the overall sampling plan.

At the lowest level of information submitted to the GKB—that of data entities—data standards should define the minimum metadata requirements that must accompany all submissions.

- Where appropriate, these standards should be provided with templates and tools to help data generators harvest metadata and plan experiments before data are collected.
- Much (if not most) metadata should be collected before generating experimental data, but this process must be defined and prepared prior to conducting the experiment.

Recommendation

- Proper organization and capture of metadata are essential to providing the data context consumers require. Since collection of metadata often will be viewed as burdensome, effective standards describing required metadata must be established. Plans for enforcing these standards should be introduced early in the GKB project.

Data Handling

Archiving Raw Data

Findings

- Deposition of raw data from various technologies (e.g., imaging or mass spectrometry) into a comprehensive archive such as the GTL Knowledgebase is, in many cases, impractical. For each data-generating technology, the GKB project must determine the level of resultant data that could be captured. For cases in which raw-data capture is impossible or impractical, the knowledgebase should provide references to the sources of such data.
- Many technologies employed in GTL research generate large volumes of raw data that often are processed in complex analysis pipelines to produce final data products useful to the scientific community. For example, mass spectrometry–based proteomic analyses currently can generate terabytes of raw mass spectra that are processed to produce information about the peptides and proteins present in a biological sample.
- Since the GKB would not manage all raw data, the long-term availability of such information cannot be guaranteed. Researchers thus should devise their own local data-preservation strategies at the conclusion of a project that has generated raw data potentially important for future studies.

Recommendation

- The GTL Information and Data Sharing Policy (see Appendix 1, p. 59) should define the responsibilities of data generators in managing their raw data for the lifetime of their projects and in preserving data upon project completion.

Annotating Data

A central goal of the GTL Knowledgebase is supporting annotation, with the objective of achieving improved accuracy by removing inconsistencies and reducing ambiguity. Pursuing this objective will involve coordination with DOE's Joint Genome Institute (JGI; see sidebar, Analysis and Annotation at DOE's Joint Genome Institute, p. 23).

Analysis and Annotation at DOE's Joint Genome Institute

Analysis of DNA sequence at the Department of Energy's (DOE) Joint Genome Institute (JGI) is performed through a combination of centralized data processing and distributed data analysis capabilities. Extensive sequence annotations and analyses are generated for DOE's scientific community by JGI partner labs—including the Hudson Alpha Institute for Biotechnology, Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), Los Alamos National Laboratory (LANL), Oak Ridge National Laboratory (ORNL), and Pacific Northwest National Laboratory (PNNL). Through an extensive data and computing hardware infrastructure (550 terabytes and 1600 processors at the LBNL and LLNL Production Genomics Facilities alone), analysis of genomes from a diverse cross-section of the tree of life includes rich annotation, curation, and comparative genomics studies. Results of these studies—many of which are present in a wide variety of JGI public databases and high-profile scientific publications—underlie the value of genomics to the scientific community.

Annotation of plant genomes is carried out by DOE JGI's Computational Genomics group in collaboration with other researchers at the institute and elsewhere. State-of-the-art methods for gene prediction using ab initio, homology, and expressed sequence tag (EST) data are integrated to produce gene sets. Research efforts include applying new technology ESTs to improve gene predictions and incorporating small-RNA datasets. Comparative analysis of plant genomes is facilitated by Phytozome (http://www.phytozome.net), a hub for plant genomics.

Comprising more than 75% of DOE JGI sequencing capacities, the annotation of eukaryotic microbes is a significant component of the institute's informatic and analytical activities. Annotation and analysis of these genomes are exploited by more than 80% of JGI users and result in a considerable portion of the JGI publications in *Nature* and *Science*. The success of eukaryotic annotation is based on a community annotation program that is unique among genome-sequencing centers and highly valued by user communities.

Experience with prokaryotic genome annotation and comparative genomics is prevalent among DOE JGI partners and is most evident in teams at ORNL and the Production Genomics Facility. The flow of data—from production sequencing to assembly to finishing and annotation—is producing critical information on hundreds of new bacterial and archaeal genomes. Advancements in gene models also are being achieved through manual data curation, comparative and higher-quality functional annotations, and automated metagenome and metatranscriptome analyses. These capabilities are paving the way to new discoveries underpinning DOE missions in bioenergy, carbon cycling and biosequestration, and environmental remediation.

Furthermore, DOE JGI activities and resources significantly support the goals of DOE's Genomics:GTL program (GTL). When embraced, integrated into, and further expanded on by the GTL community, JGI capabilities can help achieve the program's vision to usher biology into a new era of systems sciences characterized by predictive understanding of the interactions of biological systems—both with their environment and each other.

Assigning function to genes and gene products is the classic concept of annotation. However, a substantially broader view is needed to describe the gradual refinement of assertions and inferences. As metabolic reconstructions, regulons, regulatory circuits, dynamic models, and phenotypic measurements and predictions are introduced into the GKB, the notion of annotation and maintenance of annotations extends significantly beyond the curation of protein function. Annotation also involves detection and removal of inconsistencies at higher levels in the biological hierarchy (for example, between phenotypic measurements and hypothesized metabolic reconstructions, such as the systematic approach used by *Shewanella* (see sidebar, *Shewanella* Knowledgebase, p. 24).

The concept of high-quality genomic annotation differs between eukaryotes and prokaryotes, largely because of the difficulties in accurately identifying eukaryotic genes (whether from plants or unicellular eukaryotes). At minimum, high-quality

Shewanella Knowledgebase

The *Shewanella* Knowledgebase is designed to provide a framework for investigators to share, combine, and analyze data. The first version of the database was released to investigators in 2007. Priorities of this knowledgebase include (1) improving coverage and support for "omic" data, such as expression arrays and proteomic, physiological, and biological information; (2) improving data linkage to key investigators and developing procedures to capture their data streams; (3) developing database support for multiple *Shewanella* species and strains by providing, for example, tools for comparative annotation and gene-function editing across multiple species; and (4) strengthening links to other data resources in the scientific community and to reference materials.

Capabilities for investigating multiple species have been implemented, including construction of several ShewCyc pathway databases. Tools for species comparison at pathway and genome levels are available and improving; regulatory data from numerous sources have been integrated. The knowledgebase also includes computational predictions of *Shewanella* regulatory elements collected from published literature and Internet resources, such as Rfam, RibEx, Tractor_DB, RegTransBase, BioCyc, and PromScan. This information was analyzed to identify a set of basic regulatory classes to present in the database. Such regulatory elements include translated coding sequences, DNA regulator-binding sites, sigma-factor binding sites, transcription units, promoters, regulons, stimulons, and RNA regulators. The latter encompasses a diverse class of regulators, including noncoding and small RNAs, different types of terminators, and riboswitches.

Current efforts focus on advancing tools and interfaces for cross-species annotation of multiple *Shewanella* species and on supporting the manual curation of 20 available genomes. Results of such activities will provide a foundation for experimental studies using a comparative approach that can be applied to essentially any group of model organisms. The *Shewanella* Knowledgebase is now equipped with a publication-mining system that includes a list of journals and other sources with links to references, authors, and related knowledgebase projects, as well as a text-search function. Procedures to maintain and update this library are in place.

Data Analysis

The knowledgebase user interface has many intuitive guides for exploring *Shewanella* experimental results. Multiple analytical modules perform one-on-one analysis across diverse biological datasets, supplemented by corresponding visualization capabilities at various data aggregation levels and biological contexts. The user interface also provides a unified set of integration analysis tools that support ShewCyc pathways and pathway-group categories. Future releases include KEGG pathways, TIGR roles, and GO ontologies for exploring data.

Data Visualization

Various data visualization tools display *Shewanella* experimental results. One such tool compares relative expression data at the gene level, while others compare averages or percentages of under- or overexpressed genes in a pathway or pathway group. These data viewers are cross-referenced to Pathway Tools software, which contains reference pathways for multiple *Shewanella* strains.

Web Portal

The *Shewanella* Knowledgebase Web portal is a data and knowledge integration environment enabling investigators to query across *Shewanella* datasets, link to *Shewanella* and other community resources, and visualize data in a cell-system context. The Web portal offers several ways to access and analyze data. Users can download data to their computers in the original format, and various data navigation features enable data exploration on the server. The knowledgebase's infrastructure is coupled with a powerful system-wide search feature that includes *Shewanella* data and publications. The user interface of the knowledgebase is built using a combination of Web 2.0 presentation layer technologies. Its Web portal is built with HTML 4.0, CSS, and Script.aculo.us javascript library. Generally, content is dynamically generated using Java server pages standard tag library.

prokaryotic annotations must include accurately identifying genes, assigning correct functional roles to gene products, and providing estimates of operons. For a growing number of prokaryotic genomes, reasonable estimates of metabolic networks and regulons also can be included in annotations.

In eukaryotes, the process of identifying genes and assigning meaningful descriptions to particular DNA segments (referred to as gene calling) is far more challenging than prokaryotic annotation. Much focus centers on overcoming this difficulty given that gene calls form the foundation for more advanced annotations. High-quality eukaryotic gene calls will need to incorporate cDNA data, including expressed sequence tags (ESTs), which—for some protists with high gene overlap—must be directional for effective use. These gene calls also should include sequence similarity and computational predictions based on the recognition of probable splice sites. As with prokaryotes, once reliable eukaryotic annotations have been established, the next goal is placing gene products in a larger context (e.g., within a metabolic pathway, complex, or nonmetabolic subsystem). Issues relating to cellular location and tissue specificity become important, but many are just beginning to be explored. Rapid progress is anticipated, however, as access to more genomes and expression data increases. This expanded accessibility will enhance opportunities for comparative analysis and will support, in particular, gene calling.

Findings

Accurate annotation of thousands of microbial genomes and a rapidly increasing number of plant genomes is a central goal of the GTL Knowledgebase. Achieving this goal would require the following:

- Incorporation of new empirical data and inferences.
- Detection of inconsistencies across a wide variety of data types.
- Logging of each inconsistency and the change introduced to correct it.
- Collection of such logs as a source of data to streamline annotation.

Recommendations

- The GTL Knowledgebase should support development of tools to refine and expand the concept of annotation. Doing so would establish consistency and remove ambiguity in assigning function across the hierarchy of biological components and systems—from DNA to proteins to pathways and networks.
- This process ultimately must be anchored in the characterization of phenotype, which includes environmental influences. Establishing protocols to control the annotation process will be essential to GKB viability.
- The GTL community also will need to agree on cultural strategies to move beyond "expert owner–based" curation.

Supporting Creation, Storage, and Maintenance of Inferred Data

Inferred data will be produced by comparative analysis, modeling and simulation. Since one central goal of the GKB is to support derivation and validation of inferred data, the project must include standards for defining provenance, attachment of appropriate metadata, and integration with experimental data.

Findings

- Accessing and using models to make and store inferences are emerging capabilities that increasingly more knowledgebase users will employ. The GTL Knowledgebase needs to support not only development of subsystems and models but also access to the models themselves.
- Researchers will extract data entities from the GKB, process them through analysis pipelines and workflows, and generate new entities that then will be submitted to the knowledgebase.
- The GKB needs to capture appropriate metadata and provenance information.

Recommendations

- The knowledgebase community should determine how to manage and control the introduction of inferred data and models.
- In addition to data, the GKB needs to provide software tools to the GTL community and link to other sites containing relevant software of interest. Such tools would include applications that facilitate capturing and recording inferred data and provenance.
- The knowledgebase should encourage open access to applications developed within the GKB framework, thereby connecting the GTL community.

Making Quality Control and Assurance Integral Parts of GKB Data Input, Annotation, and Modeling

Findings

- Quality control (QC) and assurance (QA) are critical aspects of incorporating new data into the GKB. Controlling quality occurs at the following two levels:
  - QA establishes processes to ensure the quality of the overall experimental program that generates data flowing into the knowledgebase.
  - QC screens data to reject faulty data.

These same quality processes can be applied effectively to inferred data and the mechanisms producing such data.

- Existing data providers may have their own protocols to control the quality of their data. However, sources often provide processed data to users as a "black box," meaning little relevant metadata are easily accessible. Such metadata provide information on how data have been processed and normalized, including the tools, parameter values, and versions of resources used.
- Another aspect of quality control involves changes to annotations and models. Such changes often result from supporting evidence discovered through a tedious, manual curation process. Often, however, neither the process itself nor the evidence is propagated in a coupled way with annotation changes.
- Managing data quality through QA processes and QC protocols and retaining and communicating information about such quality require the following:
  - Data quality and information must be established for all data products and should be communicated with all exchanges of such products. This approach

would enable users to efficiently complete evaluations of data products extracted for a specific use.

- Assembling and retaining quality information (e.g., QA processes and QC protocols) in a manner not overwhelming to data generators and consumers are significant topics to be resolved in the GKB design process.

Recommendations

The GTL Knowledgebase should provide a systematic approach for controlling the quality of data flowing into the GKB.

- Data must undergo appropriate QC protocols at their originating source. Although this responsibility for compliance lies with the source, the source also should provide metadata describing the data-processing workflow that can easily be queried, accessed, and summarized by GKB users.
- Establishing QC standards and protocols as they relate to annotation and inference must be the responsibility and an essential component of the GTL Knowledgebase. Adhering to GKB standards and implementing required protocols must be the responsibilities of data producers and users. Minimally, changes to data must be logged and detected conflicts updated and managed appropriately.
- GKB infrastructure should enable users to access and contribute to the evidence behind each act of curation. For example, an assertion of the presence of a given variant of a subsystem should be accompanied by users' ability to relate it directly to phenotypic measurements, expression data, and the functions associated with a set of proteins and to record this ensemble as evidence supporting an annotation change. The real power of data integration is manifest in these capabilities, which represent a major step forward for systems research. As such, they should be integral parts of the GKB process.
- Knowledgebase infrastructure also should provide mechanisms to quantify and record uncertainty (and dependencies) at all levels of analysis and propagate it in a consistent, probabilistic, and Bayesian manner. Doing so would involve, for example, characterizing and quantifying errors and biases across different metagenomic sequencing technologies.

Findings

- Curation is a long-lived process. Knowledgebase design should comprise methods for maintaining this process over the long term.
- Good stewardship of GKB information requires robust, ongoing curation accompanied by a mandatory independent assessment of knowledgebase data.
- In this context, curation includes—as an early step—tests to ensure data are complete, meet minimum reporting requirements, and have no obvious mistakes such as format problems and count errors.
- Testing for consistency will range from manual curation to automated checking.
- Documentation for inferred and assumed data entries must be rigorous.

Using Data Standards

Standards are a mechanism for capturing information in a form easily shared and integrated with other data or data types. Using data standards to capture data entities is the

foundation for comparative analysis and integration of information. These standards also define the minimum level of metadata that must be associated with a data entity. For some data, in the early stages of development, standards may be unavailable or even competing. Establishing and maintaining standards for incoming data thus should be essential components of the GTL Knowledgebase.

Findings

- Contributions of data and information to the GKB will depend on the adoption of existing standards or the development, when necessary, of new standards, and also on defined submission procedures.
- The user community must be involved in both the adoption of existing standards and the development of new standards, so that the standards enable reliable access to data and information deposited in the knowledgebase.
- One advantage of having standards is that they can be used to automate data extraction. For example, standardized metadata and provenance information can be obtained using automated tools and processes that remove the burden of data extraction from users.

The development of data standards often proceeds best within the auspices of international working bodies. The GTL program should establish a strong policy of adopting existing standards and seeking out emerging ones. This policy will have advantages for users, helping them develop internal data standards, and may even lead to the GTL Knowledgebase spearheading efforts to promote community-wide standards.

- The GTL Knowledgebase should use best practices to build consensus on internal consistency of data collection and management.
- Examples of existing and emerging standards to consider for the GKB are the Minimum Information about a Microarray Experiment (MIAME), Minimum Information about a Proteomics Experiment (MIAPE), and Minimum Information about a Genome Sequence (MIGS).
- Examples of organizations actively involved in developing data standards include the Proteomics Standards Initiative (PSI), Genomic Standards Consortium (GSC), Consortium for the Barcode of Life (CBoL), and International Census of Marine Microbes (ICoMM).

For models and algorithms, SBML (an extension of XML) and other markup languages have made substantial inroads in standardizing models in systems biology, at least in the context of dynamical systems modeling (Slepchenko et al. 2003; Hlavacek et al. 2006).

Recommendations

- The GTL Knowledgebase will need a standards committee to define minimum requirements, recommend adoption of community-developed standards, and initiate drafting of GKB standards as the need arises.
- This committee should be empowered to institute standards quickly and thus must establish a clear set of principles and decision-making processes to be followed. The following are operational examples:

- The GTL Knowledgebase must include enough information for skilled practitioners to reproduce any available data. Achieving this goal requires adopting and developing appropriate schemas.
- Data requirements need to address uncertainty propagation, so that all types of output data have a confidence limit, confidence interval, or other uncertainty field.
- Data input tools should be developed to ensure a model or algorithm meets all minimum requirements prior to submission to the knowledgebase.

Data Integration

Data integration is a feature that clearly expands the role of the GTL Knowledgebase (GKB) beyond an archive to a dynamic systems biology resource for progressively increasing scientific understanding. The GKB is envisioned to contain data, such as those described below, on thousands of complete genomes and thousands of metagenomic and transcriptomic samples.

- Data for each complete bacterial and archaeal genome should include estimates of gene function and regulons as well as detailed metabolic reconstructions.
- For each metagenomic sample, the GKB should provide estimates of microbial population and data on metabolic potential.
- For each of the more complex eukaryotic genomes, such as those of plants, protists (including algae), and fungi, data should include detailed estimates of genes and metabolic reconstructions for different tissues (e.g., root versus stem). These data also must cover various stages of development (e.g., in meristems and seeds), some of which have been extremely well elucidated at the molecular level.

To develop accurate and predictive models, the GKB needs to capture additional data for a limited (but increasing) number of organisms. These data include phenotypic, metabolic, expression, protein-protein, and protein-DNA measurements. Incorporating such data in the knowledgebase would advance the development of stoichiometric and regulatory models, leading to improvements in metabolic reconstructions that would be propagated to all genomes in the GKB.

The GTL Knowledgebase should integrate genomic, metabolic, regulatory, and phenotypic estimates under continual revision. Integrating such data would require ongoing curation by the GKB to ensure increased data consistency among a growing body of measurements, which would enhance the predictive capability of models and provide a new resource for the study of organisms.

Core applications within DOE are intended to drive the knowledgebase initiative. These applications typically revolve around macro processes (e.g., the carbon and nitrogen cycles). A key GKB requirement thus would be to couple these flows with modeling of individual organisms and communities of organisms. Seeking insight relating to processes that operate at widely separated temporal and spatial scales will be extremely challenging. Although many studies develop hypotheses for organisms' potential functional roles and interactions with their environment, few studies have tested such hypotheses. Because of system complexity, obtaining these measurements is a major challenge. Nonetheless, empirically determining process rates in complex systems is essential and should include appropriate experimental scaling that allows measured rates to be related to the genetic and regulatory bases for processes.

An example of research for which integrated process and activity rates are needed involves photosynthesis by marine microbes. Little is known about the rates at which different organisms (e.g., cyanobacteria versus the wide variety of protistan primary producers) take up CO_2 or the impact of competition on these organisms' performance. Similarly, much remains to be learned about the subsequent fate of photosynthetically fixed carbon as it is respired by organisms or exported to the deep ocean for long-term storage. Some of these microbes (and consequently their carbon) can descend to the deep ocean on their own; others must be consumed by

larger organisms to sink; and still others are lysed by viruses. As a result, little carbon reaches ocean depths because some is lost at each step of the way as organisms consume or degrade organic carbon and remineralize it to CO_2 through respiration. These mechanisms play central roles in oceanic carbon flow and sequestration. Hence, the GKB should capture and integrate measurements of these activities. A systems biology approach would facilitate such integration and thus enable predictive modeling (Azam and Worden 2004).

To support both the development and use of detailed models to infer properties of organisms and resolve model ambiguities, the GTL Knowledgebase should incorporate increasingly diverse data types. This growing body of measurements should be directly integrated into GKB estimates based on genomic data. To be effective, expression data should be related to metabolic reconstructions and estimates of regulons, and these relationships must undergo continual curation.

As a secondary service, the GTL Knowledgebase also should support archiving a diverse collection of data types. For example, GKB structure must allow heterogeneous structural and dynamic process data to be scalably and consistently captured, reconciled, and related, as well as retrievably stored. The knowledgebase also should include tools that will enable extraction, integration, and modeling to draw inferences and make predictions that can be incorporated into the GKB as additional data.

Defining the Scope of Data Integration

The GKB's data integration effort should have the capacity to access and interrelate thousands of prokaryotic genomes, hundreds of eukaryotic genomes, and thousands of environmental samples. This service also should provide capabilities for integrating associated phenotypes, other high-throughput data, and metadata. Furthermore, data integration needs to support effective comparative analyses, modeling and simulation, and inference of data hierarchies based on an increasing wealth of information.

This ambitious GKB integration strategy poses many challenges, two of which are particularly significant: data quantity and incorporation of numerous data types.

1. The sheer volume of data is imposing and includes both genomic and other information required to support comparative analysis and interpretation. A large, diverse collection of reference genomes clearly will form the essential framework needed to support more focused investigation of the biological communities relevant to DOE mission-application areas.
    a. A primary knowledgebase requirement is the ability to integrate an exponentially growing body of data and to provide reasonably accurate initial annotations, ongoing refinements, metabolic reconstructions, and estimates of regulons.
    b. Inclusion of data and information for thousands of metagenomic samples ultimately may constitute the bulk of the GKB's computational load. The rapid accumulation of a collection of well-annotated reference genomes will represent an essential asset supporting the interpretation of more complex metagenomic data. Central to the overall goals of the knowledgebase is the need to absorb this quickly expanding flow of new data into a framework offering tools for convenient comparative analysis. Similar to the complexity of metagenomes will be the integration of higher plants with distinct tissue types and their linkage to particular environmental locales and responses.

2. The second challenge lies in using the planned integration to support extensive incorporation and reconciliation of numerous types of data. These data range from genes and estimated gene products to metabolic reconstructions and models of regulatory circuitry.
   a. In addition to integrating large numbers of reasonably well annotated genomes, another GKB objective would be to select a limited set of organisms with specific relevance to DOE missions and to develop predictive models of them.
   b. Developing these models will impose consistency among the models, metabolic reconstructions, and experimental data that will form the foundations for biological research in this century.
   c. However, imposing consistency on these elements necessarily implies the ability to make and maintain numerous changes to widely shared and deeply interdependent data.

Today's architectures are capable of supporting the data structures and integrations envisioned for the GTL Knowledgebase. Existing data systems clearly support the feasibility and utility of an ambitious integration effort. None, however, currently addresses the opportunities introduced by recent advances in both microbial modeling and the ability to obtain and analyze metagenomic sequences.

Core Requirements for Data Integration

Improving the Quality of Data Annotation through Continuous, Semiautomatic Curation

Findings

Incorporating data annotations at various scales and resolutions is one objective of the envisioned GTL Knowledgebase. Achieving this goal would require addressing several challenges associated with the expanding scope of annotation.

- Assigning function to genes and gene products is the classic view of annotation.
- A substantially broader concept of this process is emerging, however, in the context of systems biology. This wider view includes annotated models of metabolic pathways and regulons, protein interactions and interaction networks, and three-dimensional protein structures.
- Many annotations—computationally derived from uncertain, noisy, incomplete, and complex data—contain various inconsistencies, ambiguities, and gaps in knowledge.

The infrastructure of GKB's data integration service presents a unique opportunity for improving annotation quality.

- Increasingly, research groups are successfully using integrative approaches to significantly improve the quality of data annotation. For example, the *Shewanella* Federation has demonstrated a systematic approach to detect inconsistencies between phenotypic measurements and hypothesized metabolic reconstructions (see section, Illustration of Use Case Scenario 1: Integrated Approach to Reconstruction of Metabolic and Transcriptional Regulatory Networks in Bacteria, in Appendix 2, p. 74).

- Similarly, incorporating information on three-dimensional protein structure has been highly valuable in annotating hypothetical genes (i.e., those without functional assignments) identified by genomics-based annotation pipelines. For example, of the unannotated proteins in *Halobacter* NRC-1, Bonneau et al. (2004) assigned functions to about half and reconstructed metabolic pathways by combining structural-functional predictions from the Robetta server (see sidebar, Example Analysis and Integration, p. 35) with genomic-context data and a variety of experimental information.
- While highly promising, many of these approaches significantly rely on tedious manual curation.

Recommendations

The GTL Knowledgebase should provide semiautomatic tools to expert curators to help them more efficiently improve the quality of annotations. These tools would support several activities.

- Incorporating new empirical data and inferences.
- Detecting inconsistencies across a wide variety of data types.
- Logging each inconsistency and the change introduced to correct it.
- Collecting such logs as a source of data to streamline annotation.

As the GTL Knowledgebase incorporates increasingly higher levels of data—such as metabolic reconstructions, regulons, regulatory circuits, dynamic models, and phenotypic information—the concept of GKB annotation would need to expand to encompass and maintain these entities. This expansion would require creating various resources and protocols to improve data quality, for example, the following:

- Tools to support consistency and improve confidence, particularly as the scope of data widens.
- Mechanisms to efficiently link existing knowledgebase annotations to emerging and newly published experimental evidence (e.g., from mutagenesis studies or expression profiles) that refines or confirms such annotations.
- Protocols to control annotation.

Facilitating Data Integration through Standards, Controlled Vocabularies, and Ontologies

Findings

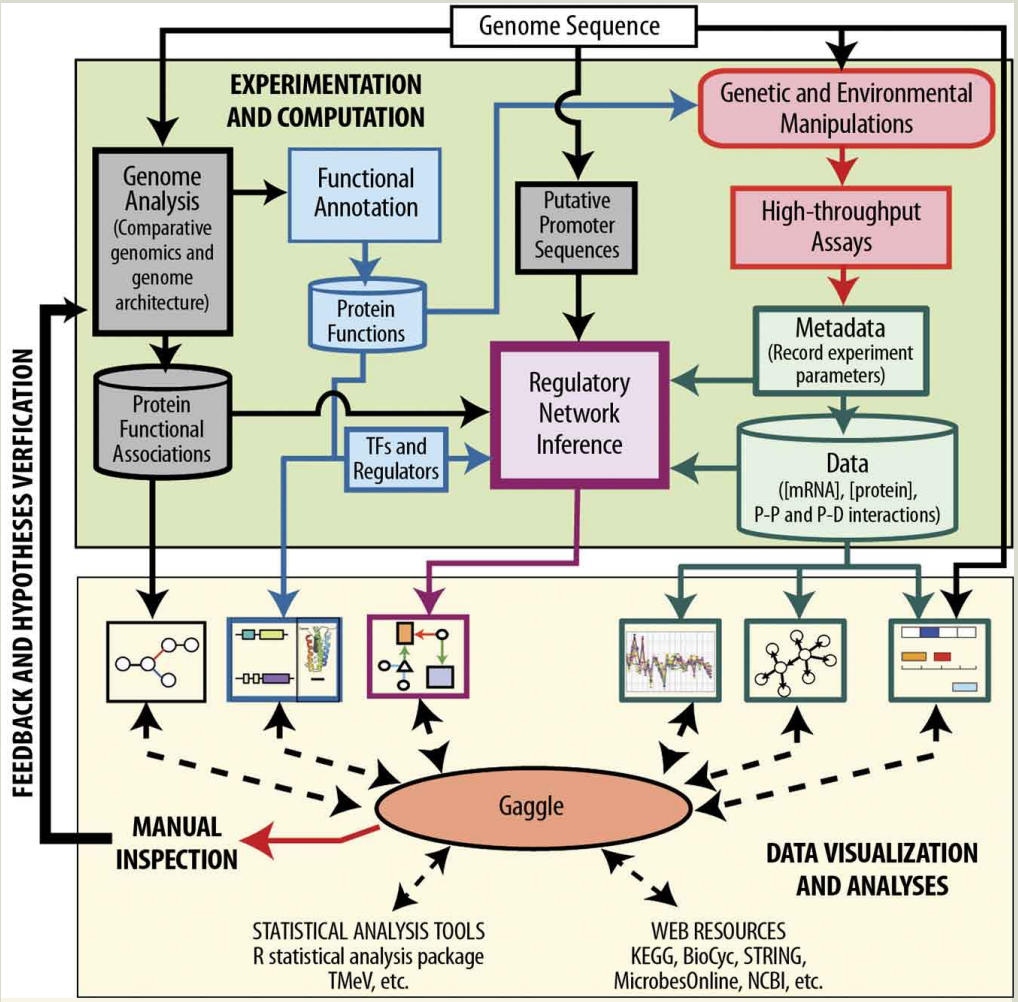
Data integration and model development in systems biology largely are hampered by the lack of semantically consistent naming conventions.

- Although different annotation systems depend on each other, they often use inconsistent definitions, resulting in decreased quality of the systems and their annotations. For example, genome annotation pipelines may use gene-function definitions inconsistent with the controlled vocabularies used by systems that annotate metabolic pathways.
- Such semantic ambiguity and inconsistency probably lead to holes in reconstructed metabolic pathways.

Example Analysis and Integration: The Process of Generating Models of Metabolic and Regulatory Networks

The ability to generate accurate and predictive models of organisms' metabolic and regulatory circuitries represents a substantial advancement in systems biology. Developing such models may be viewed as a process that produces, as a by-product, consistency among protein functions, metabolic reconstructions, and derived models. The need is to have massive data-driven and falsifiable (testable) hypotheses. The "trivial" underlying hypothesis is, "Can a network model represent the available datasets?" The ultimate driver of these models is the need to generate new predicted hypotheses that can be tested in silico and in vivo. Deriving these models requires the following data:

- Annotated genomes (including genes, transcription start sites, and operons).
- Detailed metabolic and regulatory reconstructions.
- Initial estimates of regulons.
- A list of binary associations between proteins, reflecting existing data on protein-protein interactions, relationships inferred from phylogenetic profiles, and co-occurrence information. (The number of data sources providing evidence of protein associations clearly will increase over time.)
- Estimates of transcription factors.



Example Analysis and Integration: The Process of Generating Models of Metabolic and Regulatory Networks. [Source: Adapted with permission from Elsevier. From Bonneau, R., N. Baliga, et al. 2007. "A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell," *Cell* **131**(7),1354–65 (http://www.sciencedirect.com/science/journal/00928674).]

Generating a model of an organism's regulatory circuitry involves designing manipulative experiments that induce genetic or environmental perturbations and recording measurements of the resultant changes through high-throughput assays. These measurements include (at minimum) expression data, protein-DNA binding, protein-protein interactions, and protein modifications. Each perturbation is described in a controlled vocabulary, measurements are recorded and normalized, and the resulting data pairs (i.e., the induced perturbation coupled with the observed outcome) become input for an inference process. This process involves ever-improving algorithms that use pair sets to infer aspects of an organism's regulatory circuitry. Producing an accurate model then requires (1) iteratively examining the derived regulatory circuitry; (2) reconciling it with known phenotypic data; (3) gradually understanding the sources of inconsistency; and (4) changing asserted protein function, metabolic reconstructions, and proposed circuitry to reconcile inconsistencies.

Several existing efforts support semantic normalization and standardization.

- The Open Biomedical Ontologies (OBO) Foundry (http://www.obofoundry.org) has emerged as a framework for community-ontology development that conforms to a set of principles and best practices (see Table 3.1. Open Biomedical Ontologies Foundry, p. 37).
- The Foundry and Gene Ontology (GO; http://www.geneontology.org) include many emerging GTL-relevant ontologies, such as those for plants (PO), environment (ENVO), phenotype (PATO), chemicals (ChEBI), proteins (PRO), and metagenomes (MIMS).

Recommendations

To promote interoperability and facilitate data integration, the GTL Knowledgebase should identify, adopt, and develop common standards, controlled vocabularies, and ontologies.

- Annotation systems that semiautomatically populate their inferences into GKB infrastructure should clearly define and post their underlying ontologies.
- These interdependent annotation systems should be part of a consolidated effort to conduct semantic consistency checks. Proper semantic mapping functions, synonymous lists, and controlled vocabularies need to accompany data and annotations released from and propagated into the GKB.
- For GTL researchers who are not currently served by a vocabulary or associated ontology that describes their area of research, new ontologies might need to be defined or the scope of existing ontologies expanded with input from the user community.

Increasing the Efficacy of Complex Queries from Integrated Data

Findings

Public query engines—such as NCBI's Basic Local Alignment Search Tool (BLAST), which enables analysis of genomic sequences—have significantly increased the throughput of data extraction and made this information routinely accessible to all types of users, including computationally skilled scientists and lay people.

- However, performing similar queries of integrated systems biology data still is in its infancy.
- Despite the potential to enable new discoveries, using integrative systems biology approaches to extract information of interest is a painstaking and time-consuming task that only a few scientists endeavor on their own.

The heterogeneity and complexity of data to be integrated into the GTL Knowledgebase will be substantial.

- Effective use of these data thus will require the GKB to support highly diverse advanced queries that few existing databases have encountered. Box 3.1, Typical Complex Queries, p. 38, lists several examples of such queries.
- Supporting these types of advanced integrative queries in a user-friendly, automatic, and routine manner comparable with BLAST will transform future systems biology studies.

Table 3.1 Open Biomedical Ontologies (OBO) Foundry*

Granularity	Continuant				Concurrent
	Independent		Dependent		
Organ and organism	**Organism** (NCBI taxonomy)	**Anatomical entity** (FMA, CARO)	**Organ function** (FMP, CPRO)	**Phenotypic quality** (PaTO)	**Organism-level process** (GO)
Cell and cellular component	**Cell** (CL)	**Cellular component** (FMA, GO)	**Cellular function** (GO)		**Cellular process** (GO)
Molecule	**Molecule** (ChEBI, SO, RNAO, PRO)		**Molecular function** (GO)		**Molecular process** (GO)

*Aiming to create a suite of orthogonal interoperable reference ontologies to support integration and analysis of biological data, the OBO Foundry ontologies are organized along two dimensions: (1) granularity (from molecules to populations of organisms) and (2) relation to time (a distinction between entities that undergo changes through time and the entities—processes—that *are* such changes). [Source: Adapted by permission from Macmillan Publishers Ltd. From Smith, B., et al. 2007. "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration," *Nature Biotechnology* **25**(11), 1251–55 (http://www.nature.com/nbt/).]

Recommendations

The GKB should provide easy-to-use interfaces to significantly increase the throughput of predictive inferences resulting from queries of integrative data by lay users.

The knowledgebase should support both "vertical" and "horizontal" queries.

- Vertical queries span data levels (e.g., from correlating climate data and habitats to genes found in different samples).
- Horizontal queries associate equivalent data entities across species, samples, or habitats (e.g., homologous genes between species, community composition across samples, and abundance or enrichment of metabolic pathways across habitats).

So-called canned queries in the GKB should support systems biology modeling tasks performed by a broad community of users. Both generic and model-specific information need to be automatically retrieved in response to relatively simple inputs provided by users. For example, when a user selects an organism to query, the knowledgebase should automatically compute and retrieve (in a structured and downloadable format) relevant information for the specified metabolic model of interest. This information should include the following components:

- A list of proteins (e.g., enzymes and transporters), inferred reactions, and metabolites.
- All associated information and features, including functional assignments (from various sources) and evidence; association with protein families (e.g., phylogenetic profiles); multiple alignments and phylogenetic trees for each family; domains, motifs, and structural features (known or predicted); genomic context (e.g., operons and regulons); functional context (e.g., associated pathways and subsystems); gene expression data (users may choose from integrated or uploaded datasets); proteomic data; associated reactions and metabolites; and other types of data relating to specific genes.
- Clusters (lists) of functionally coupled genes (e.g., stimulons) with a detailed correlational analysis (e.g., linkages between gene expression and pathways or between gene expression and protein levels).

Box 3.1

Typical Complex Queries

1. Which genes in genome X are known to be essential, and which are predicted to be essential? Display the differences.
2. Which of the genes in X have relatively solid annotations? Which are less reliable but have estimates of function, and which are completely uncharacterized?
3. Given genome X, what is a working estimate of the regulons in X?
4. In genome X, list—in a convenient format—the sets of genes believed to be coregulated. Then, given microarray MA, list the sets of expressed genes that agree with existing estimates of regulons and display discrepancies.
5. Which functional roles are used in model M of genome X but are not yet mapped to any specific gene or genes in X?
6. Does model M predict that organism X can sustain growth with just Y as a carbon source based on the organism's genome and other data?
7. Given the phenotype of a metabolic pathway, which genes and gene products are probably active in the steps of the pathway, and which are likely rate limiting?
8. Which transporters are required by model M? Which are mapped to specific genes, and which have been supported by experimental evidence but have not yet been mapped to specific genes?
9. Given metagenomic sample S, what is the existing best estimate of the microbial population (i.e., which operational taxonomic units make up the sample and in what relative abundances)?
10. Given two metagenomic samples, what distinguishes them? Similarly, given two sets of metagenomic samples, what distinguishes one set from the other? Given a set of genomes, which genes are common, and which distinguish one genome from the other?
11. Given a set of genomes, which subsystems are common, and which distinguish one from the other?
12. Given two sets of genomes, G1 and G2, which subsystems distinguish G1 from G2 genomes?
13. Given two different models, M1 and M2, which experimental measurement would help differentiate models? Alternatively, list differing phenotypic predictions based on M1 and M2.
14. Given a dataset, do the data have biological or technical replicates?
15. Given a particular gene, what are all its associated annotations?
16. Integrate and compare proteomic and transcriptomic datasets for the same experimental condition.
17. Conduct visualization analyses of data or model simulations (e.g., onto pathway maps).
18. Query metadata and conduct fuzzy matching of such data.
19. Which proteins have been observed for a particular organism and also across all organisms? How do protein profiles correlate with phylogenetic differences? Determine conservation of post-translational modification across organisms.
20. Obtain upstream sequences for a coding region.
21. What is the location of a protein under a specific condition?
22. Determine conservation of regulation across species.
23. Horizontal gene transfer: Which genes have been horizontally transferred?
24. How many genes relating to photosynthesis and nitrogen fixation are present in metagenomic data?

Streamlining GKB Incorporation of Dynamically Changing Biological Data

Findings

The GTL Knowledgebase should seamlessly incorporate new classes of data and models to meet the demands arising from continuous advances in both the experimental technologies producing data and the informatic methods deriving predictions from such data.

- Knowledgebase integration would involve inputs from two basic categories of data sources.
  - Projects producing initially processed data.
  - Curated information from other public data resources (e.g., UniProt, KEGG, NCBI, and topically oriented databases).

Critical to GKB integration efforts, the first category would be responsible for initial processing of experimental data, which should be normalized and condensed into a form directly incorporable into the knowledgebase.

- The most obvious example of such processing is genome sequence data, which should be incorporated into the GKB as assembled contigs, not raw reads.
- Similarly, microarray data should be normalized by their sources and accompanied by descriptions of the experiments from which they were derived; such inputs would not include images.
- To support modeling efforts, phenotypic data also should be condensed into a form suitable for GKB integration.

Enabling Integrative Capabilities for Data Analysis and Visualization

Findings

Although significant progress has been made in developing bioinformatic tools that derive predictions from individual data types, there is an emerging and critical need for tools that support comparative analysis and visualization of the results. The significance of advances in interface conception and implementation are obvious. Comparative genomic tools such as those available through KEGG, the SEED, the Expert Protein Analysis System (ExPASy), or NCBI provide good examples of integrated and easily accessible capabilities. However, while the ability to visualize data in these resources has advanced, it is far from optimal.

The variety of genomic and comparative genomic tools can be attributed to the availability of such resources on the Web. However, similar capabilities for quantitative proteomics, metabolomics, or transcriptomics are just emerging. Moreover, these tools typically are presented as stand-alone applications, making their adoption by the biological community problematic.

Recommendations

The utility of the GTL Knowledgebase would be enhanced substantially by the availability of easy-to-use, broadly accessible, and predictive visual analytical environments. GKB infrastructure should make visual analytics an integral component of the core tool set for data integration. The following three core activities should be supported to achieve this objective:

- **Adapting existing data analysis tools for ease of use and accessibility.** GKB should pursue a systematic effort to make the evolving set of analysis tools for various omic data accessible through user-friendly knowledgebase portals.
- **Extending GKB infrastructure to add new tools for data analysis and visualization.** The GKB community should develop guidelines allowing developers of data analysis tools to contribute to GKB infrastructure through easy-to-use plug-in interfaces. While Web services offer a mechanism for providing third-party tools to the scientific community, their inherent limitations make it desirable that such tools be easily downloadable and readily accessible as integral parts of GKB infrastructure.
- **Enriching GKB capabilities for integrative data analysis and visualization.** The GKB community should develop new comparative and integrative data analysis and visualization tools for creating more predictive models. Specifically needed are the development of and increased accessibility to tools for relating quantitative proteomic and transcriptomic data, finding conserved evolutionary network motifs, and linking transcriptional and metabolic network models.

Provenance

Researchers conducting laboratory experiments routinely control and record all aspects of their experimentation environment and manipulations for description in scientific publications. Similarly, biologists using the GTL Knowledgebase should be able to capture details of the in silico experimental process used to derive their results. These details, known as provenance data, include experimental information about the datasets used, the software models and tools that processed the data, and the resultant information that eventually is added to the knowledgebase. Provenance data will allow biologists not only to visualize the experimental processes used to reach a particular conclusion but also to potentially reproduce the results of a specific experiment.

Capturing provenance information in large-scale data management repositories can lead to an exponential explosion of GKB data. Knowledgebase designers thus should devise a novel, scalable strategy for provenance capture and visualization that meets specific needs of the GTL community. Various potential solutions exist (see Fig. 3.1. Example of Provenance Browser in Taverna, p. 41), and GKB planners need to evaluate them to design an architectural component that satisfies GTL requirements.

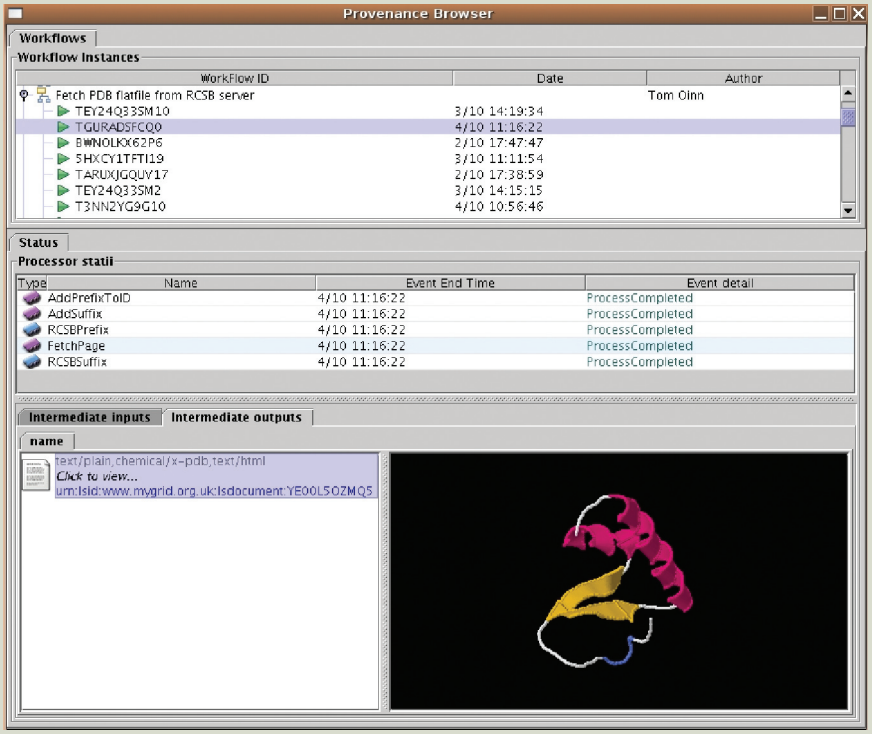


Fig. 3.1. Example of Provenance Browser in Taverna (http://www.taverna.org.uk). This feature provides a way for biologists to view the origins of data.

Technical Components of the GTL Knowledgebase

Database Architecture and Infrastructure

Rapidly advancing available and emerging technologies for computation, data storage, and communications promise a wealth of aggressive and high-performance options for establishing the GTL Knowledgebase (GKB). To take full advantage of these opportunities, GKB technical requirements and operational needs must be well defined. In addition, decisions on system architecture and infrastructure will be influenced substantially by institutional requirements manifest in the GKB governance and management model and by the resultant roles of data providers, integrators, and users. Moreover, resources for creating, maintaining, and using the knowledgebase will arise from various elements of GTL research initiatives and from computing and informatics programs and institutions. These research and computing programs in turn will influence the choices and locations of hardware and software efforts and assets.

Building a successful knowledgebase will require evaluating and implementing numerous design choices for system architecture that impact models for GKB development, scalability, and GTL-relevant use cases. Investigation of these choices and requirements has revealed several viable options for GKB architecture (each with distinct strengths and weaknesses) that could meet at least part of the data needs of the GTL community. However, GKB architectural design ultimately must satisfy the full range of GTL researchers' data requirements and provide a foundational software platform for cost-effective software development and operations. In these capacities, system architecture is fundamental to the GTL Knowledgebase project.

To meet various user and operational requirements, optimal GKB architecture most likely would be a hybrid design combining elements of several basic architectural options. This solution could link, for example, a central data model (perhaps supported at multiple sites) with more heterogeneous and distributed data and analysis support accessible through a Web services model. Existing, proven architectural designs set precedents for the success of such a venture.

Since the early development of genome databases, system architectures have undergone revolutionary changes that the GTL Knowledgebase should exploit. Major features of these changes follow:

- *Unifying algorithms and data by integrating programming languages with a database system.* This creates an extensible object-relational system in which nonprocedural relational operators manipulate object sets.
- *Integrating Web services with a core database management system (DBMS).* Such integration has significant implications for how applications are structured, with DBMSs functioning more like object containers. Online analytic processing is now integrated into most DBMSs, and service-oriented architecture (SOA) models based on Web services can be leveraged successfully in this approach.
- *Progressively incorporating new services into DBMSs.* More of these systems now have frameworks for data mining, machine-learning algorithms, decision trees, visualization, clustering, time-series analyses, and modeling—with flexibility for adding novel, integrated analytical tools.
- *Increasingly using distributed and parallel approaches based on federated or clustered architectures.* Clustered architectures, in particular, have the advantage of removing

query bottlenecks by providing shared access to numerous data storage units from multiple clients.

- *Advancing object-oriented and deductive databases.* These databases enable many concepts to be integrated into mainstream object-relational systems.

GTL Knowledgebase developers must assess the value of a spectrum of choices in designing GKB architecture. The following sections outline some of these choices and describe how they might provide important capabilities for the GTL program. Since there are significant tradeoffs among various architectural options, establishing specific requirements with which to evaluate these options is critical. Typical architecture considerations include cost, scalability, flexibility, security, query performance and reliability, and management and sociological factors.

Architecture Design Driven by User Requirements

Findings

- The envisioned GTL Knowledgebase would support existing scientific communities (e.g., biologists as well as computational and information scientists) and would help foster the growth of a much larger research community: computational biologists.
- GKB architecture thus must address a wide range of user needs—an effort requiring constant interactions between the knowledgebase design team and scientists from all associated fields. In the past, enabling computational groups to apply analytical methods to biological data has been cumbersome because data were stored in multiple locations, saved in various formats using nonoverlapping systems of identifiers, and had widely varying levels of quality.

Recommendation

- The knowledgebase should enable computational groups to easily apply new algorithms on diverse platforms to GKB data.

Basic Architecture Choices

Designers of the GKB architecture must make fundamental choices regarding where and how data for the GTL program would be collected, organized, managed, and maintained. These capabilities conceivably could be focused at a single, central site or through coordination at multiple locations. A centralized approach generally has significant limitations because it requires building a huge infrastructure at one site. Moreover, the chosen site must employ—in one location—staff having all the biological expertise necessary to organize, represent, and curate every aspect of biological data and information. Such centralized groups also have to keep track of customer requirements and avoid becoming insular.

Finding

- Given these limitations, a distributed approach to knowledgebase creation is highly desirable.

One option for GTL Knowledgebase development would involve integrated research teams such as the *Shewanella* Federation. In this approach—called a federated system—several sites specialize in a biological domain, but all members share a single data model

or global schema (e.g., a homogeneous distributed system; see Fig. 4.1. Federated Database System, this page). This enables efficient data queries and retrieval across multiple sites without the need to translate formats across different schemas. Furthermore, this approach allows each site to establish expert curators in specialized areas of biology such as proteomics or metabolic pathways. Individual sites also can use the level of resources or even parallelization each needs to support query loads. Within this framework, the efficiency of knowledgebase development is generally good because the system is divided into smaller, more manageable parts (partitions). In this model, data queries can be highly efficient within a single site's biology domain but could be somewhat slower across multiple sites. Moreover, redundancy and fault tolerance are possible within a federated database system.

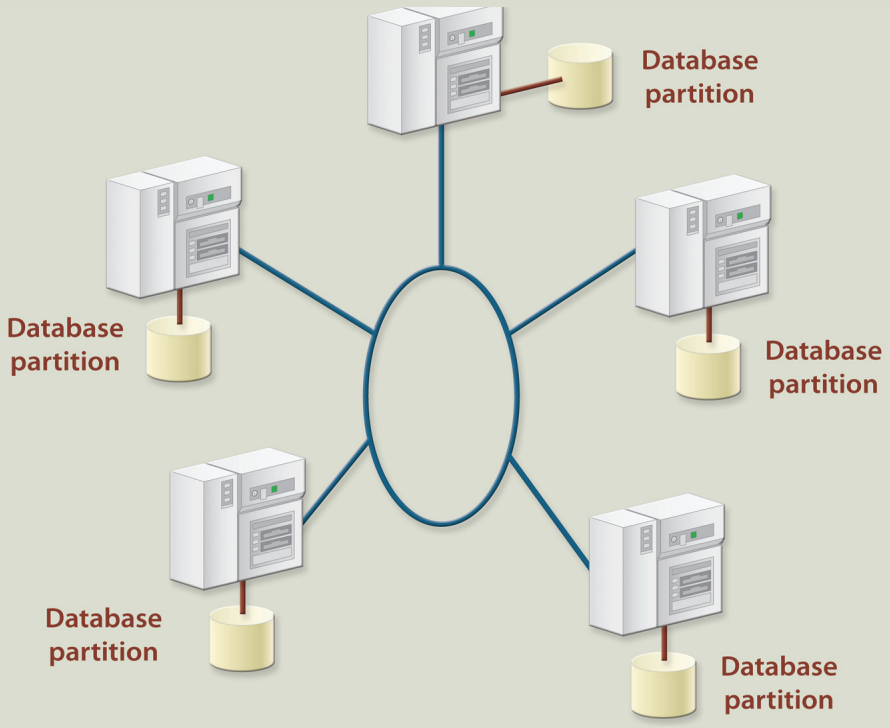


Fig. 4.1. Federated Database System. In this system, all sites share a common data model but "own" a particular part of the biology domain (i.e., horizontal partitioning) and have separate data. Queries can be directed against one or multiple sites as needed using the network.

A somewhat different approach to a distributed data system would involve a clustered architecture, in which multiple sites would mirror each other and have complete and equal access to all data through a shared group of data stores (see Fig. 4.2. Clustered Architecture Data System, this page). Such a framework would decrease redundancy and improve query efficiency, enabling optimal shared storage.

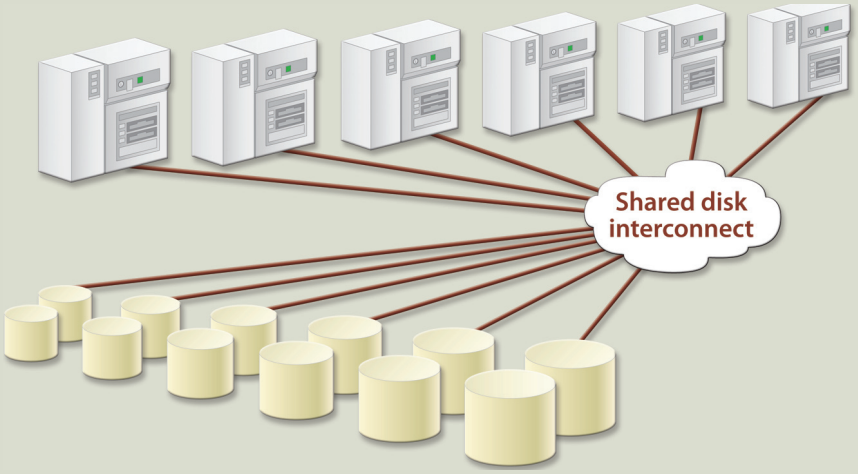


Fig. 4.2. Clustered Architecture Data System. A distributed clustered architecture has multiple server mirrors that all access a complete and shared data store. In addition, all database mirrors share a complete common data model and schema. Combinations of federated and clustered configurations also are possible and may have some advantages for knowledgebase design.

Although developers may be able to adopt a single data model for certain core information relevant to GTL, knowledgebase services would derive partially from linkages to sources of data and analysis tools outside GKB control. For example, GTL investigators could benefit from various external community databases potentially useful to their research and from other relevant analysis tools accessible through the Internet. Incorporating both external resources and the core data model into GKB architecture would require a logical partitioning of data and services as illustrated in Fig. 4.3. Conceptual Overview of GKB Architecture, p. 46. Core GTL data likely would be well understood and stable in terms of the data model; external data and services, however, are subject to faster evolution and potential instability that would need to be tracked by the GTL Knowledgebase. Community development of standards and ontologies is thus necessary for easy access and meaningful use of these external resources.

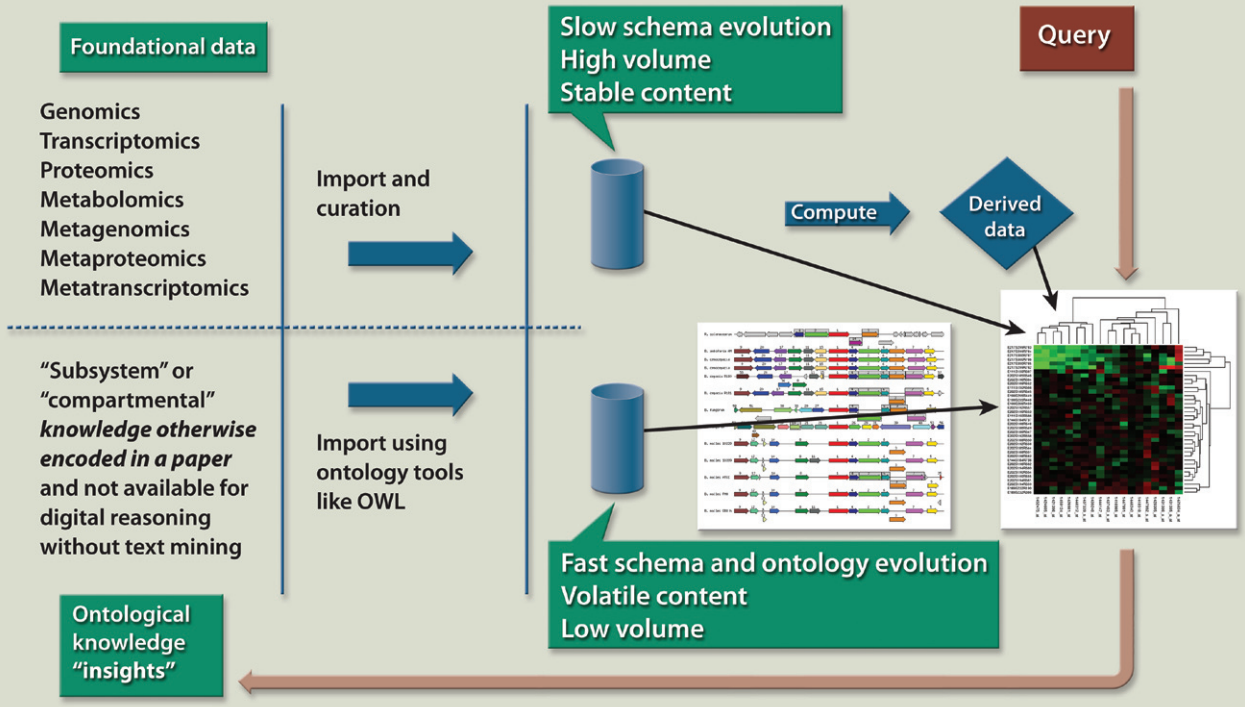


Fig. 4.3. Conceptual Overview of GKB Architecture. Partitioning of data and services between the components sharing a single data model defined in GTL and external heterogeneous data and services.

Findings

- A principal GKB requirement is the need for *separation between high-volume data* (usually from high-throughput experiments) *and low-volume data* (e.g., information on protein structures and transmission electron microscopy images).
- Another important knowledgebase requirement is the need to maintain large bodies of derived data to support queries on vast amounts of information. Good examples of such information are the data needed to rapidly display large sets of chromosomal clusters in prokaryotes; the volume of these data exceeds that of input data by two orders of magnitude.

Recommendations

- For both high-volume and low-volume data types, the GKB should provide capabilities for performing machine reasoning and user-driven queries (e.g., via a simple Web interface). Data volume will constrain storage and query mechanisms for high-volume information to a more limited set of possible implementations (see Fig. 4.3, this page).

Service-Oriented Architectures and Ontologies

Findings

To take advantage of a wide array of Web-based resources such as data stores, visualization environments, and analysis tools, architectures based largely on Web services models—so-called service-oriented architectures (SOAs)—have evolved and are significantly applicable to the GTL Knowledgebase. Driven by massive commercial data stores like

Amazon and Google and by the need to represent and present unusual data types (e.g., multimedia), this architectural trend considers Web content and services as databases. In Web services models, additional logic beyond that which resides in standard database engines is built to access distributed Web resources. Such models have many attractive features for distributed biological data and services and thus offer the potential for biologists to create analysis pipelines that automatically link experimental data to multiple computations, resulting in new insights. One example of such a resource is MeDICi (Middleware for Data-Intensive Computing; Gorton et al. 2008), which represents a workflow tool for biologists based on SOAs.

SOA-based models are not without drawbacks. For example, they are subject to failures of individual Web resources on which queries depend and sometimes are associated with query performance problems in accessing heterogeneous Web resources. However, well-understood design approaches and supporting technologies can address SOA drawbacks and could be leveraged to build successful SOA features into the GTL Knowledgebase.

While many existing biology databases use a simple architecture, the GKB would require a combination of architectures, including SOA for Web services; application programming interfaces (APIs) for data retrieval; database clusters; online analytical processing; and carefully crafted, flexible data models. Current data systems employing this combination or hybrid approach are, for example, the *Shewanella* Knowledge Base and MicrobesOnline, both of which integrate several data resources (see sidebar, *Shewanella* Knowledgebase, p. 24, and Fig. 4.4. Data Types and Resources Integrated by MicrobesOnline, this page).

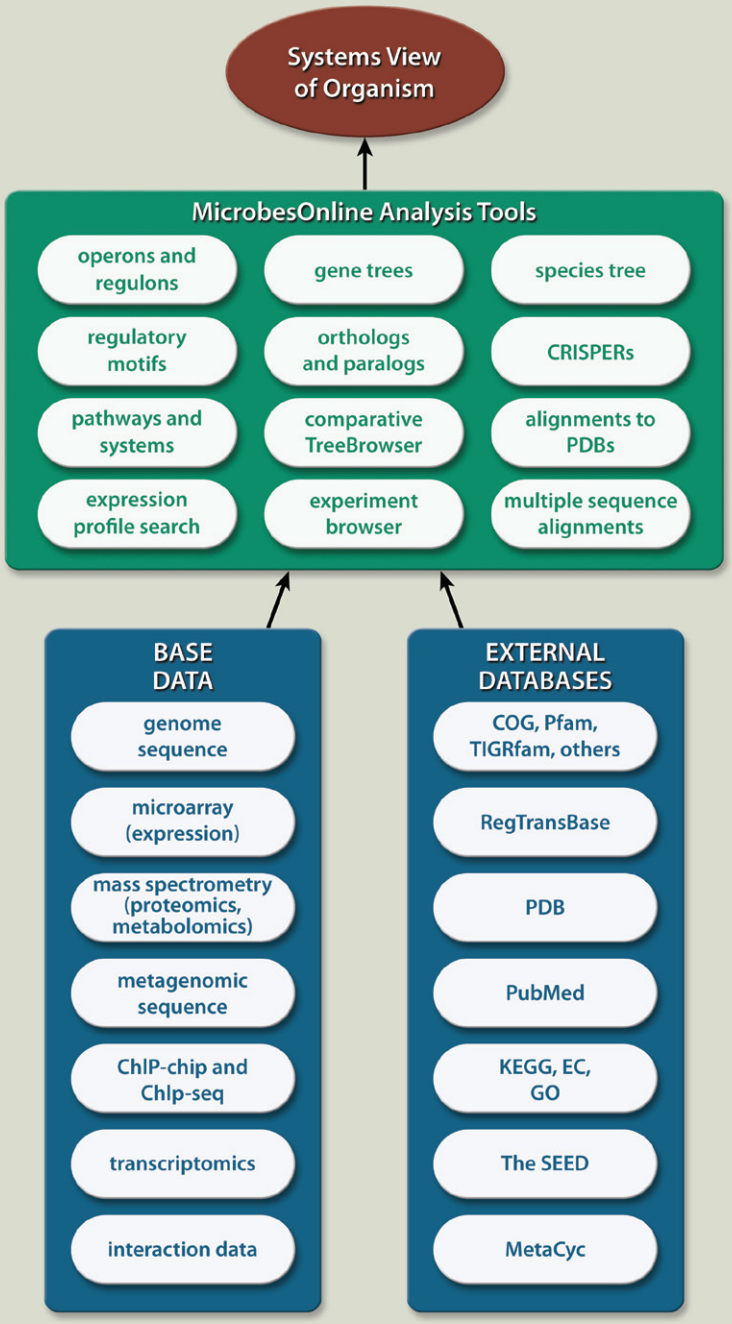


Fig. 4.4. Data Types and Resources Integrated by MicrobesOnline. A hybrid, distributed, and Web services model for data integration and management.

Knowledgebase planners also anticipate that semantic Web technologies can be employed to augment core capabilities of GKB architecture. Such technologies include standard ways for defining Web services using controlled vocabularies (e.g., with UDDI or SOAP) and ontologies for describing data objects (e.g., based on OWL). These semantic Web capabilities make access to distributed knowledgebase services technically easier and more meaningful for researchers. Furthermore, with such technologies, query and retrieval tools can intelligently determine which information and services on the Web have data relevant to a query because knowledge in each domain has been described using a formal ontology. For example, Web resources describe themselves with rich semantics amenable to reasoning by external automated agents, and machines can assume much of the burden of data and

service discovery, engagement, heterogeneous assimilation, and integration. Theoretically, scientists thus can devote their time to more creative decision making in advanced stages of the information-gathering process.

Developing robust, expressive ontologies and using them efficiently are extremely challenging endeavors requiring the collaboration and commitment of computer scientists and highly qualified domain experts. For these activities, only limited advancements have been achieved globally, and such efforts are nascent in the biology community. However, experts have created a strong and ever-expanding family of ontologies. These vocabularies range from environmental markup language (EML), which is highly relevant to numerous emerging high-throughput datasets of environmental sequences, to relatively new protein functional annotation (PRO) ontologies. Based on its pervasiveness among PubMed references, Gene Ontology (GO)—which describes the molecular and cellular functions of protein complexes—is the most successful ontology in biological research today. An ontology achievement related to DOE's Bioenergy Research Centers involves carbohydrates in plant cell walls. Formal description languages have been created for carbohydrate structures (e.g., at the University of Georgia's Complex Carbohydrate Research Center, http://www.ccrc.uga.edu), and these ontologies potentially could be extended to encompass the composition and structure of the entire plant cell wall. This would enable descriptions of analytical results on biomass and detailed queries of observed cell-wall structure and dynamics.

Integrating bioinformatics software with databases will be essential. The Gaggle (Shannon et al. 2006), for example, uses standardized descriptions of data items and Web services to integrate data resources and tools to support biologists' analytical needs (see Fig. 4.5. Communication in the Gaggle, p. 49). This approach represents a small-scale prototype of the kinds of systems the GTL Knowledgebase should incorporate to support data-driven analysis, modeling, and visualization. Other useful concepts include notification services that provide updates to biologists if data relevant to their research change or become available.

Additionally, Web-accessible systems increasingly are handling enormous datasets (e.g., in Earth fly-by or geographic information systems and multimedia). These have given rise to an architectural trend characterized by very large databases having simple schemas and high efficiency. The GTL Knowledgebase inevitably will contain datasets to which such technology can be advantageously applied.

Recommendations

- A GKB architecture study group should be established to analyze options and priorities for designs. The study group should include stakeholders from various GTL-relevant areas of biology, representatives of DOE funding entities, developers of software tools, and the core GKB architecture team. Additionally, an architecture configuration board should be established to oversee system choices and performance.
  - Knowledgebase technical and operational requirements must be analyzed in detail to fully exploit the tremendous opportunities offered by computing, informatics, and communications technologies.
- Well-proven architectural technologies and configurations should be used whenever possible to reduce costs and increase system robustness.

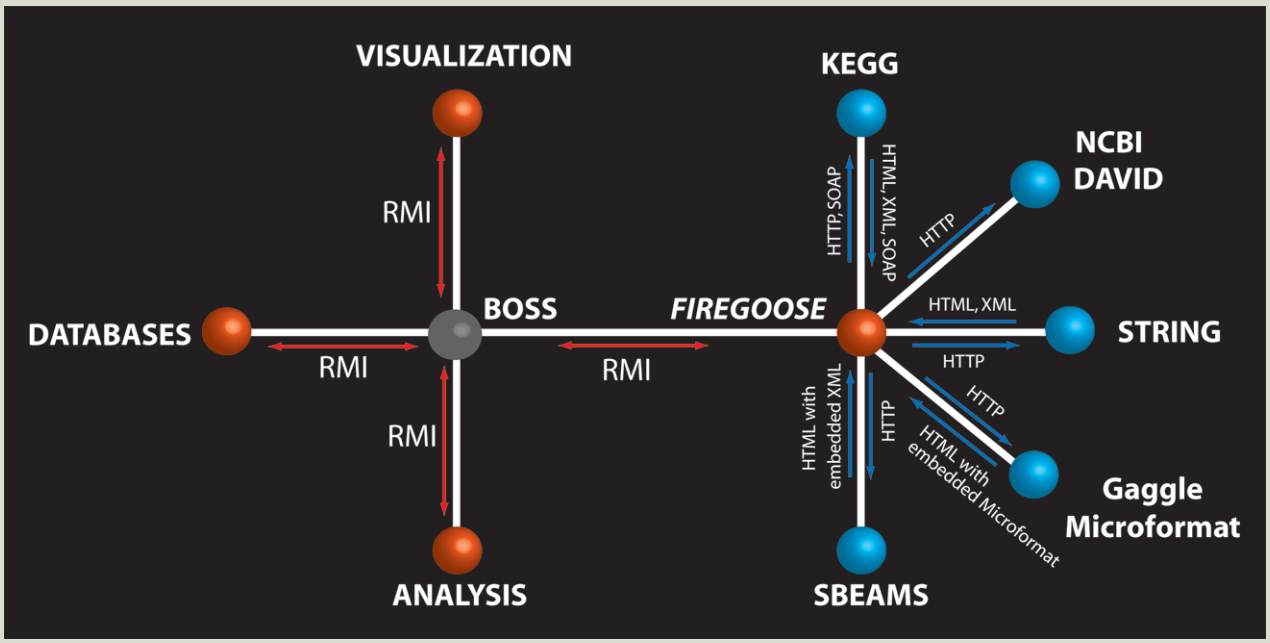


Fig. 4.5. Communication in the Gaggle (http://gaggle.systemsbiology.net/docs/). Software and databases shown as red dots send and receive broadcasts via Java remote method invocation (RMI). The blue nodes are Web resources connected to the Gaggle through the Firegoose and accessed using HTTP with other protocols and formats such as HTML, XML, and SOAP layered over top. Analysis tools in the Gaggle framework include R, MatLab, and MeV; the visualization tools include Cytoscape, BioTapestry, DMV, and Genome Browser. A central strength of Gaggle and Firegoose is the ease with which they can be extended to include third-party tools and databases that have been developed using varied platforms and programming languages. [Source: Adapted from the following two documents: Bare, J. C., et al. 2007. "The Firegoose: Two-Way Integration of Diverse Data from Different Bioinformatics Web Resources with Desktop Applications," *BMC Bioinformatics*, **8**(456). Shannon, P. T., et al. 2006. "The Gaggle: An Open-Source Software System for Integrating Bioinformatics Software and Data Sources," *BMC Bioinformatics* **7**(176).]

- GKB performance, scalability, and latency requirements must be carefully defined and analyzed.
- The GKB should be designed to facilitate cost-effective upgrades associated with anticipated changes in requirements.
- Detailed data requirements—such as rapidly evolving versus stable schemas and large versus small volumes of data—must be defined and the underlying architecture and transport mechanisms built accordingly.

Data Access and Security

Findings

The envisioned GKB would promote the formation of collaborative groups that both informally and formally share data and insights to advance their scientific investigations. Such collaboration is extremely important for integrating analyses of large datasets across multiple groups and allowing sensitive, accurate curation and analysis of data prior to public release. These activities will facilitate the construction of various user interfaces ranging from simple Explorer-type tools to next-generation collaborative tools comparable to contemporary social networking sites such as Facebook.com.

Recommendations

- GKB architecture must implement security policies and practices supporting GTL data and information sharing. These should include procedures for protecting pre-public data for periods specified within architectural guidelines.
- Also, users should be able to incorporate into the GKB additional private data based on their analyses and protect this information using security mechanisms provided by knowledgebase architecture.
- Furthermore, the GKB security model should allow biologists to share their private data with a selected set of collaborators in the GTL community.

GTL Knowledgebase Community and User Issues

For users, the GTL Knowledgebase (GKB) would be an important tool for accelerating discovery and hypothesis-based fundamental research and rapidly translating this research into critical, practical solutions for global climate change, environmental remediation, energy independence, and alternative fuels. The GKB significantly would influence both basic and applied science, ultimately providing a valuable resource in numerous areas of industrial research. Furthermore, the knowledgebase would lead to transformative technologies, serve as a tool for the overall biological research community, and provide a key methodology for transferring emerging knowledge to industry.

The GKB would uniquely assist the GTL and broader research communities by integrating environmental and biological information into a unified system enabling users to extract existing knowledge, formulate hypotheses, create new data networks, and generate models of complex biological systems. Achieving these envisioned capabilities would require the GKB not only to carry out its data archive mission but also to become a working environment for testing hypotheses using shared data.

Without effective access to information, even the most highly integrated, standardized, complete, and correct set of analytical data is unusable by the broader research community. The knowledgebase project would provide such access by serving as a focal point for data sharing and information exchange within the GTL community (see Appendix 1. Information and Data Sharing Policy, p. 59). The GKB would facilitate these exchanges by supporting a wide variety of data types generated by the general research community and then integrating the data into a common framework linking otherwise disparate systems. Thus a significant challenge for the GKB would be to provide a robust public resource that allows researchers to access GTL data in diverse and flexible ways.

The GTL Knowledgebase also should develop and enforce a data sharing policy that both protects individual researchers' data and ensures the broader scientific community has easy and open access to GKB information. This would require the GKB to actively remove obstacles that typically impede data access and to work vigilantly with scientists to monitor and improve the knowledgebase over time. In addition to initial development of this unprecedented system, the GKB—through outreach, training, and survey-based performance evaluations—would need to continuously assess the critical DOE- and GTL-relevant data needs of researchers and the knowledgebase capabilities for providing them.

Knowledgebase User Community

Findings

An advantage of the GTL Knowledgebase would be having a manageable community of potential users showing early interest in the GKB. The following are general classifications of these target user groups:

- **Data users.** Knowledgebase design should enable users to easily and quickly find data and tools relevant to their research. Such access to GKB data could be achieved

through Web interfaces, bulk downloads, and machine use via Web services. GKB features beneficial to researchers include user feedback and ratings showing new users which knowledgebase tools other scientists have found useful. In addition to data, the GKB should provide users with estimates for errors in data and with information about the experiments from which data were derived (e.g., the parameters used on equipment, scientific questions data producers sought to answer with the data, replicate information, and the methods employed). Data users would interact with GKB information in various ways. Some researchers might be interested in qualitative information (e.g., binary data such as presence or absence and protein localization); others might wish to conduct more quantitative analyses (e.g., exploring gene expression levels across different conditions).

- **Data producers.** The GTL Knowledgebase should provide a mechanism enabling data generators to easily deposit their data and metadata into the GKB. This might require the knowledgebase to accept multiple data formats. In addition, data producers may need embargoes allowing them to incorporate their data into the GKB for queries prior to publication, but this feature must be weighed as part of overall operational requirements. Furthermore, producers would like the ability to determine how their data are used. The GKB could provide this service by allowing users to track publications resulting from the use of knowledgebase data. Encouraging GKB participation among producers likely would require offering clear benefits for data deposition, such as GKB graphics and tools for creating publication-ready figures (as in Cytobank) and the ability to use knowledgebase resources and software for analyzing new datasets prior to publication. Moreover, straightforward data analysis tools would serve as incentives for data producers to become GKB users.
- **Software and tool developers.** Developers of software and tools integrated into or connected to the GKB would like to track and evaluate the usefulness of their tools (e.g., via feedback from other users) and the publications resulting from the application of such resources.
- **Data producers and users.** Many GKB users also would be data producers. As they deposit data, models, and tools into the GKB, these users would use resources others have added. Systems biologists, for example, probably would be both producers and users of data.
- **Industry.** Researchers representing the industrial community would take advantage of GKB data and tools to use research results to make significant contributions to DOE mission areas (e.g., climate change research, bioenergy, and carbon cycling).
- **DOE program staff.** DOE program staff likely would be interested in knowing which data and tools were available in the knowledgebase. Participation from this group of users could facilitate evaluation of data-release policies and provide an incentive for data producers and software and tool developers to contribute to the GKB.

Recommendations

- To optimize effectiveness and service to constituents, GKB planners should clearly determine members of the research community who would use the knowledgebase and the science and technology groups who will develop it.
  - Knowledgebase developers could identify the GKB user community by evaluating and pursuing several sources, including current and former GTL grants, publications, conference attendees, DOE phone directories, target institutions (e.g., DOE's Bioenergy Research Centers, JGI, and national laboratories), and Google analytics.

- Contact lists for probable GKB user groups should be developed.
  - These files can be assembled from several sources, including conference abstracts, author lists from publications representing various disciplines, DOE personnel databases, principal investigators awarded DOE research grants, and scientists who contact the GKB website.
- During design and operation phases, the GTL Knowledgebase should survey individuals from each target user group to measure their requirements for new functionality and GKB adequacy.
- User groups associated with large GTL efforts such as centers and confederations should be enlisted to help develop GKB.

[Source: DOE Joint Genome Institute.]

Knowledgebase Interfaces and Portals

Findings

- All knowledgebase constituents—including data producers and users, tool developers, industry representatives, and DOE staff—would access and interact with GKB through several interfaces. These could include the Web, file transfer protocol (FTP) servers, application programming interface (API), Web services, wikis, videos, online tutorials, and software (e.g., MeV, Quackenbush, R/biocurator, and Cytoscape). Knowledgebase interfaces would be nested in various layers such as an entry Web portal leading to GKB subpages with links to analysis tools and FTP servers for data deposition. Organization of these interfaces and portals should be fairly intuitive to meet the needs of both expert and novice users. Furthermore, users accessing GKB data might generate new types of data or refine existing information, facilitating the iterative process of knowledgebase data improvement.

Recommendation

- The GTL Knowledgebase should engage the help of identified user groups, not only to ensure these communities have effective and easy access to the GKB, but also to constantly assess knowledgebase performance in supplying needed and facile services.

Knowledgebase Outreach

Findings

- Outreach activities are excellent not only for fostering knowledgebase awareness, but also for obtaining feedback on the quality and efficacy of GKB resources. Moreover, these activities provide mechanisms for users to be involved in system development. Basic modes of outreach include presentations at scientific conferences, articles in scientific journals, engagement of researchers in individual laboratories, electronically based announcements (e.g., wikis and mailing lists), multiday workshops at conferences, Web-based instruction (e.g., tutorials, webinars, and bulletin boards), and internships for visiting postdoctoral researchers or students. The GKB also would need to target specific areas for focused outreach.

Recommendations

- Significant GKB resources should be dedicated for outreach and mobilization of the scientific community.
- The GKB should engage user groups to ensure effective, easy access and assess knowledgebase performance.
- GKB developers should engage in active communication with the scientific community to obtain input for GKB development and promote widespread and ongoing knowledgebase use. GKB developers can achieve this through the following activities:
  - Seconding staff within large facilities such as DOE's Bioenergy Research Centers and JGI, and using center personnel with expertise in data management.
  - Rotating GKB staff into these centers to develop user expertise.
  - Cross-training graduate students and postdocs by offering GKB training programs in which researchers can interact with developers onsite.
  - Conducting exercises in which knowledgebase developers try to recreate figures from publications using GKB data and methods to anticipate how users might access and analyze data.
  - Providing online and active support for first-time users and new GKB analysis tools and randomly surveying users for detailed feedback.

Knowledgebase Training

Finding

- The sophistication, breadth, depth, and range of GKB capabilities would far exceed previous data resources for probable user communities. Thus users would require training and education to fully understand and exploit GKB tools, resources, and opportunities for advancing their research efforts.

Recommendations

- Pursuant to its full development and operations, the GKB should provide a comprehensive training program to user communities. Training opportunities could include 2- to 4-day modular onsite courses, half- or full-day workshops at relevant conferences, and monthly online sessions and seminars.
- Knowledgebase staff should establish a help-desk service to answer user questions.
- A user-friendly Web resource should be created and include the following features:
  - Extensive documentation of GKB services and capabilities, including quality assurance (QA) methods.
  - Detailed descriptions of standard operating procedures (SOPs) used to generate or analyze data.
  - Information on file format standards and content.
  - Descriptions of data structures.
  - Details on controlled vocabularies used by the research community.
  - Step-by-step directions for using GKB analysis tools.
  - Instructions for downloading data.

Knowledgebase Performance Assessment

Findings

- Continual assessment of GKB performance and definition of system metrics would be essential to the success of the knowledgebase. These tasks could be accomplished by conducting surveys of targeted user groups to gain feedback on GTL community resources. Such feedback would drive further development of GKB resources. Survey responses also would be used to evaluate knowledgebase enhancements as they are released and tested on each target group.
- Moreover, results of user surveys could be used to establish performance measures that would be incorporated into reports for knowledgebase staff, DOE project officers, and other individuals involved in GKB governance. These performance metrics would be the foundations for enhancing many GKB activities. Staff meetings, project plans, and other activities could be optimized to improve metrics over time.
- An iterative process of GKB releases and user feedback would enable the knowledgebase to continue to meet the data and analysis needs of each user community. Furthermore, GKB's Web presence and resources—including a help-desk email address and tools to track system bugs and improvements—would provide ample opportunities for users to offer feedback about the knowledgebase.

Recommendations

- GKB staff should serve as resources and information providers to user communities.
  - Staff presentations and workshops on GKB capabilities would create opportunities for user training and establish further contacts within targeted GKB constituents.
- The GKB user community should be surveyed and the results translated into performance measures and prioritization schemes for ongoing knowledgebase development.

Knowledgebase Data Sharing and Policy Development—Incentives for Depositing Data

Findings

- Universal, straightforward, and productive use of the GKB by the scientific community would require various incentives. In particular, alerting the GKB user community to newly contributed data or tools would provide a significant incentive for participation in the knowledgebase. The GKB could promote such resources by (1) sending users periodic emails about recently submitted data, models, and tools; (2) tracking publications arising from the use of GKB services and emphasizing knowledgebase capabilities and data most cited in scientific literature; (3) showcasing new GKB datasets and tools online on the system's homepage; (4) providing strong graphics and analysis tools to improve and facilitate the publication process; and (5) eliciting user ratings of GKB data and tools (e.g., as Amazon does).
- Researchers also would be encouraged to contribute to the GKB if DOE program managers were able to track data and tools deposited by agency-funded principal investigators. Such a capability would promote knowledgebase participation if, for

example, DOE research grants require release of data and tools to the GKB. As an additional incentive for contributing to the GKB, data and resource providers could be offered early access to newly developed tools (e.g., beta testing), as well as GKB help in maintaining and backing up contributors' data.

Recommendations

- Transparent methods, incentives, requirements, and other inducements should be established to encourage researchers to contribute to and use the GTL Knowledgebase.
- To encourage submission of data and methods into the GKB, knowledgebase designers, along with GTL research programs and their managers, should make this process easy for data producers and tool developers by establishing and supporting standard data formats for different data types.
- To further facilitate knowledgebase use, GKB should develop and provide support for software, processes, and protocols that easily convert data into formats acceptable for knowledgebase entry.

Knowledgebase Working Group

Findings

Envisioned for the GTL Knowledgebase is the complete integration of multiple partners with unique capabilities and science objectives. Successfully integrating these partnerships across DOE mission–focused laboratories, universities, and industry would require a GKB management and governance model combining flexibility and accountability. The GKB management plan likely would focus on organizational structure and reporting mechanisms, operations management, and community assessment and review (see Box 5.1, Elements of the GKB Management Plan, p. 57).

- **Organizational structure and reporting mechanisms.** Critical to GKB success would be a well-defined management team including a director, management staff, and scientific and governance boards. The director's responsibilities would include executing the overall vision of the project and overseeing project direction. The GKB director would be accountable to a governance board for ensuring timely achievement of project goals and milestones. The governance board also would be expected to manage internal and external GKB reviews. In turn, knowledgebase stakeholders in federal agencies such as DOE would oversee and receive reports from the GKB governance board and director. GKB viability also would require a scientific board that would provide vision and recommendations for new directions for the knowledgebase project.
- **Operations management.** The GKB management plan would detail the roles and responsibilities of all senior management staff, which could include a deputy director as well as scientific and technology lead directors. Operational management plans would describe challenges and opportunities associated with coordinating GKB efforts across possible distinctive sites, national laboratories, and academic settings.
- **Community assessment and review.** GKB operating procedures and the system's effectiveness in supporting scientific research would be reviewed every 3 years at minimum. Integral to the review process would be community assessment of the

Box 5.1

Elements of the GKB Management Plan

Governance models define the relationships among various aspects of a program or organization and among key personnel and operations. For the GKB, the governance model should clearly describe the functioning of communications, submission and exchange of knowledgebase information, and establishment of policies, procedures, and personnel responsible for decision making.

- **Roles and responsibilities.** The GKB management plan should define the roles of institutions, programs, and individuals and the commensurate responsibilities of each.
- **Authorities and accountabilities.** The implementation plan should ensure that each role has the requisite authority to bring needed resources to bear and carry out their functions. Accountabilities define responsibilities between parties.
- **Policies, standards, and processes.** The GKB should enlist the input of all GKB stakeholders in defining knowledgebase needs and discussing system solutions for establishing policies and standards.
- **QC and QA protocols.** The GKB's ultimate value to the research community would depend on the degree to which users could rely on the accuracy, fidelity, and completeness of knowledgebase datasets and the tools to use them. Essential to user confidence in GKB resources, therefore, would be the establishment of quality control and quality assurance protocols.
- **Resources and funding.** The GKB would rely on resources and funding from a wide variety of sources. DOE's Genomics:GTL program would provide resources focused on knowledgebase establishment, operations, and maintenance. GTL research programs and centers—each of which would contribute significantly to data management and use—would ensure that research results are rigorously incorporated into the knowledgebase and that GKB-related research directions and priorities are well defined and supported.
- **Program management and staff.** GKB management and staff—in consultation with DOE's Office of Biological and Environmental Research (OBER) and other advisers—would facilitate quality in the research conducted using the knowledgebase. OBER's program management staff would define and approve GKB policies and processes (including the governance model) and would oversee implementation of the GTL Knowledgebase project.

GKB, which would critically depend on the management staff's outreach to the biological community. Also important to knowledgebase success would be project leaders who proactively define emerging needs for GKB-relevant policies on data and information sharing. Developing such polices would require establishing a systematic method for assessing current data and information guidelines, including those adopted by the GTL program. Although all external programmatic reviews would be coordinated with the GKB governance board and director, DOE would have the responsibility of ensuring and formulating the success of such reviews.

Recommendation

- A scientific working group should be formed to work closely with the GKB project throughout the development process. This group should include the following: (1) representatives from all user communities; (2) data producers generating various types of information (e.g., environmental, proteomic, genomic, and transcriptomic); (3) researchers focused on different GTL missions (e.g., bioenergy, carbon cycling, systems biology, and environmental remediation); and (4) experts in computing, informatics, and communications technologies and systems.

Appendix 1

Information and Data Sharing Policy*

Genomics:GTL Program

Office of Biological and Environmental Research
Office of Science
Department of Energy

Final Date: April 4, 2008

Introduction

Experimental biology has evolved in the past 20 years to include a rapid-access, global scientific community hyperconnected through the Internet. The changing scope of scientific inquiry and the astonishing rate of data production drive the development of a new type of cyberinfrastructure, which in turn has promoted the formation of e-science (1). Journals, funding agencies and governments correspondingly have developed information standards and sharing policies, all of which in one way or another address research conducted in an open-access environment. A key hallmark of these policies is the requirement that scientific inquiry and publication must include the submission of publication relevant information and materials to public repositories. For the most part, the policies follow the uniform principle for sharing integral data and materials expeditiously (called UPSIDE) (2). Conversely, when research information is not made publicly available to the global scientific community, a corresponding price is paid in lost opportunities, barriers to innovation and collaboration, and the obvious problem of unknowing repetition of similar work (3).

This statement summarizes the information and data-sharing policy within the Genomics:GTL (GTL) program at the Department of Energy's Office of Biological and Environmental Research (OBER). OBER recognizes that successful implementation of this policy will require the development of new technologies such as software tools and database architectures and will be funded, as necessary, from the GTL program subject to funding availability. We affirm our support for the concept of information and data standards and sharing and we believe that a comprehensive policy can be constructed that will encourage GTL researchers to exchange new ideas, data and technologies across the GTL program and the wider scientific community.

Research information obtained through public funding is a public trust. As such, this information must be publicly accessible. The GTL information-sharing policy requires that all publication related information and materials be made available in a timely manner. All Principal Investigators (PIs) within the GTL program will be required to construct and implement an Information and Data-Sharing Plan that ensures this accessibility as a component of their funded projects.

---

*From http://genomicsgtl.energy.gov/datasharing/.

Policy Statement

The Office of Biological and Environmental Research (OBER) will require that all publishable information resulting from GTL funded research must conform to community recognized standard formats when they exist, be clearly attributable, and be deposited within a community recognized public database(s) appropriate for the research conducted. Furthermore, all experimental data obtained as a result of GTL funded research must be kept in an archive maintained by the Principal Investigator (PI) for the duration of the funded project. Any publications resulting from the use of shared experimental data must accurately acknowledge the original source or provider of the attributable data. The publication of information resulting from GTL funded research must be consistent with the Intellectual Property provisions of the contract under which the publishable information was produced.

I. Applicability

This policy shall apply to all projects receiving funding in the Genomics:GTL program as of October 1, 2008. For cases where information sharing standards or databases do not yet exist, the information sharing and data archiving plan provided by a project's PI must state these limitations. Data and information that are necessary elements of protected intellectual property and related to a pending or future patent application are explicitly exempt from public access until completion of the patenting process. Adherence to this policy will be monitored through the established procedure of yearly progress reports submitted to GTL program managers. All information regarding data shared by GTL funded research projects will be made publicly available at http://genomicsgtl.energy.gov/datasharing/.

II. Submission of Information and Data

All investigators are expected to submit their publication related information to a national or international public repository, when one exists, according to the repository's established standards for content and timeliness but no later than 3 months after publication. This includes:

- Experimental protocols,
- Raw and/or processed data, as required by the repository,
- Other relevant supporting materials.

OBER will maintain a website listing all published peer reviewed papers and published patents resulting from GTL funding, and PIs are expected to inform OBER on a regular basis when a publication appears in print. OBER is encouraged by the development of the National Institutes of Health open-access policy and, when possible, OBER will link to open-access GTL funded publications. PIs, however, are encouraged to publish in journals appropriate to their fields of research. OBER recognizes that subdisciplines and experimental technologies have varying degrees of cyberinfrastructure and standard ontology to accommodate this policy. Specific guidelines and suggestions for GTL investigators are provided below.

II.A. Nationally and Internationally-Accepted Databases and Ontologies

II.A.1. Sequence Data

The field of genomic sequencing has a very well developed mechanism for public archiving of experimental data. Nucleotide sequence data will be deposited into GenBank, and protein sequence data will be deposited into the UniProtkb/Swiss-Prot Protein Knowledge database. Investigators should report to OBER the sequence identifier including the accession number and version. In addition, investigators are encouraged to use the gene ontology annotation database (4) when possible, and OBER applauds the work of the Genomic Standards Consortium (GSC) in the development of minimum information about a genome sequence standards (MIGS).

Specifically for large-scale GTL sequencing projects, OBER will adopt the policy that whole genome sequencing data, where genome completion is the stated goal, must be made publicly available 3 months after first assembly of the sequencing reads for that genome. In the case of metagenomic sequencing, data must be deposited to the National Center for Biotechnology Information (NCBI) 3 months after completion of the last sequencing run, which must be specified in the Joint Genome Institute User Agreement. For other types of sequencing experiments, such as expressed sequence tags (ESTs), the data will fall under the guidelines for publication of relevant information and shall be deposited to NCBI 3 months after publication.

II.A.2. Three-Dimensional Structural Data

All coordinates and related information for structures of biological macromolecules and complexes are to be deposited in the Protein Data Bank (PDB) or Nucleic Acid Database (NDB) as appropriate. Accession codes are to be reported back to OBER.

II.A.3. Microarray and Gene Expression Data

The Microarray and Gene Expression Data (MGED) Society recommends the use of a MGED ontology for describing key experimental conditions as, for example, using a MIAME-compliant format (MIAME, Minimum Information About a Micro-array Experiment). OBER's policy will follow the MGED recommended ontology. We further strongly encourage GTL researchers to deposit raw and transformed data sets and experimental protocols to a public microarray database and report back to OBER the accession number and URL. Possible microarray databases for data deposition include the Gene Expression Omnibus (5), ArrayExpress (6), and the Stanford Microarray Database (7).

II.B. Information Sharing Systems and Databases Under Development

II.B.1. Proteomics

The Proteomics Standards Initiative (PSI), a working group of the Human Proteome Organization (HUPO), recently outlined two standard proteomics ontologies: minimum information about a proteomics experiment (MIAPE) (8) and minimum information required for reporting a molecular interaction experiment (MIMIx) (9). Because this is an evolving initiative and the field is still immature, we cautiously encourage GTL

proteomics researchers to adopt the use of MIAPE and MIMIx in their research. We are further encouraged by the development of public proteomics repository databases such as the Open Proteomic Database (10) and PEDRo (Proteome Experimental Data Repository) (11) and encourage GTL researchers to engage with these databases. However, we recognize that standards and ontologies will evolve within the proteomics community, and GTL's policy will follow guidelines set forth by HUPO as they develop.

II.B.2. Other Technologies

GTL research makes use of a large variety of technologies for which there are, as yet, no national or international information standards and archival formats. Scientists in the GTL program are encouraged to participate in the efforts of research communities to develop such standards for enabling information sharing.

GTL's long-term objective is to encourage the development of infrastructure for technologies that do not yet have nationally or internationally accepted information sharing standards. In cases where there are no public repositories or community driven standard ontologies, OBER recommends that these types of data and information be made publicly available by the PI.

III. Protection of Human Subjects

Research using human subjects provides important scientific benefits, but these benefits never outweigh the need to protect individual rights and interests. OBER will require that grantees and contractors follow DOE principles and regulations for the protection of human subjects involved in DOE research. Minimally this will require an IRB review. These principles are stated clearly in the Policy and Order documents: DOE P 443.1A and DOE O 443.1A, which are available online at http://www.directives.doe.gov.

IV. Systems Biology and the GTL Knowledgebase

A long-term vision for the Genomics:GTL program, as outlined in the 2005 roadmap, is an integrated computational environment for GTL systems biology (12). OBER affirms our support for the development of an integrated framework to provide for data sharing, modeling, integration, and collaborations across the program. OBER also recognizes that continued support for development of community driven standard ontologies and data-sharing policies is inherent to the successful implementation of a systems biology network.

V. Computational Software

The International Society for Computational Biology (ISCB) recommends that funding agencies follow ISCB guidelines for open-source software at a "Level 0" availability. ISCB states that research software will be made available free of charge, in binary form, on an "as is" basis for non-commercial use and without providing software users the right to redistribute. OBER will follow ISCB recommendations at a Level 0 availability. OBER recommends that research software developed with GTL funding that results in a peer-reviewed software publication be made accessible through either an open-source license (http://www.opensource.org) or deposited to an open-source software community such as SourceForge.

VI. Laboratory Information Management Systems (LIMS) for Data Management and Archiving

GTL systems biology research projects involve high-throughput, data intensive research that necessitates use of a data management system to automatically handle this pipeline of data. OBER's goal is that researchers within the GTL program utilize a LIMS system for managing their research data and information. Because different research agendas require different information management systems, an overarching and restrictive policy could place an undue burden on PIs. Therefore, we expect that research projects that involve more than one senior investigator will be required to implement a LIMS or a similar type of electronic system for data and information archiving and retrieval. This plan should balance the clear value of data availability and sharing against the cost and effort of archive construction and maintenance.

VII. Summary

This document outlines the Genomics:GTL program policy and will require GTL funded principal investigators to construct an information and data-sharing plan as a component of their projects. The policy requires information to conform to existing community recognized standard formats wherever possible, to be clearly attributable, and to be deposited, in a timely manner, within a community recognized public database(s) appropriate for the research conducted. OBER is committed to encouraging development of public repositories and standard ontologies for the GTL research community. OBER recognizes that this policy necessarily will be revised to include new standards, data types, and other advances that are pertinent to maximizing availability of data and information across the GTL program. This information and data-sharing policy and related materials can be found at http://genomicsgtl.energy.gov/datasharing/.

References

1. E-science refers to large-scale science distributed through global collaborations and enabled by the Internet (see http://www.rcuk.ac.uk/escience/default.htm).
2. National Research Council, *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*, The National Academies Press, Washington, D.C., 2003.
3. Uhlir, P. F., and P. Schröder, "Open Data for Global Science," *Data Science Journal* **6,** OD36–53 (2007).
4. Camon, E., et al., "The Gene Ontology Annotation (GOA) Database: Sharing Knowledge in Uniprot with Gene Ontology," *Nucleic Acids Research* **32,** D262–66 (2004).
5. Edgar, R., M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository," *Nucleic Acids Research* **30,** 207–10 (2002).
6. Brazma, A., et al., "ArrayExpress—A Public Repository for Microarray Gene Expression Data at the EBI," *Nucleic Acids Research* **31**, 68–71 (2003).
7. Sherlock, G., et al., "The Stanford Microarray Database," *Nucleic Acids Research* **29,** 152–55 (2001).

8. Taylor, C. F., et al., "The Minimum Information about a Proteomics Experiment (MIAPE)," *Nature Biotechnology* **25,** 887–93 (2007).

9. Orchard, S., et al., "The Minimum Information Required for Reporting a Molecular Interaction Experiment (MIMIx)," *Nature Biotechnology* **25,** 894–98 (2007).

10. Prince, J. T., et al., "The Need for a Public Proteomics Repository," *Nature Biotechnology* **22,** 471–72 (2004).

11. Garwood, K., et al., "PEDRo: A Database for Storing, Searching and Disseminating Experimental Proteomics Data," *BMC Genomics* **5,** 68 (2004).

12. U.S. Department of Energy Office of Science Office of Biological and Environmental Research. *Genomics:GTL Roadmap: Systems Biology for Energy and Environment,* 2005 (see http://genomicsgtl.energy.gov/roadmap/index.shtml).

Appendix 2
Use Case Scenarios of Systems Biology Investigations Utilizing the GTL Knowledgebase

The workshop systems biology group identified a set of high-priority scientific and engineering research examples based on the analysis of priority DOE applications, unmet needs, and feasibility. Each of these use case scenarios was selected to match projected phases of GTL Knowledgebase (GKB) development.

Use Case Scenario 1

Use Case 1 has two objectives:

- Support the capability to rapidly assess the metabolic potential and regulatory features of any culturable or sequenced prokaryote of primary importance or relevance for all DOE GTL focus areas.
- Map parts (genes) and modules (pathways, subsystems, and regulons) that constitute the core of life across thousands of diverse species within Use Cases 1 and 3.

Identification and accurate functional assignment of genes involved in the key cellular processes of any organism with a completely sequenced genome would allow assessment of the organism's metabolic and regulatory capabilities with respect to their applications. This information would provide a foundation for further detailed reconstruction and modeling and allow assessment of the organism's role and interactions within the community. Although many components of the required workflow (including a substantial body of annotated genomes and tools) already exist in the public domain, a considerable effort would be required to automate and scale up the process and, at the same time, maintain and improve coverage, quality, and consistency of annotations.

Understanding and integrating thousands of diverse genomes and the associated nongenomic information—inferences (gene annotations, subsystems and pathways, and regulons) within a framework (tools)—are critical for their assessment and comparative analysis. Although the microbial sequencing projects throughout the world have created a rich, diverse collection of microbial genomes, strong biases are evident in what has been sequenced thus far. Use Case 1 would be an extension of ongoing work seeking to understand related species, as outlined in the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project, which is aimed at systematically filling in sequencing gaps along the bacterial and archaeal branches of the tree of life. Sequencing large numbers of diverse microbial genomes has been a mission focus of DOE for several years. A comprehensive goal would be to sequence a representative of every species discussed in Berghey's manual. The GEBA project is a piloted, modest sequencing effort along this line. This project represents a new paradigm in using the tree of life as a guide to sequencing the target selection. Supporting this level of analysis, which we term "draft reconstruction" or stage 1 analysis, for thousands of diverse genomes—regardless of the current perception of their immediate importance for DOE applications—is crucial for creating several resources:

- Rich pool of diverse metabolic and other features for future applications, not all of which can be foreseen (e.g., new engineering needs and synthetic biology).
- Comparative framework for comprehensive and accurate functional annotation of application-related organisms, including complex eukaryotes.
- Reference set of sequences for metagenomic data analysis.

Issues and Requirements

The analysis under consideration would be coordinated with ongoing efforts in gene sequencing and annotation at the DOE Joint Genome Institute (JGI). The assessment level would include the following critical aspects, which require robust informatics support either not provided or only partially covered by existing efforts.

- Accurate gene-function assignments (annotations) supported by various lines of evidence, including homology-based projection from experimentally characterized genes; delineation of structural domains and conserved motifs; genomic context (conserved operons, regulons, and phylogenetic occurrence); phylogenetic profiling; and functional context (role in a relevant pathway, subsystem, or complex).
- Functional predictions for previously uncharacterized gene families. For example, gene candidates would be proposed to fill in gaps in known pathways. A critical investigation into experimental testing of functional inferences also must be considered high priority.
- A controlled vocabulary (for function definition) and connection with a collection of analyzed reactions and metabolites for consistent propagation of annotations and their further use for reconstruction and modeling.
- Curation of gene annotations and pathways.

The genomic reconstruction of regulons (for a recent review and many examples of references therein, see Rodionov 2007) includes identification and capturing of DNA and RNA regulatory motifs (e.g., promoters, operators, attenuators, and riboswitches), along with respective regulatory factors (e.g., transcription factors, regulated genes, and effectors).

Transcriptomic and proteomic data that would become available for some organisms should be minimally analyzed to extract information about which genes are expressed (and under what conditions), which proteins are produced, and protein features such as localization and post-translational modifications.

Typical uses of a knowledgebase for supporting rapid assessment of metabolic and regulatory features include the following:

- **Target:** Organism (or group of related organisms) with completely sequenced genomes. **Use:** Reconstruct a chosen aspect of metabolism and infer some phenotypic properties for further testing. Example: Reconstruct carbohydrate utilization machinery in *Shewanella* spp., including prediction and verification of novel genes, pathways, and physiological properties.
- **Target:** Group of related organisms with completely sequenced genomes. **Use:** Reconstruct a substantial fraction of metabolic regulons.
- **Target:** Set of desired metabolic and regulatory properties important for a certain class of applications (e.g., in bioenergy). **Use:** Identify a group of optimal candidate species.

- **Target:** Defined metabolic function (e.g., enzymatic reaction). **Use:** Identify known and putative candidate genes performing this and other related functions in all organisms, including supportive evidence (experimental, homologous, genomic, and functional context).
- **Target:** Selected organism. **Use:** Automatically compute a list of proteins (enzymes and transporters), inferred reactions, and metabolites for use as a first step to building a metabolic model.
- **Target:** List of genes (proteins) in a target organism. **Use:** Provide all associated information and features including functional assignments from various sources and evidence; association with protein families (phylogenetic profiles); multiple alignments and phylogenetic trees for each family; domains, motifs, and structural features (known or predicted); genomic context (operons and regulons); functional context (associated pathways and subsystems); gene expression data (chosen from integrated or uploaded datasets); proteomic data; associated reactions, metabolites; and other types of data connecting to specific genes.
- **Target:** Experimental "omic" data (e.g., gene expression). **Use**: Identify clusters (lists) of functionally coupled genes (e.g., stimulons), retrieve their properties as described above, and support a detailed correlational analysis (e.g., assess correlations between gene expression and pathways or gene expression and protein levels).

Use Case Scenario 2

Use Case 2 has one objective:

- Support the capability to predict and simulate microbial behavior and response to changing environmental or process-related conditions for target sets of prokaryotic species.

Implementation of this objective would require all components employed in Use Case 1, but additional data would be needed for detailed predictive modeling. These capabilities would be acquired and integrated to allow for the development of computational models, so all relevant data and inferences would be provided in a format for modeling. Among required capabilities of this use case would be the framework and tools that support iterative improvement of models through comparison of model predictions and experimental data. Approaches would be needed to identify inconsistencies between model predictions and experimental observations and to automatically generate hypotheses that would resolve inconsistencies through novel components or component interactions. An example case would be to provide guidelines for organism engineering and even de novo design for tasks both fundamental (e.g., proof of understanding and novel model systems) and applied (e.g., metabolic engineering for biofuel production).

Issues and Requirements

I. Detailed Reconstruction and Modeling of Metabolic Networks

Deliverables described above in Use Case 1 would provide a foundation (i.e., "draft reconstruction") from which detailed models for any species or groups of species selected for specific applications can be developed. The workflow in this case would require additional manual effort aimed to fill in gaps in the metabolic reaction networks, reconcile inconsistencies, and account for published legacy data and accumulated additional (omic

and phenotypic) data; iterative gap filling; model simulations; experimental validation and refinement of the model; and final application to address scientific and engineering objectives. Such detailed analysis would be a focused effort applied to dozens rather than thousands of species. It would proceed in concert with the experimental work, including systematic generation of large volumes of postgenomic data.

Typical uses of a knowledgebase to predict and simulate microbial behavior and response to changing environmental or process-related conditions include the following:

- Quantitative assessment of metabolic capabilities (e.g., production yield or metabolic flux distribution in the network) for a selected organism as a function of environmental and growth conditions.
- Support of optimization tasks (e.g., fermentation conditions optimized for production).
- Guidelines for rational engineering of industrial organisms, including assessment of the potential for directed evolution.

Scope and Requirements

This analysis, which already has been successfully prototyped, would require a level and type of informatics support not currently provided by existing public domain resources. It would aim to cover the following:

- Detailed metabolic and genetic network definition (genes, proteins, roles, reactions, and compounds).
- Stoichiometric matrices and sets of constraints for modeling.
- Predictive computational models and modeling tools.
- Inferred fluxes, phenotypes, growth and application-related properties.
- Experimental validation of models.

Example Uses Employing Complex Queries from the Integrated GKB

- **Target:** Organism with a completely sequenced genome, draft reconstruction (Use Case 1), and collection of additional data (biomass composition, medium composition, and growth characteristics). **Use:** Build a detailed and consistent metabolic reconstruction.
- **Target:** Detailed metabolic reconstruction. **Use:** Apply modeling tools (e.g., flux-balance analysis) to test whether the model is consistent with known physiological properties and growth characteristics; refine the model.
- **Target:** Validated metabolic model. **Use:** Address application and optimization tasks. For example, estimate the maximal yield of the desired product or optimize the growth medium.
- **Target:** Validated metabolic model. **Use:** Predict which genes are essential and dispensable under given growth conditions (e.g., as a way to test the model or to support engineering goals).
- **Target:** Validated metabolic models of two or more organisms. **Use:** Compare their metabolic capabilities with respect to a desired application.
- **Target:** Validated metabolic model. **Use:** Suggest re-engineering strategy (e.g., gene elimination, addition, deregulation, or amplification) to improve organism properties with respect to application.

II. Integrative Modeling of Transcriptional Regulatory Networks

Rationale (modified from Bonneau, Baliga, et al. 2007)

Rapid DNA sequencing technology has provided access to a large number of complete genome sequences from diverse and often poorly characterized organisms. Use of this information is expected to help engineer new biotechnological solutions to diverse problems spanning bioenergy and environmental remediation. In principle, re-engineering new processes by selectively combining otherwise distinct biochemical capabilities encoded in different genomes is a reasonable expectation. In reality, however, this will be possible only when we have a sophisticated understanding of how RNAs and proteins encoded in each individual genome dynamically assemble into biological circuits through interactions with the environment. Given that more than 500 genomes already have been sequenced and that little biological information exists for most of these organisms, a classical gene-by-gene approach is inefficient. Furthermore, since every organism is unique, it is impractical to rely on accumulated sets of known interactions from select model systems to construct really detailed models. A data-driven systems approach, on the other hand, is ideally suited to tackle this problem.

An important goal of applying systems approaches in biology is to understand how a simple genetic change or environmental perturbation influences the behavior of an organism at the molecular level and, ultimately, its phenotype. High-throughput technologies to interrogate the transcriptome, proteome, protein-protein, and protein-DNA interactions present a powerful toolkit to accomplish this goal (DeRisi, Iyer, and Brown 1997; Eichenberger et al. 2004; Laub et al. 2000; Liu, Zhou, et al. 2003; Masuda and Church 2003). However, each of these individual data types captures an incomplete picture of global cellular dynamics. Therefore, these data need to be integrated appropriately to formulate a model that can quantitatively predict how the environment interacts with cellular networks to effect changes in behavior (Facciotti et al. 2004; Faith et al. 2007; Kirschner 2005; Kitano 2002). Accurate prediction of quantitative behavior—the ultimate test of our understanding of a given system—will enable re-engineering of cellular circuits for specific applications relevant to DOE missions.

Scope and Approach

The goal would be to develop a framework that would enable the type of integrative analysis necessary to accomplish reconstruction of predictive models of cellular behavior. The approach (illustrated by the work of Bonneau, Baliga, et al. 2007) would involve genetically or environmentally perturbing the cells, characterizing their growth and survival phenotype, quantitatively measuring steady-state and dynamic changes in mRNAs, assimilating these changes into a network model that can recapitulate all observations, and experimentally validating hypotheses formulated from the model. This type of approach would require the integrated development and implementation of computational and experimental technologies (see Fig. A2.1. Systems Approach for Predictive Modeling of Cellular Responses, p. 70) and would comprise the following steps:

1. Sequence the genome and assign functions to genes by using comparative genomic approaches (for example, protein sequence and structural similarities).
2. Perturb cells by changing relative concentrations of environmental factors and gene knockouts.

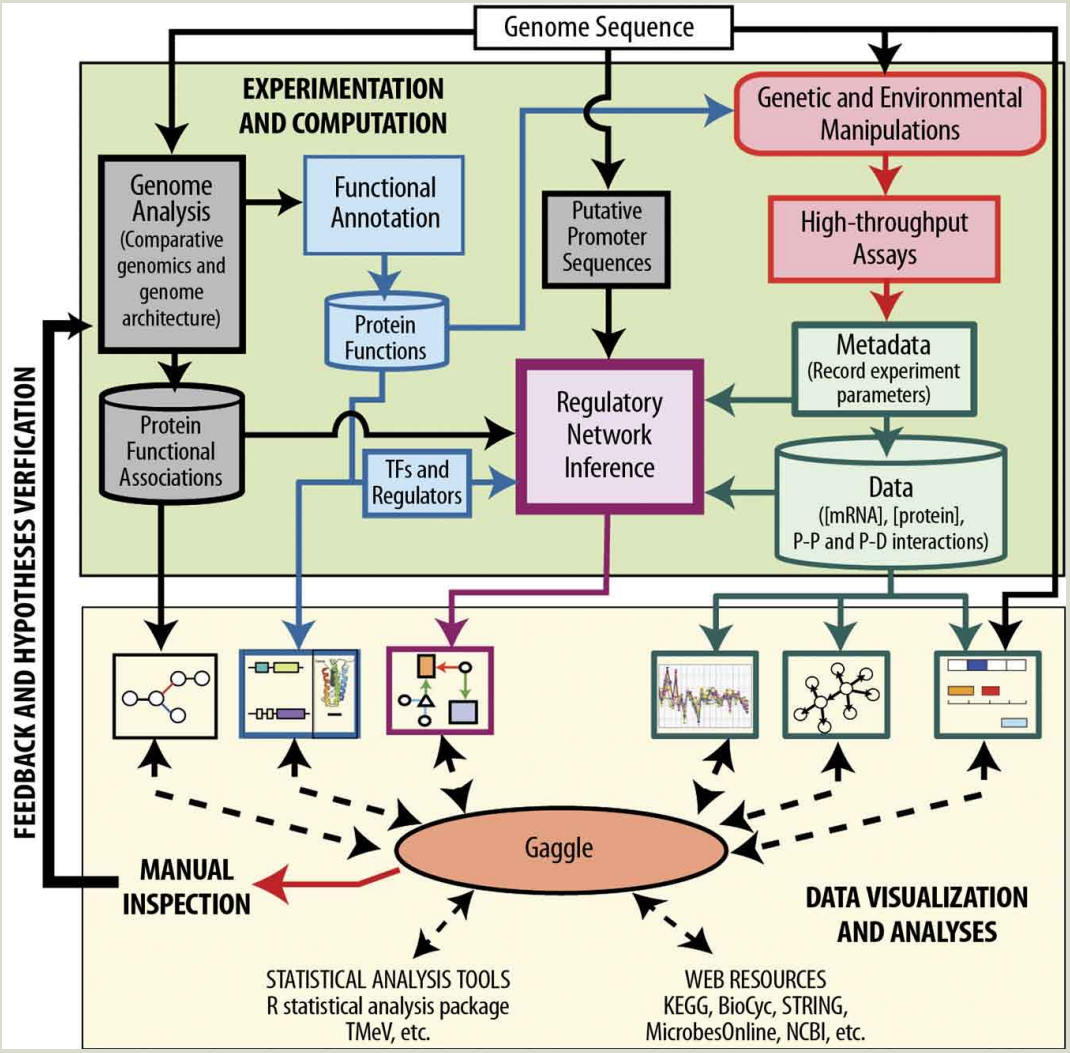


Fig. A2.1. Systems Approach for Predictive Modeling of Cellular Responses. After genome sequencing, two major interconnected and iterative components—experimentation and computation—are followed by data visualization and analyses. Within the first component, major efforts needed include computational genomic analyses for discovering functional associations among proteins (black boxes); putative functional assignment to proteins using sequence- and structure-based methods (blue boxes); and high-throughput microarray, proteomic, and ChIP-chip assays on genetically or environmentally perturbed strains (red boxes). All data from these approaches, along with associated records of experiment design (green boxes), are analyzed with network inference algorithms (purple box). The resulting model is explored with underlying raw data, using software visualization tools within a framework (yellow box) that enables seamless software interoperability and database integration. The interface should be extensible to provide a cost-effective interface to third-party tools and databases. This manual exploration and analysis enable hypothesis formulation and provide feedback for additional iterations of systems analyses. [Source: Adapted with permission from Elsevier. From Bonneau, R., N. Baliga, et al. 2007. "A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell," *Cell* **131**(7), 1354–65 (http://www.sciencedirect.com/science/journal/00928674).]

3. Measure the resulting dynamic and steady-state changes in gene expression, protein-protein interactions, protein-DNA interactions, and protein modifications. Archive these measurements along with digital metadata that capture the genetic and environmental context.

4. Integrate diverse data such as gene expression, evolutionarily conserved associations among proteins, metabolic pathways, and *cis*-regulatory motifs to reduce data complexity and identify subsets of genes that are coregulated in certain environments (biclusters) (Reiss, Baliga, and Bonneau 2006).

5. Using a machine-learning algorithm such as Inferelator, construct a dynamic network model to predict the influence of changes in environmental factors and transcription factors on the expression of coregulated genes (Bonneau et al. 2006).

6. Explore the network in a framework for data integration and software interoperability (Shannon et al. 2006) to formulate and then experimentally test hypotheses to drive additional iterations of steps 2 through 6.

From a practical standpoint, the following are the types of activities associated with the approach described above:

- Extract function information; microarray, proteomic, and metabolomic data; and physical interactions (protein-protein and protein-DNA) from different databases.
- Submit these data to one or more algorithms (written in different computational language environments: R, Matlab) that can infer operational relationships among the genes.
- Interactively visualize, explore, and analyze the inferred network model in the context of underlying raw data to gain biological insight and discover inconsistencies to drive new experiments.
- Record new insights and curate function information to propagate knowledge.

Unmet Technical Needs

The GTL Knowledgebase would need to address the following technical needs during Use Case 2 activities:

- Mapping schema across databases.
- Standardized normalization of data (e.g., quantitation by sequencing versus arrays, two channel versus one channel).
- Standardized statistical models that capture uncertainty.
- Meta-information concerning data collection (e.g., perturbation, growth parameters, and metadata genotype).
- User interface to algorithms to adjust parameters.
- Algorithms that accept standardized data formats and output in standardized formats compatible with visualization software.

Use Case Scenario 3

Use Case 3 has one objective:

- Expand Use Cases 1 and 2 toward key application-related aspects of microeukaryotes (e.g., fungi and algae) and plants.

Implementation of this use case would require a substantial increase in the volume of eukaryotic genomic data, and important issues pertaining to eukaryotes may have requirements that extend in scope and complexity beyond those for prokaryotes. Therefore, the GTL Knowledgebase development strategy for the eukaryote case would combine the following:

- Limiting the initial scope of modeling by key aspects of obvious applied value (e.g., primary and secondary metabolic pathways and selected categories of enzymes); extensively studied model organisms to train the tools; and several most important and tractable target organisms to address actual application issues.
- Providing the foundation (data and tools) for developing new research tools and modeling techniques to expand the initial scope's limit toward the behavior and responses of more complex systems and organisms.

This strategy would allow work to begin immediately on some priority application tasks associated with fungi and plants, in parallel with other developments.

- **Target (Near term):** Mission-relevant and model plants and microeukaryotes—parts inventory and draft metabolic reconstructions. Some fungi and microalgae: detailed reconstruction of core metabolism, key biochemical pathways (i.e., lipid biosynthesis), selected functional enzyme categories (i.e., cellulases), and regulatory circuits.
- **Target (Long term):** Reconstruction of secondary metabolism and identification of potential natural products and secondary metabolites. Development of whole-organism metabolic models. Understanding of molecular bases of plant-microbe and microeukaryote-bacteria interactions (symbiosis and mutualism); coupled organism models.

Use Case Scenario 4

Use Case 4 has one objective:

- Expand Use Cases 1 and 2 toward progressively more complex communities and ecosystems.

As Genomics:GTL science moves toward research involving complex organism interactions and eventually communities, the GTL Knowledgebase would need to evolve capabilities, many of which are difficult to predict at this early stage. Not only different modeling styles and approaches but different types of data would be applied as complexity increases in this objective. The following are some first steps that could be performed within the GKB to enable this area of science.

- **Target:** Well-defined mixed cultures of two or several model prokaryotic species. **Data and modeling needs:** Complete individual genomes, individual and combined metabolic and regulatory reconstructions, and models; obtain omic data for individual organisms and for a mixed culture; assess cross-species signaling, communication, and physical interactions; and obtain imaging data and physiological measurements.
- **Target:** Eukaryote-prokaryote interactions (such as plant-microbes). **Data and modeling needs:** Repeat as above with the addition of at least partial reconstruction of plant metabolism.
- **Target:** Phylogenetically complex microbial communities. **Data and modeling needs:** Obtain metagenomic data (including gene expression, proteomic, and metabolite); inferred metabolic potential; and metabolic coupling. Physiological measurements, including the assessment of carbon and nutrient flow (isotope labeling) in support of metabolic modeling, are particularly important.
- **Target:** Ecosystems and environments. **Data and modeling needs:** Repeat as above but at a substantially larger scale, including the analysis of multiple samples from different locations and measurements of various environmental parameters in conjunction with biological measurements.

Potential impacts of this objective range from fundamental understanding and the ability to simulate and predict global processes associated with carbon cycling and climate changes to the rational control of these global processes. At this stage, the main limiting factor appears to be the lack of modeling techniques adequately reflecting even simple mixed cultures.

Issues and Requirements

Metabolic properties of complex communities would need to be assessed by using metagenomic and related approaches. Significant scientific opportunities exist in an open research area in which many concepts and approaches to data analysis and visualization are yet to be developed. Among specific challenges are the following:

- Data often intrinsically incomplete and unstable.
- Bottom-up approach—importance of reference genomes to infer parts and modules.
- Binning of shotgun sequences to reconstruct phylogenetic groupings.
- Inferences from community composition (e.g., 16S RNA).
- Novel type of probabilistic and intentionally ambiguous assignments (e.g., *either this or this but not that*).
- Estimation of metabolic potential (as opposed to detailed reconstruction).
- Assessment of expressed metagenomes (RNA and proteins) and meta-metabolomes (metabolites).
- Importance of metadata describing samples and features of the environments from which they were collected.
- Remote capture and collection of spatially and temporally heterogeneous environmental and biological data to assess metabolic and biogeochemical processes.

Modeling a Fully Defined Microbial Community

Microbial community modeling is another open research area. Even if we had all the underlying data, we would not know how to model even the simplest and best-defined community (several species, for example). New approaches to modeling microbial interactions and the functions of communities should be developed and tested using simple (reduced) or reconstructed communities. Investigations of cocultures in bioreactors, where factors that limit or shift populations can be tightly controlled and responses monitored, would be a good starting point. Single-cell analyses (i.e., transcriptome, proteome, and investigation of metabolic interactions) also would contribute to focused model development on a system that is not overly complex.

Given the challenge of applying metagenomics to complex communities, nongenomic or targeted genomic approaches should be considered (i.e., process-level models that are genome informed). One example would be a top-down approach (e.g., experimental assessment of carbon flux, other biochemical measurements, and respective modeling techniques).

A long-term challenge will be to bridge the gap between a bottom-up approach, which would develop a metabolic model of an individual organism by using detailed genomic and metagenomic information, and the top-down approach, which would involve measuring and modeling coarse-grained processes such as nutrient and metabolic fluxes at macroscopic scales.

Illustration of Use Case Scenario 1: Integrated Approach to Reconstruction of Metabolic and Transcriptional Regulatory Networks in Bacteria

Comparative genomic and experimental techniques are combined to support the mapping of metabolic and regulatory networks in hundreds (soon to be thousands) of diverse species with completely sequenced genomes. Dissection of cellular processes to compact subsystems allows researchers to accurately project functions of individual genes (enzymes, transporters, and regulators) and elementary modules (pathways) from model organisms to those of others. The growing collection of curated subsystems in the SEED genomic integration (http://theseed.uchicago.edu/FIG/index.cgi; Overbeek et al. 2005) provides a framework supporting the inference of metabolic scenarios (DeJongh et al. 2007) and regulatory circuits (Rodionov 2007).

Comparative subsystem analysis reveals a remarkable picture of conservations and variations in the central machinery of life. Most notably, the observed metabolic diversity appears to emerge, by a combinatorial explosion, from a limited set of operational subsystem variants (stable gene patterns) implemented in different species (Ye et al. 2005).

Application of genome context analysis techniques (e.g., delineation of conserved operons and regulons in prokaryotes) allows researchers to predict previously uncharacterized genes and pathways (Osterman 2006; Osterman and Overbeek 2003; Osterman and Begley 2007). Focused validation experiments are used to test the most critical functional predictions. Among the recently published examples applying this approach are prediction and verification of the novel N-acetyglucosamine utilization pathway in *Shewanella oneidensis* (Yang et al. 2006); genomic discovery and reconstitution of a complete biosynthetic pathway for the osmolyte di-myo-inositol-phosphate in *Thermotoga maritima* (Rodionov et al. 2007); reconstruction of catabolic pathways driven by the novel glycerate-2-kinase in *T. maritima* and other bacteria (Yang et al. 2008); genomic reconstruction and validation of nicotinamide adenine dinucleotide (NAD) metabolism in *Synechocystis* sp. (Gerdes et al. 2006); genomic reconstruction and experimental validation of novel regulons controlling NAD biosynthesis in diverse bacteria including *S. oneidensis*, *Synechocystis* sp., and *T. maritima* (Rodionov et al. 2008a; Rodionov et al. 2008b); bioinformatic prediction and experimental characterization of the novel class of multicomponent vitamin transporters (Rodionov et al. 2009); and novel lactate utilization machinery predicted from *Shewanella* studies and experimentally validated (Pinchuk, Rodionov, Yang, et al. 2009).

Case Study 1: Genomic Encyclopedia of Carbohydrate Utilization

The integrated approach was applied to systematically map mono- and disaccharide catabolic pathways in 15 species of *Shewanella* with completely sequenced genomes. This project is a component of the *Shewanella* Federation studies aimed at better understanding the ecophysiology and speciation of this important genus.

A bioinformatic approach established in this study is briefly illustrated in Panel A of Fig. A2.2. Genomic Reconstruction and Experimental Testing of Sugar Utilization Pathways in *Shewanella* spp., p. 75. A set of ~480 protein families (FIGfams)—including transporters, transcriptional regulators, and several classes of metabolic enzymes (e.g., sugar kinases, oxidoreductases, epimerases, and aldolases) from a

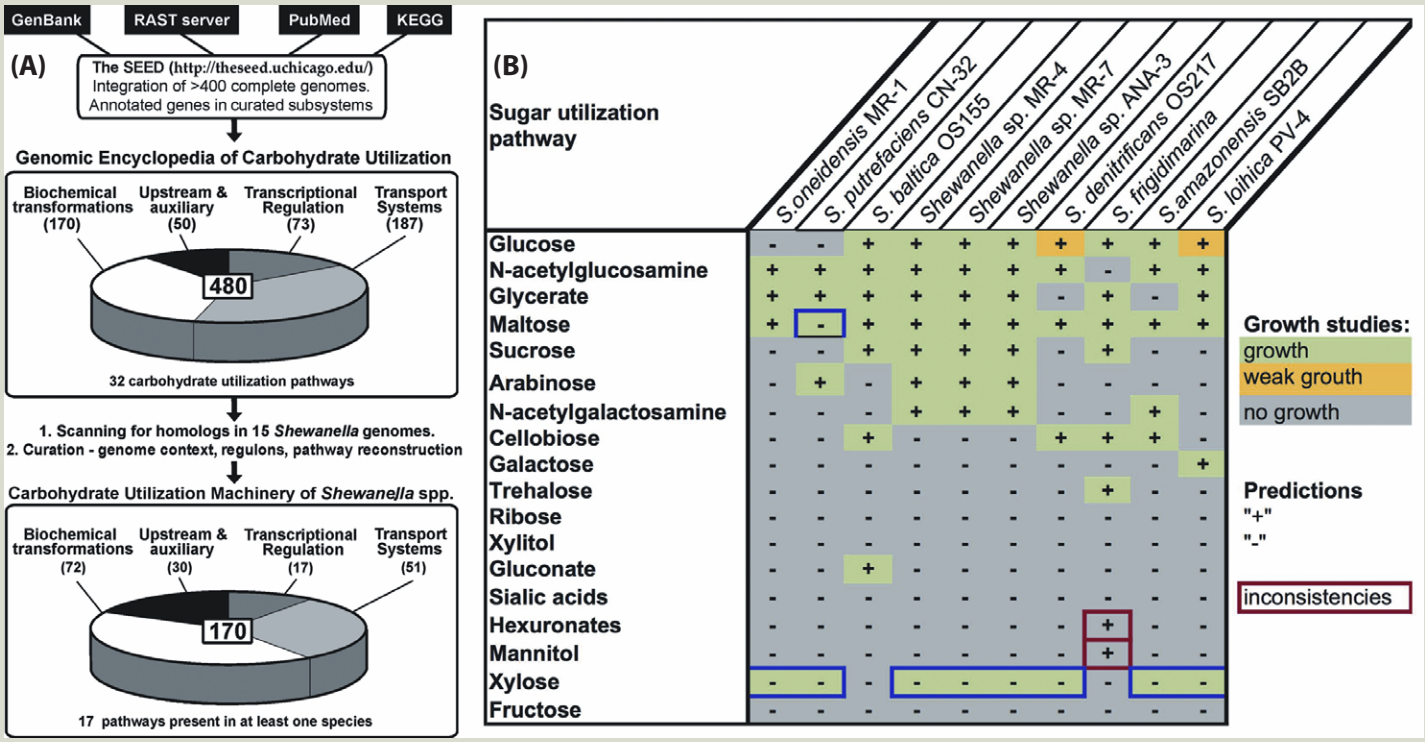


Fig. A2.2. Genomic Reconstruction and Experimental Testing of Sugar Utilization Pathways in *Shewanella* spp. (A) A subsystems-based approach. **(B)** Growth studies illustrated for a subset of 10 *Shewanella* spp. on 18 individual sugars (shown by color) in comparison with bioinformatic predictions (shown by + or –). A few observed inconsistencies are outlined by boxes. For example, many strains were able to grow on xylose, whereas no genes corresponding to any known version of xylose utilization pathways could be identified, suggesting a topic for further research. Some of these results were reported in Fredrickson et al. (2008) and Yang et al. (2006).

collection of carbohydrate utilization subsystems in the SEED—was used to scan for homologs in 15 *Shewanella* genomes. Identified candidate genes were the subject of further genome context analysis for their accurate functional assignment and reconstruction of respective pathways. This analysis detected substantial variations in a sugar diet among different *Shewanella* species, reflecting various aspects of their ecophysiology and evolutionary history. More striking, however, are the differences revealed by comparison with a classical model system of *Escherichia coli.* These differences are manifested at various levels, from the presence or absence of certain sugar catabolic pathways to a dramatically different organization of transcription regulatory networks in the central carbon metabolism. The results of this analysis included prediction of several novel variants of carbohydrate utilization pathways (e.g., for N-acetyglucosamine, sucrose, cellobiose, arabinose, and glycerate) and tentative functional assignments for previously uncharacterized gene families (e.g., within GlcNAc operon SO3503-3507 in *S. oneidensis*, sucrose operon Sfri3988-3991 in *S. firgidimarina*, and cellobiose operon Sbal0541-0545 in *S. baltica*). Overall, of the 170 identified protein families involved in sugar utilization pathways in various strains of *Shewanella*, 60 families may be considered novel because their specific functional assignments were previously unknown. Representatives of eight predicted families in four pathways were verified experimentally by genetic complementation and biochemical characterization of purified recombinant enzymes. The most important validation of inferred pathways, however, was obtained by experimental testing of predicted phenotypes by growth studies. A high level of consistency between predicted and observed phenotypes (the ability to grow on a panel of individual sugars)

is illustrated by the table in Panel B of Fig. A2.2, this page. In addition to the specific knowledge of carbohydrate catabolism in the *Shewanella* genus, this study led to a substantial expansion of the current version of the *Genomic Encyclopedia of Carbohydrate Utilization*. A systematic iterative application of this approach to multiple taxonomic groups of bacteria will further enhance this knowledgebase, providing adequate support for efficient analyses of newly sequenced genomes as well as of emerging metagenomic data.

Case Study 2: Genomic Reconstruction of Metabolic Regulons in *Shewanella* spp.

An integrative comparative genomic (knowledge-driven) approach was used to infer transcriptional regulatory networks (TRN) in 13 species of *Shewanella* spp. To accomplish this goal, we combined the identification of transcription factors (TF), TF-binding sites (TFBS), and cross-genome comparison of regulons with the analysis of the genomic and functional context inferred by metabolic reconstruction (see Panels A and B of Fig. A2.3. Genomic Reconstruction of Transcriptional Regulons in *Shewanella* spp., p. 77). The reconstructed TRNs for the key pathways involved in central metabolism, production of energy and biomass, metal ion homeostasis, and stress response provide a framework for the interpretation of gene expression data. This analysis also helped to improve functional annotations and identify previously uncharacterized genes in metabolic pathways. Overall, this approach allowed us to identify candidate TFBSs for ~80 TFs from the *Shewanella* group. For ~30 described regulons, TFs were conserved between *Shewanella* and *E. coli*, whereas most others were characterized for the first time. Among many other observations, this analysis revealed a substantial rewiring in the TRN of the central metabolism between *Shewanella* and a classical model system of *E. coli* K12 (see Panel C of Fig. A2.3, p. 77).

Acknowledgments. Manuscripts describing Case Studies 1 and 2 are in preparation. Some of these results were presented at the DOE GTL 2008 meeting and briefly described in Fredrickson et al. (2008). These research results are from the work of the *Shewanella* Federation project (J. Fredrickson, principal investigator), supported by a DOE GTL grant. The examples in these sections were provided by A. Osterman and D. Rodionov of the Burnham Institute, La Jolla, California.

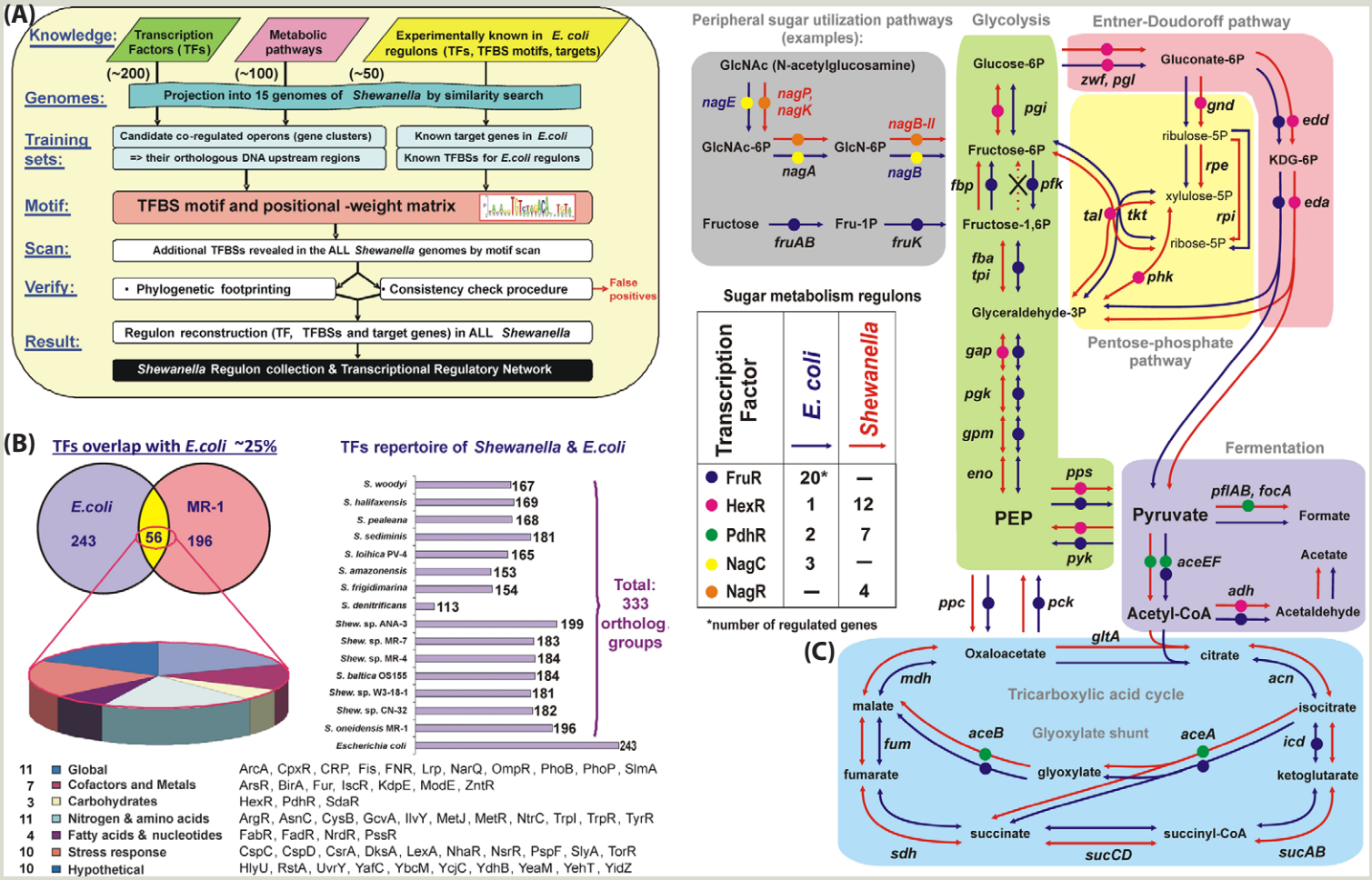


Fig. A2.3. Genomic Reconstruction of Transcriptional Regulons in *Shewanella* spp. (A) A knowledge-driven approach. **(B)** Survey of transcription factor repertoire in *Shewanella* spp. **(C)** Rewiring of regulatory network for central carbohydrate metabolism in comparison of *Shewanella* spp. and *E. coli* K12. For additional information, see Fredrickson et al. (2008). [Source: C portion of figure adapted by permission from Macmillan Publishers Ltd. From Fredrickson, J., et al. 2008. "Towards Environmental Systems Biology of *Shewanella*," *Nature Reviews Microbiology* **6**(8), 592–603 (http://www.nature.com/nrmicro/).]

Appendix 3
Systems Biology for Bioenergy Solutions

Defined here are the GTL Knowledgebase (GKB) requirements to aid and accelerate the understanding and engineering of biological systems for biomass conversion to bioenergy.

Background

The development of renewable alternatives to fossil carbon–based transportation fuels has become an urgent national priority. One of the most promising options for near-term, commercial-scale deployment is biofuel from lignocellulosic biomass (wood chips, grasses, cornstalks, and other inedible plant-based materials). Additionally, biological systems offer multiple paths to diverse bioenergy products—biodiesel from algal or plant biolipids, microbial methane production, or algal production of biohydrogen from sunlight and water.

The scientific breakthroughs needed to make lignocellulosic biofuel a cost-effective alternative to petroleum will require coordinated investigations of plant, microbial, and enzyme systems that span many orders of complexity and scale. Although some challenges are common to biological research across all DOE mission areas—such as noise, complexity, size, dynamic nature, lack of standardization, and heterogeneity of biological "omic" datasets—several issues are unique to bioenergy. While both carbon cycle and environmental remediation research are focused on understanding biological systems in their natural environments, bioenergy research seeks to understand and engineer these systems to work in highly controlled, production-oriented environments ranging from 96-well plates to 100,000-L fermentors or 1000-acre fields.

In addition to the core omic datasets resulting from the analysis of biological systems across multiple DOE missions, other key data unique to bioenergy include linked imaging and chemical characterization data for analyzing lignocellulose structures, imaging interactions within natural and constructed microbial communities, chemical analyses of chemical structure and breakdown intermediates of biomass, and a variety of datasets arising from sustainability research that examines the links among carbon, nutrient, and water cycles, as well as the environmental, economic, and societal impacts of bioenergy technologies being developed in the laboratory.

Biofuels: Grand Challenges for Biology

The ultimate goal for fundamental research in bioenergy, including the three DOE Bioenergy Research Centers, is to understand the biological mechanisms underlying biofuel production so well that those mechanisms can be redesigned, improved, and used to develop novel, efficient bioenergy strategies. Research undertaken by the centers and smaller endeavors will create the knowledge underlying three grand challenges at the frontiers of biology:

- Development of next-generation bioenergy crops.
- Discovery and design of enzymes and microbes with novel biomass-degrading capabilities.
- Discovery and design of microbes that will transform the production of fuel from biomass.

The science to solve these challenges is inherently complex, requiring multiple interdisciplinary teams that approach the same problems from different directions to accelerate scientific progress. Research details from DOE's Biomass to Biofuels Workshop can be found in the report, *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda* (U.S. DOE 2006, http://genomicsgtl.energy.gov/biofuels/). The following sections explain some scientific issues related to these challenges.

Development of Next-Generation Bioenergy Crops

Bioenergy crops are grasses, trees, and other plants grown specifically to provide cellulose and other complex cell-wall compounds, such as lignin, that strengthen and support plant structure as feedstocks for bioenergy. Plant cell walls are so complex that several thousand genes, few of which are known or understood, are thought to be involved in their synthesis and maintenance. By understanding the genes and mechanisms that control cell-wall synthesis in plants, scientists could develop new energy crops with altered biomass composition, modified links within and between cell-wall components, or increased accumulation of easily convertible starches and oils in plant tissues.

In addition to altering biomass composition, other energy-crop improvements include increasing biomass productivity per acre, increasing resistance to pests and drought, and decreasing the application of fertilizers and other inputs. Availability of more plant genome sequence information can accelerate the development of DNA markers used to identify and isolate the many genes associated with traits that can improve energy crops and speed identification of desired genetic variants and production of new energy crops. To obtain enough biomass for large-scale production of biofuels, a variety of perennial grasses and woody crops well suited to different geographies and regional climates across the nation will need to be grown.

Discovery and Design of Enzymes and Microbes with Novel Biomass-Degrading Capabilities

Nature uses both enzymes and multienzyme complexes called *cellulosomes* to break down cellulosic biomass. The performance of the limited number of biomass-degrading enzymes and cellulosomes studied thus far is much slower than many other types of enzymes as a result of a combination of the nearly impenetrable architecture of plant cell walls, chemical and physical changes to biomass during pretreatment, and structural features of the enzymes.

Each DOE Bioenergy Research Center is searching diverse biomass-degrading environments, from hot springs to rainforests to the guts of leaf-eating insects, to identify bacteria that specialize in producing enzymes that degrade biological materials.

Molecular-level understanding of how enzymes and cellulosomes degrade biomass is a prerequisite to designing improved enzymes for processing. Each center is integrating different combinations of methodologies to identify, characterize, and rationally optimize important factors influencing the rapid deconstruction of plant materials into sugars and other energy-rich components that can be converted to biofuels. These methods include high-throughput screens for proteins and metabolites, chemical analyses, state-of-the-art imaging technologies, and computational modeling.

Discovery and Design of Microbes That Will Transform the Production of Fuel from Biomass

In addition to cellulose, other carbohydrates (collectively called *hemicelluloses*) in plant cell walls are broken down into fermentable sugars when biomass is pretreated with heat and chemicals. Although cellulose is made of one type of six-carbon sugar (glucose) that is readily converted into ethanol and other products, microbial fermentation of the five- and six-carbon sugar mix from hemicelluloses is less efficient, thus representing a key area for improvement. En route to the fermentation tank, biomass currently is subjected to physical, chemical, and enzymatic processing steps that can create by-products and conditions that might inhibit microbial conversion of sugars into biofuels. Ethanol and other biofuel products also inhibit microbial fermentation at high concentrations. Consequently, developing microbes robust enough to withstand the stresses of industrial processing and tolerate higher ethanol concentrations is another important research area. Consolidated bioprocessing (CBP) is a more distant research target that could dramatically simplify the entire production process.

Consolidated Bioprocessing

The strategy of consolidated bioprocessing combines cellulose deconstruction and sugar fermentation into a single step mediated by a single "multitalented" microbe or stable mixed culture of microbes. CBP requires a redesign of microbial systems far more extensive than conventional genetic engineering approaches involving only the modifications of a few genes associated with microbial production of a single drug or other biochemical product. A successful CBP microbe or specially designed microbial consortium may be required to produce a variety of biomass-degrading enzymes; produce minimal numbers of molecules that inhibit the overall process; ferment both five- and six-carbon sugars; and thrive in industrial reactors with high temperature, low pH, and high concentrations of biofuel products.

Investigations of Bioenergy Systems Utilizing the Knowledgebase and Systems Biology Methods

Extremely complex phenotypes or functional characteristics important to bioenergy production—plant cell-wall biosynthesis, makeup, and structure; biomass degradation; and product tolerance and toxicity for biofuel-producing microbes—result from the protein products of numerous genes working together to control mechanisms at molecular, cellular, and higher levels. For example, within a plant genome are hundreds of genes participating in cell-wall biosynthesis, and the genomes of certain biomass-degrading microbes encode dozens of genes for hydrolyzing specific plant cell-wall polymers under different environmental conditions.Two general approaches are used to link variations in genes, pathways, and cellular mechanisms to particular phenotypes:

- Top-down: from known phenotype to understanding its cellular mechanisms and the bases for phenotype improvements.
- Bottom-up: from genomic characteristics of the sequenced bioethanol-producing microorganisms to specific phenotypes by quantification of molecular functions and their network clustering from omics technologies, as depicted in Fig. A3.1. Comparing Genomics to Phenotypic Characteristics, p. 82.

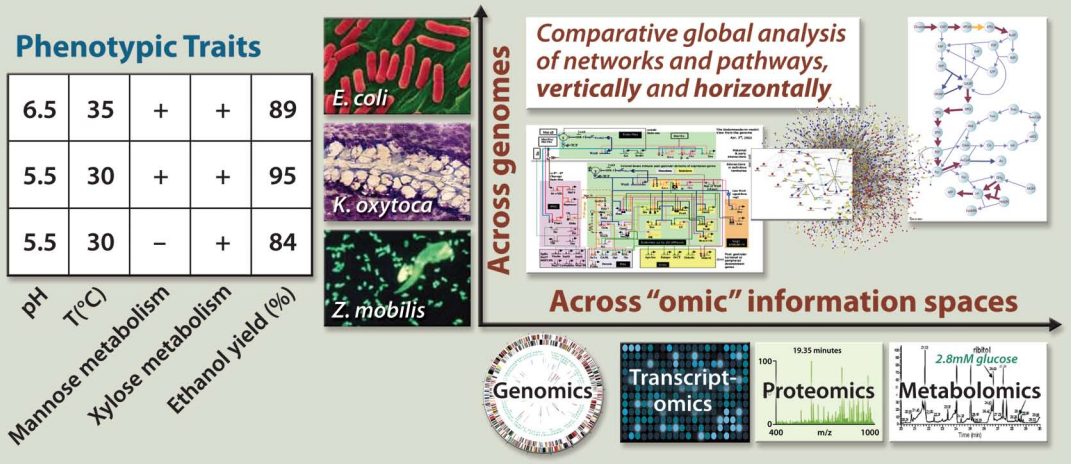


Fig. A3.1. Comparing Genomics to Phenotypic Characteristics.

Plant feedstocks involve additional layers of complexity at the tissue, organ, and whole-organism level that must be considered when studying phenotypes such as high productivity, partitioning to biomass, drought resistance, growth rate, diameter, morphology, field conditions, nutrient uptake rates, and yield. Analyzing observable plant phenotypes from specific natural variations or genetic modifications requires longer time frames for organism growth and development (months or years rather than hours for microbes) and largely manual interpretation by researchers.

Determining the mechanistic underpinnings of important bioenergy phenotypes will require systems biology approaches for the global analysis of completely sequenced organisms, analysis of microbial communities through metagenomics, and more focused investigations to characterize the structures and functions of plant biomass polymers and the enzyme systems that degrade those polymers.

Systems Biology: Enabling a Predictive Understanding for Sequenced Bioenergy Organisms

Applying systems biology approaches to bioenergy research challenges will require core knowledgebase components based on microbial and plant genomes. The completely sequenced genomes of organisms with capabilities relevant to DOE missions would form the foundation of the GTL Knowledgebase.

Microbial Genome Core Component

The microbial knowledgebase component would be a comprehensive repository of systems biology data for sequenced microbes with capabilities relevant to bioenergy production. Some of these microbes could serve as prototypes for platform microbes that are readily engineered to synthesize different fuel compounds. This component would require improved datasets of predicted gene calls and functions from genome sequences, an increased understanding of a minimal set of cellular functions needed to extract energy from substrates and generate desired fuels, and robust understanding of metabolic and regulatory networks that contribute to or detract from any given biofuel pathway. This activity would include quantitative omics (e.g., transcriptomics, proteomics, and metabolomics), transcription factors, flux analyses, and data to trace the flow of carbon, energy, and nutrients through critical pathways. Computational methods are needed that generate accurate predictive models—inferences for how to maximize transformation of renewable resources into fuels (e.g., improvement on conversion rate and conversion yield or the amount of fuel per equivalent substrate consumed).

Plant Genome Core Component

Genomic research can play a major role in developing new crops optimized for biofuel production without decades of agronomic research. Because of the incredible complexity of plant biology at all levels of organization from genome to cell to whole organism, systems biology for plants lags far behind current efforts for microbes. Although the number of sequenced plant genomes continues to grow, the volume of data associated with each plant is much larger than that of a microbial prokaryote.

Screening Plant Genomes for Bioenergy-Related TFs and Binding Sites

To develop plant feedstocks, a more comprehensive knowledge of transcription factors (TF) and the sites they bind throughout a genome is needed to narrow the list of potential genes associated with a specific phenotype such as increased differentiation of plant cells into xylem. Xylem cell walls in plant tissue are the primary source of cellulose used to make cellulosic biofuels, yet many genes linked to xylem differentiation are unknown. By screening the entire genome for new genes regulated by the same transcription factors that control known xylem differentiation genes, researchers can identify a short list of coregulated gene targets to be functionally characterized by experimentation.

Each cell within a plant contains the same genome, but the regulatory controls, subsets of expressed genes and proteins, and collections of metabolites can vary greatly for each cell type and even for subcellular compartments. In addition, multiple growth conditions, different tissues and organs, dynamicism in developmental states, and longer life cycles for plants result in more extended time frames for experimentation and practically endless combinations of variables to explore systematically. The research community is defining the experimental space most relevant to developing model bioenergy crops.

Dealing with biological complexity is a major obstacle, but another challenge is the limited availability of tools for analyzing plant systems data. By building on high-throughput technologies currently available for human and microbial systems, the first steps toward achieving predictive modeling of plant biology for bioenergy applications are within grasp. One initial goal would be to establish a bioenergy plant *phenome* database that identifies and defines the most important plant phenotypes relevant to bioenergy production and links these traits to genomic- and molecular-level functional information.

Improving Functional Annotations for Microbial and Plant Genomes

The availability of complete genome sequences for plants and microbes is the foundation for a broad range of approaches that can be used to characterize genes of unknown function and identify gene products with bioenergy-relevant functions. A genome encodes an organism's complete set of metabolic pathways, yet many key steps in these pathways may involve genes for which no functional information exists. Identifying subsets of genes that are coexpressed and regulated by the same elements (e.g., transcription factors; see sidebar, Screening Plant Genomes for Bioenergy-Related TFs and Binding Sites, this page) is one approach to discovering new genes involved in pathways that control such complex phenotypes as plant cell-wall biosynthesis.

Integrating Different Biological Datasets from Genome-Wide Analyses

The ability to integrate and compare orthogonal datasets would be central to scientific discovery from the GTL Knowledgebase. Relevant tools could either be built into the GKB or be external tools enabled by the GKB (e.g., MicrobesOnline,

http://www.microbesonline.org). The myriad unanticipated ways of combining data types point to the value of allowing for the latter approach—strong support for a tool-development community external to the knowledgebase team itself.

Of primary importance in systems biology research is improving the signal-to-noise ratio. Integration of multiple orthogonal datasets can accomplish this objective. As an example, one might remove spurious components of comparative gene neighbor–based regulon predictions by combining gene expression or protein-level data, protein-interaction data, and transcription factor binding-site motif detection. An additional goal of systems biology is to discover novel and unanticipated relationships. An example might be adding genes to a subsystem or pathway by combining a comparative phylogenetic footprint, expression profiles, and knockout assays of function. Numerous other ways could be used to integrate orthogonal datasets, and the value of such approaches will continue to grow as we devise new combinations.

The following are specific challenges in bioenergy research that might be addressed by integrating orthogonal data.

- Determining the necessary and sufficient subset of genes for elevated ethanol tolerance by combining expression analysis with comparative gene-content analysis of ethanol-resistant microbes.
- Determining the impact of stress conditions on metabolic pathways involved with biofuel synthesis by combining expression and protein-level data with data on metabolites and metabolic flux.
- Engineering novel metabolic pathways for biofuel synthesis from sugars or removal of pathways wasteful or detrimental to producing the desired end product, requiring measurements of the impact on modified pathways by combining expression and protein-level data on metabolites and metabolic flux.

Characterizing and Modeling Microbial Communities

In addition to characterizing omic data from completely sequenced organisms, the GTL Knowledgebase also would need to handle data from the metagenomic analyses of microbial communities. Metagenomics combined with environmental transcriptomics and functional assays for lignocellulose degradation would permit identification of the novel glycosyl hydrolases, transferases, and other important proteins involved.

GTL Knowledgebase Components Unique to Bioenergy Research

In order to identify the genetic basis for the recalcitrance of biomass to deconstruction by enzymes and microbes, omic data must be integrated, compared, and correlated with heterogeneous data. The data should come from a broad range of analyses to identify the most efficient biomass-degrading enzymes and characterize the composition and structural features of plant cell walls.

Linking Imaging Data to Other Experimental Data

Imaging biological systems over a wide range of spatial and temporal scales is an essential component of GTL bioenergy research because this capability provides a method for linking genomic and molecular information to complex biological functions (e.g.,

Integrating Existing Resource Data into GKB Component on Carbohydrate-Active Enzymes

To create a GTL Knowledgebase component focused on carbohydrate-active enzymes, a new integrated database for existing and newly discovered GHs and GTs will pull data from a variety of bioinformatics resources. Among those are the following:

- GenBank for genomic sequences (http://www.ncbi.nlm.nih.gov/Genbank/)
- MicrobesOnline (http://www.microbesonline.org) for comparative genomic and phylogenetic analysis
- Robetta structure prediction server (http://robetta.org) for structural model building
- RCSB Protein Data Bank (PDB, http://www.rcsb.org/pdb/) for structure data
- Gene Expression Omnibus (GEO) for microarray data (http://www.ncbi.nlm.nih.gov/geo)
- BRENDA (http://www.brenda-enzymes.info/), CAZy (http://www.cazy.org), and the UniProt Knowledgebase (http://www.uniprot.org) for guides that may be used more directly as repositories for enzyme functional parameters

The GDB carbohydrate-active enzyme component also will capture experimental conditions, protocols, and results to develop a cross-referenced database indexed to enzyme sequence and, where possible, high-resolution crystallographic structures.

Datasets from Sustainability Research

Sustainability research to analyze the potential economic, environmental, and social consequences of different pathways to large-scale biofuel production will play a critical role in determining the viability of new bioenergy technologies arising from systems biology research.

changes in microbial community behavior, enzyme-biomass interactions, and plant cell-wall structural information). Improving biomass conversion efficiency is key to biofuel production and requires the integration of multiple technologies to correlate changes in chemical properties with observed changes in cell-wall structure during the degradation process (see sidebar, Plant Cell-Wall Characterization and Visualization, beginning on p. 86, for a description of diverse data types generated by different techniques used to analyze biomass structure and chemical composition at multiple scales of space and time).

A major bioinformatics challenge for the GTL Knowledgebase would be developing efficient strategies for analyzing, storing, and sharing the vast amount of image data generated by the characterization of biomass and enzyme structures. Storage of all raw data within the GKB would be impractical, so another requirement would be tools for reducing noise and retaining only the most relevant, high quality data needed for subsequent analyses and visualization. To enhance interpretation and evaluation of image data, methods for associating experimental conditions with images and extracting and annotating the most biologically meaningful image features also would be needed.

Tools for Rapid Identification of Important Bioenergy Functions from Large-Scale Datasets

One of the most pressing needs in advancing production of lignocellulosic biofuels is the development of a comprehensive and robust knowledgebase of carbohydrate-active enzymes such as glycoside hydrolase (GH) and glycosyl transferase (GT) enzymes (see sidebar, Integrating Existing Resource Data into GKB Component on Carbohydrate-Active Enzymes, this page). This GKB component would combine the evolutionary history of GHs and GTs with structural, thermodynamic, and kinetic characterization to produce a more robust data and training set and comprehensive resource for the GH and GT research communities. This activity would include the development of

normalized functional assays for single and multiplexed GH and GT enzymes, as well as substrates for targeted reaction schemes and products.

Combining this work with comparative genomic and metagenomic studies would go a long way toward realizing the goal of establishing and predicting GH and GT libraries that would permit engineering of custom activities. Also important would be determining supporting genes within such systems as well as the combinations of enzyme subfamilies within a single organism or in combination. Unquestionably, accomplishing this goal will occur only by integrating the above data types to determine the rules governing which sequence begets a specific structure. Key data to be

Plant Cell-Wall Characterization and Visualization

Progress in understanding and manipulating the many steps involved in biofuel production will require a broad range of information regarding the chemical composition and molecular ultrastructure of the cell walls that make up the bulk of biomass (see figure, Switchgrass—Fluorescence Microscopy, facing page). Ready access to this molecular phenotype information will allow the genetic basis to be determined, providing a much deeper level of understanding than bulk analysis (e.g., mass and energy transfer and efficiency assessments).

Rapid development of plant cultivars whose biomass is more readily or efficiently deconstructed to simple sugars requires in-depth knowledge of cell-wall biosynthesis. Genetic analysis alone cannot provide this knowledge because the physical, chemical, and biological consequences of genetic manipulation must be identified. Quantitative data will be much more useful than qualitative data in this context. For example, quantitative analytical data describing the cell wall's physical and chemical features can be used to map these features to plant genetics (e.g., by quantitative trait loci analysis). Such data also are invaluable for forward and reverse genetic experiments. Furthermore, identifying molecular mechanisms that lead to increased efficiency of pretreatment chemistry, enzymatic deconstruction, and biological deconstruction requires quantitatively accurate analysis of chemistry and ultrastructure of cell walls in biomass before and after chemical or biological processing. Such analysis will, for example, provide information regarding the amount, molecular accessibility, and chemical and enzymatic susceptibility of various polymeric constituents of the biomass and how these parameters change as a function of genetic manipulation, development, and environment.

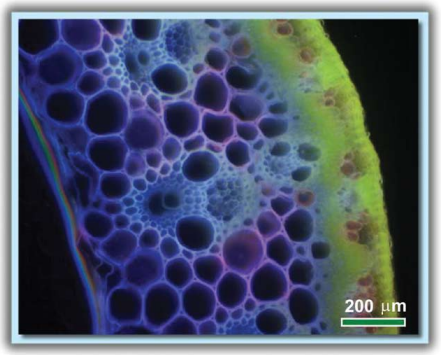
A wide range of analytical data will become available in the near future, including the following:

- **Wet chemistry analyses.** These investigations include carbohydrate content, monosaccharide composition, lignin content, and monolignol composition.
- **Spectroscopic analyses such as mass spectrometry (MS), infrared spectroscopy, and nuclear magnetic resonance (NMR) spectroscopy.** These techniques provide specific chemical and structural information, which often takes the form of "spectral features" that can be mapped to specific "structural features." Numerous examples of this type of data include pyrolysis molecular beam MS (PMBMS), which provides lignin and carbohydrate compositional data; solid-state NMR analysis of biomass for information on the crystallinity and allomorph composition of cellulose in the biomass; and high-resolution solution-state NMR analysis of solubilized biomass for data on the composition and detailed primary structural features of lignin and hemicellulose components of the biomass.
- **Surface chemistry analysis.** Novel spectroscopic and imaging methods are being developed to identify and quantitate specific molecules present at solvent-accessible surfaces. For example, MALDI-TOF microscopy provides information regarding the distribution of small molecules or molecular fragments at the sample surface.
- **Molecular mass analysis of polymeric and oligomeric components.** This analysis can be performed using various combinations of chromatographic and light-scattering technologies, as well as by MS methods.

considered would be which set of conditions, in the presence of supporting proteins and other factors, possesses the desired specificity and reaction kinetics for a given sugar biopolymer moiety. Proteomic and genomic tools are available and considered mature technologies. Bioinformatic tools that record structure-function relationships are making advances, but computational modeling tools to improve enzyme performance (e.g., sequence and structure based) are still extremely variable in terms of reliably identifying mutations that modify specificity and improve performance and stability. To provide a reliable tool for these computational efforts, researchers would need to integrate more biochemical experimental data with this sequence and structural information.

- **High-resolution imaging and diffraction methods.** These methods provide information regarding the molecular ultrastructure of polymeric components of biomass. This information includes the organization of cellulose microfibrils and spatial distribution of hemicelluloses and lignin in biomass.
- **Miscellaneous measurements.** These include physical properties such as porosity, density, compressibility, and heat capacity.

The huge amount of raw data recorded using these plant cell-wall characterization techniques can be interpreted by only a few experts. Therefore, the end user will require processed data to extract the most relevant parameters. For example, solid-state NMR spectra include significant information regarding cellulose crystallinity, but the end user probably is interested in a concise specification of the parameters that describe crystallinity. Because of the possibility of processing errors in such tabular data, robust provenance information must be included. For example, the biological source (e.g., genetic background of the plant cultivar) and physical and chemical processing history of the sample analyzed to produce the raw data should be readily available.



Ding et al. unpublished results.

Switchgrass—Fluorescence Microscopy. Fluorescence signals primarily come from chlorophyll, lignin, carotenes, and xanthophylls in plants, each with a different wavelength (color); lignin fluorescence is blue-greenish. Cell lignification is determined by using different filter sets. [Source: National Renewable Energy Laboratory and the DOE BioEnergy Science Center.]

Explicit storage of all raw data is impractical because of data-storage capacities and increase in the amount of data irrelevant to most knowledgebase users (i.e., noise). Only high-quality raw data should be included, for example, to be used for algorithm training.

A broad range of analytical and visualization tools will be required to effectively utilize this diverse cell-wall structural data. Evaluation of microimaging data will require visualization tools to render, zoom, and rotate images and to identify, demarcate, and quantitate relevant structural elements represented in these images.

In addition, a robust object model is required for the abstraction of structural features. That is, information about the same structural features (e.g., cellulose crystallinity) may be obtained using different analytical techniques. Nevertheless, the same formal representation of *cellulose crystallinity* should be used in structural feature tables obtained by different methods. A truly useful object model would be concise enough to ensure efficient querying, yet expressive enough to include all salient features required for meaningful analysis and comparison of data. Fundamental relationships such as partonomy (e.g., "a cellWall hasPart cellulose") and connectivity (e.g., "a sideChain isConnectedTo a polysaccharideBackbone") would provide deeper context for data retrieval and evaluation.

Appendix 4

Opportunities and Requirements for Research in Carbon Cycling and Environmental Remediation

Underlying problems in understanding the global carbon cycle and processes important to environmental remediation are similar in that both topics involve the mechanistic understanding of the fate and transport of species as they flow through ecosystems. Both topics are challenged by the classic scaling problem—connecting spatial and temporal scales of molecular processes to the macro-scales of ecosystems and beyond.

Carbon Cycle

A pressing national need is for comprehensive understanding of the planetary carbon cycle across terrestrial and ocean environments to provide scientific underpinnings for more robust modeling of climate change and to define carbon biosequestration options over the coming decades. To reach a consensus on projections for future climate, the scientific community needs a better understanding of fundamental mechanisms that control carbon sources and sinks. In the past two decades, much progress has been made in understanding historical trends in atmospheric CO_2, and biogeochemical modeling of carbon in oceans and terrestrial systems continues to advance. The current state of carbon cycle science, however, still cannot quantitatively address several key questions:

- Where are major carbon sinks in ocean and terrestrial ecosystems?
- Which mechanisms control behaviors of carbon sinks and sources?
- How long will biologically sequestered carbon remain stored?
- Will current carbon sinks persist or become carbon sources in a warmer, higher-CO_2 world?
- How do human activities impact carbon storage and release in ecosystems?

Biological processes are fundamental to planet-wide carbon cycling. Thus a mechanistic, systems-level understanding of complex biogeochemical processes at multiple scales will be essential for predicting climate-ecosystem feedbacks. Key issues revolve around photosynthetic productivity; partitioning of photosynthate into energy or biomass pathways; respiration mechanisms; paths to recalcitrant compounds and structures with long environmental residence times; and the effect of environmental variables, nutrients, and water in the context of climate change. Research details from DOE's Carbon Cycling and Biosequestration Workshop can be found in the report, *Carbon Cycling and Biosequestration: Integrating Biology and Climate Through Systems Science* (U.S. DOE 2008, http://genomicsgtl.energy.gov/carboncycle/).

Environmental Remediation

DOE is faced with a daunting legacy of the Cold War: environmental management of the nuclear weapons complex consisting of more than 5000 surplus facilities and associated land located at 144 sites in 31 states and a U.S. territory. More than 380,000 m^3

of high-level radioactive waste are estimated to be stored at several sites. Furthermore, the total volume of soil, groundwater, and sediment present in highly diverse environments and contaminated with complex mixtures of radionuclides, metals, and organic contaminants may exceed 1800 million m^3. Waste treatment, environmental remediation, and long-term stewardship of these sites are estimated to require hundreds of billions of dollars and major breakthroughs in science and technology.

A key to making informed decisions regarding DOE site remediation and stewardship is to understand the interdependent biological, chemical, and physical processes that interact at multiple scales to control contaminant form and transport in the environment. New knowledge and tools are evolving rapidly from DOE Office of Science programs and initiatives to address important gaps in our understanding of the molecular sciences and the complex machinery of life and to develop the computational power and infrastructure for simulating and scaling complex interdependent phenomena. Taking advantage of these emerging resources will be critical to providing the scientific foundation for environmental remediation and stewardship of DOE sites.

Similar Requirements for Common Elements of Domains

Research problems in carbon cycling and environmental remediation have many common elements that render their knowledgebase requirements very similar, if not identical. Ultimately, a major goal is to understand biogeochemical processes at a level required for predictive modeling of both carbon cycling and contaminant fate and transport at the field scale. Examples follow.

- Scaling of processes occurring at the molecular and microscopic scales to the field scale is critical for understanding and predicting biogeochemical processes involving carbon cycling and contaminant transformations.
- Both domains deal with complex environmental systems in which organisms interact intimately with and are affected by their surrounding chemical and physical environment.
- Key processes and properties controlling carbon cycling or contaminant solubility may be occurring in only a small subset over possible domains in a given environment, yet these microscale processes and properties exert a disproportional influence on system behavior.

Enabling the connection of various omic measurements to biochemical and biogeochemical processes is a universal need in the environmental sciences. To this end, the Genomics:GTL program (GTL) has identified as one of its major mission-related goals the development of methods to relate genomics-based microbial ecophysiology (functionality) to the assessment of global carbon biosequestration strategies and climate impacts. The problem can be stated simply as the need to move from sequence to physiology to activities.

Example Problems and Key Scientific Issues

Carbon Cycling in Ocean and Terrestrial Ecosystems

Genomics governs the molecular processes that control cellular, organism, ecosystem, and, ultimately, global phenomena. Connecting mechanistic understanding at the molecular level to the physiological changes observed in organisms, ecosystems, and the global

climate represents a major challenge (see Box 1.1, Global Carbon Cycling Research, beginning on p. 10). Historically, the climate, ecosystem, and molecular biology research domains have had limited overlap because they differ widely in experimental and modeling approaches used, and results, in many cases, do not translate well across scales.

Marine Environments

The following are key questions surrounding the carbon cycle:

- How do metabolic processes of microbial communities in marine habitats link to the global carbon cycle, with special attention to integration of processes across genetic, organismal, community, and ecosystem scales?
- What are the links among the composition of dissolved organic matter, nutrient limitation, and the structure of heterotrophic microbial communities in marine systems?
- How do environmental, ecological, and physiological factors interact to set the pathways and regulate the flows of carbon and other elements through upper-ocean ecosystems?

Phytoplankton (microscopic marine plants) and photosynthetic bacteria convert dissolved CO_2 into organic compounds in surface waters. By reducing the partial pressure of CO_2 in the upper ocean, photosynthetic marine microbes enhance oceans' physical absorption of CO_2 from the atmosphere. Without phytoplankton photosynthesis, atmospheric CO_2 concentration would be 150 to 200 ppmv higher (Laws et al. 2000). Large oscillations in phytoplankton abundance, therefore, significantly impact the oceans' ability to take up atmospheric CO_2. Several challenging issues, such as the following, need study:

- Understanding the composition of microbial communities that dominate primary production in oceans (in a beginning phase).
- Determining differences in functional potential and metabolic processes of various types of photosynthetic microbes (poorly understood).
- Predicting how communities might be affected by climate change and its impact on the marine carbon cycle (difficult).

Metagenomics and related omic measurements have the potential to provide detailed insights into the structure and function of marine phytoplankton communities.

Terrestrial Environments

In terrestrial ecosystems, plants also use photosynthesis to convert atmospheric CO_2 into organic compounds for building plant biomass and driving metabolic processes. Key issues include the following:

- How can we better distinguish between regulatory systems and molecular controls for partitioning carbon among plant structures versus cellular respiration among different soil pools, and how can we represent this new knowledge in models?
- How will climate change influence enzymes and biochemical reactions underlying water use efficiency, nutrient uptake, and many other processes? These processes control photosynthetic productivity as plants are subjected to levels of atmospheric CO_2 and other conditions that have not existed for the past 650,000 years and possibly millions of years.

- How can we better quantify photosynthesis, respiration, and other biological processes that influence carbon cycling? The problem is that the metabolic flux of material and energy through cells, organisms, and ecosystems is tightly linked to the abiotic environmental factors (e.g., temperature, precipitation amounts and timing, geographical features, nutrient availability, length of days and seasons, and sunlight exposure) that define a particular region. Climate is both a product and a catalyst of interactions between a region's physical environment and the biosphere, all of which are driven by the sun and affected by human activities.
- How do soil microbes and their activities respond to climate change, and what are the consequences for carbon cycling processes?

Critical for global-scale climate and biogeochemical models are accurate estimates of process rate constants, which influence biochemical functionality in organisms. Biochemical functionality generally is defined in terms of a catalytic property plus its rate constant (V_{max}) that can be plugged into models operating at larger-system scales.

A key question is the extent of sequence divergence in orthologs (e.g., in extracellular hydrolases) related to variability in V_{max} (rate per unit biomass). In particular, to what extent is phylogenetic information (SSU rRNA and MLST) related to functionality (both substrate catalysis and environmental stress responses)? A key need is to be able to use omic information to provide estimates of catalytic rates and to inform process type and mechanism (see Fig. A2.1. Systems Approach for Predictive Modeling of Cellular Responses, p. 70).

Current measurements of genomes, transcriptomes, and proteomes can give, at best, relative abundances of molecules whose function is largely inferred from sequence homology. This is a major problem that will require a concerted, extensive effort as well as new and innovative approaches, technologies, systems for data and information sharing, and models. The potential payoff for success is tremendous, not just for carbon cycling but for all DOE science missions, environmental remediation, bioenergy, and beyond.

Optimizing Productivity and Carbon Biosequestration in Managed Ecosystems

Managed lands account for about 30% of current global terrestrial net primary production (NPP), and ongoing land-use changes will cause the proportion of global NPP from managed lands to continue to increase. By the end of the 21st Century, managed ecosystems will dominate the planet. To establish a basis for optimizing carbon fixation and biosequestration in this context, a fundamental approach is needed to provide a molecular mechanistic set of options. Along with natural ecosystem behaviors, these options will form a substantial component of Earth System Models. Many of these studies have a common basis and data management needs similar to those of bioenergy studies. Objectives include the following:

- Identify basic processes that underlie gross and net primary production (GPP and NPP, respectively) of terrestrial plants; examine molecular controls on above- and belowground components of NPP; and assess areas in which knowledge gained through mechanistic studies could lead to enhanced carbon biosequestration in plant biomass or soils.
- Consider how efficient acquisition and utilization of resources (e.g., nutrients and water) contribute to maximizing rates of GPP and NPP in terrestrial plants;

identify the molecular basis of efficient resource utilization; and assess interactions between carbon and other resources that might be important in determining the rate, magnitude, or sustainability of biosequestration.

- Evaluate how GPP and NPP could be maximized in plant populations and communities and consider the role of genetic diversity and resource utilization in carbon biosequestration. The objective is to maximize NPP and litter input to soils, for example, over a growing season.
- Generate dynamic models (in silico leaf and in silico plant) that predict how changes in genetic regulatory networks can be used to enhance GPP or NPP by altering metabolic and developmental pathways in response to external perturbations or genetic manipulation.

Leaf-Level Strategies

Emergent mechanistic and systems-based models of GPP provide potential opportunities to substantially increase carbon fixation in managed ecosystems, with impact on both DOE carbon biosequestration and biofuel strategies. The following are examples:

- **Modifying the diffusion resistance to CO_2 transport in leaves.** Mesophyll resistance is a significant limitation on carbon acquisition (24% reduction) and on water and nutrient use efficiencies.
  - **Taking steps to suppress or bypass photorespiration.** RuBisCo (Ribulose-1,5-bisphosphate carboxylase/oxygenase) evolved without the ability to discriminate between its primary substrate, CO_2, and the wasteful reaction with oxygen (a 35% reduction in carbon capture).
  - **Engineering maladapted RuBisCo in plants.** RuBisCo in current C_3 plants is optimized for historic concentrations of CO_2, 200 ppmv. Introducing RuBisCo into C_3 plants from other species that have a higher catalytic activity (and are better suited for higher CO_2) would dramatically increase carbon gain despite less ability to discriminate for CO_2 over O_2.
  - **Optimizing the distribution of nitrogen within the photosynthetic apparatus.** Nearly half of nitrogen invested in soluble protein in leaves is in RuBisCo. Manipulating the partitioning of nitrogen resources (e.g., in the regenerative phase of the Calvin cycle) could greatly increase the potential for carbon acquisition without any increase in the total nitrogen requirement.

Plant-Level Strategies

- **Minimizing carbon-sink limitations and negative feedback on photosynthesis.** Source-sink interactions have a significant impact on photosynthesis and plant growth. Limited sink capacity results in decreased photosynthetic rates in leaf tissue. Photosynthetic activity is tightly regulated by sink demand. Therefore, increased productivity may be achieved by reducing sink limitations on photosynthetic rates. Opportunities to achieve these reductions come from recent experiments suggesting that sink regulation of photosynthesis is mediated by alterations in phloem loading (Chiou and Bush 1998; Vaughn, Harrington, and Bush 2002).
- **Optimizing carbon-nitrogen metabolism for increased plant productivity.** The interaction between CO_2 and nitrogen assimilation is of key importance to productivity. Assimilation of inorganic nitrogen into organic nitrogen requires photosynthetically derived carbon skeletons to serve as backbones for assimilation

of nitrogen into amino acids. Attempts to improve productivity or alter partitioning will be informed by improved understanding of central metabolism.

- **Partitioning carbon into organs and soil organic matter.** Modification of plant morphology and phenology can have substantial impacts on productivity by, for example, enhancing gas exchange in the shoots and increasing nutrient and water acquisition in the roots.
- **Identifying genes controlling biomass by using genetic screening approaches (forward and reverse genetics and natural variation).** The unifying question for this approach is whether previously unidentified genetic loci control plant productivity. Possible targets might include novel aspects of photosynthesis (from light reactions to RuBisCo and carbon metabolism) and water and nutrient use efficiencies. Completely novel pathways and master regulatory genes also may emerge from such genetic screens.
- **Optimizing biomass production versus respiration.** Partitioning of carbon between respiration and dry matter production is variable at the ecosystem, population, organismal, and tissue levels. Thus, establishing a mechanistic picture will enable optimization at all of them.

Environmental Remediation

DOE contaminant fate and transport and biogeochemistry research focuses on natural microbial communities that significantly enhance or decrease the environmental mobility and toxicity of contaminants. The interdependent metabolic survival strategies used by microbial communities can directly or indirectly alter contaminant behavior in the subsurface or transform toxic contaminants into relatively benign forms. For example, *Shewanella* and *Geobacter*, two model dissimilatory metal-reducing microbes, currently are the subjects of systems-level research in GTL and OBER's Environmental Remediation Sciences Program. Via direct enzymatic reactions or indirectly via the generation of Fe(II), these microbes can reduce a range of radionuclides and toxic metals that are constituents of DOE wastes. Under anoxic conditions, for example, they can transform U(VI), which is relatively soluble and therefore can migrate in groundwater, to U(IV), which is insoluble and precipitates as a poorly soluble nanocrystalline solid (UO_2). Studying these model microbes at a systems level is the first step in expanding our understanding of the structure, function, metabolic activity, and dynamic nature of microbial communities that have an important role in contaminant biogeochemistry. This knowledge is needed to predict microbe-mediated contaminant fate and transport and to understand how microorganisms may affect key biogeochemical processes in the environment. Of importance, however, is recognizing that these organisms do not function in isolation but rather are components of heterogeneous populations or microbial communities that may interact to various degrees and, in some cases, even exhibit interdependence.

A critical problem in biogeochemistry and environmental remediation science is the lack of understanding about how microbial processes are coupled to geochemical and hydrological processes influential in contaminant behavior and how these processes are scaled in heterogeneous environments. In addition, new tools are needed for measuring key microbial, geochemical, hydrological, and geological properties and processes in these systems. Less than 1% of all microorganisms collected at only a few DOE sites have been cultured and characterized in any great detail, and only a small

fraction of those have been sequenced. Even less is known regarding the interactions of microorganisms in communities. Metabolic processes observed in the subsurface often are the result of unique interactions between the microbial community and subsurface geochemistry. We have only begun to appreciate the existence of such systems, let alone understand them sufficiently to take advantage of their diverse capabilities and predict how they may influence contaminant behavior. Efforts are under way at several field sites (e.g., the Oak Ridge National Laboratory Field Research Center) to define the genomic potential of microbial communities using metagenomic and other molecular approaches to provide initial culture-independent insights into potential microbial functions such as the following:

- Linking genome sequence to functional potential, as described below, remains a key issue that must be resolved if the promise of genomics for understanding the function of microbial communities is to be realized.
- Collecting robust environmental data that can be linked directly to metagenomic sequence also is a critical issue. Environmental context is very important for interpreting such data.

Other processes important in modifying contaminant form and transport and in developing environmental remediation strategies include microbe-mineral interactions and resulting molecular structural and charge-transfer responses; microbial community responses (e.g., signaling, motility, biofilm formation, and other structural responses); and ensuing community functionality. The mechanistic linking of metabolism to contaminant transformation will represent an important advance from previous contaminant-fate models.

Examining the Common Challenges of Carbon Cycling and Environmental Remediation

Microbial communities inhabiting terrestrial and aquatic environments are major players in the global carbon cycle and environmental remediation, but the organisms and biogeochemical processes they catalyze remain poorly understood.

While the genomes of hundreds of microorganisms from a range of terrestrial and aquatic habitats have been fully sequenced, they represent only a small fraction of total microbial diversity. New technologies such as metagenomics, metatranscriptomics, and metaproteomics offer a window into the metabolisms and lifestyles of the vast diversity of microbes, including uncultivated organisms. However, most successful applications have been applied to relatively "simple" microbial populations; the daunting complexity of most terrestrial and aquatic communities thus far has not yielded data that are easily translated into functionality.

Annotation

A large generic annotation problem remains in genomics: predicting protein function from sequence and homology. In some cases, defining a general functional class of a specific protein, such as amino acid transporter, is relatively easy, but identifying substrate range (i.e., which amino acids it transports) can be extremely difficult. These issues can be important for answering ecophysiology questions and for determining function within metabolic networks. Even more challenging are situations in which homologies are poor or nonexistent. A potentially powerful approach for determining gene function and ultimately improving prediction is the combined

comparative genomic and experimental approach used by Yang et al. (2006) to characterize the previously unknown *N*-acetylglucosamine utilization pathway in *Shewanella*. This approach is eminently scalable and would benefit from high-throughput protein production and characterization capabilities currently being developed within GTL. A complementary approach would be to specifically target enzyme systems that process important extracellular compounds key to carbon cycling in terrestrial and marine systems.

Linking genomics-based information to function requires both genome-scale data generation and systems biology tool development. On the data-generation side, data collected at transcriptomic, proteomic, and metabolomic levels are critical parameters that need to be assayed and quantified. On the computational side, a pressing need is to develop tools to integrate genome-scale data over time courses of treatments and to develop predictive modeling tools.

Scaling

The general challenge of scaling across process, spatial, and temporal scales is at the heart of both global carbon cycle and environmental remediation problems. Environmental scientists measure biogeochemical functions and phenomena at a range of scales but often have difficulty relating their results to higher and lower scales and extrapolating behavior outside the range of observations. Scientists work at certain scales and tend to think that those working at scales above them do not capture enough detail to say anything definitive. People working at higher scales do not think that those working on smaller scales can provide the information needed to parameterize a model at a higher scale. This contention has led to a disconnect in research across scales. Caution must be exerted in regard to generating large volumes of data that do not impact the macroscale where most effects concerning carbon cycling or contaminant transport are manifested. Ultimately, we want to conduct predictive modeling—to predict how a system will react under specific conditions—and not simply reproduce what already has been established via experimentation or through observations. To do so, we must be able to populate models with increasing levels of detail over different space and time scales and to develop innovative approaches and means for upscaling processes and properties from molecular to field scales (see figure, Scales and Processes, p. 11).

The length and time scales of the carbon cycling process to be included in future climate change and contaminant fate and transport models are likely to range from molecular to microscopic to aggregate (millimeters to centimeters) to field and beyond. In particular, they will include data yielding not only fundamental understanding of complex biological systems but also parameterized dynamic models capable of quantitatively predicting system response to ecosystem changes and disruptions. These data include not only single cells but also microbial communities, organisms with multiple cells and tissues, and complex communities and ecosystems with many species. Such model development will require measurements and quantitative characterization of biogeochemical processes that in some cases will be needed specifically for model parameterization and validation (as opposed to meeting needs of general scientific interests and relevance). Also needed are new methods for coupling informatic data, whether derived from ecological observations or high-throughput systems biology investigations of genomically informed cellular processes, to parameterized models of system response at various levels of complexity and scales. In particular, incorporating

increasingly sophisticated and detailed models of complex processes contributing to and ultimately governing carbon cycling to produce increasingly quantitatively predictive models will require addressing model scalability and the coupling of mathematically heterogeneous representations. Furthermore, while current climate change models have "hooks" to incorporate parameterized models employing increasingly detailed carbon cycling data, next-generation models are likely to require new methods for submodel parameterization and coupling that would rely on GKB data resources.

Connecting Data to Function—Dealing with Complexity

To overcome the obstacles of translating omic data into function, researchers will need to develop techniques to enable targeted metagenomic (or other omic) research. Methods such as stable isotope probing or metabolic labeling with bromodeoxyuridine will allow us to effectively isolate important segments of the total microbial community without cultivation and thus begin to understand the functional roles of different community segments. Metatranscriptomics and metaproteomics, which target primarily the "active" microbial community and their expressed macromolecules, will result in unraveling complexity and provide insight into actively occurring processes. Single-cell genomics, using cells obtained via flow sorting or micromanipulation, has potential for even more targeted analyses of community members and for further reducing the impact of complexity on metaomic approaches.

While the native communities in soils and oceans are complex, techniques and approaches under development, such as those described above, can begin to overcome some of the technical issues associated with complexity. Additionally, understanding entire communities associated with key environments would be invaluable as a baseline.

As DNA sequencing becomes ever more accessible and less expensive, we can envision a human genome–type project such as that suggested in a recent National Academy of Sciences report, *The New Science of Metagenomics* (http://www.nap.edu/openbook.php?record_id=11902&page=R1), to target the microbiome in a spectrum of representative habitats. The National Institutes of Health's Human Microbiome Project (http://nihroadmap.nih.gov/hmp/) and the Global Ocean Sampling (GOS; http://collections.plos.org/plosbiology/gos-2007.php) serve as models for this type of large-scale project. GOS is a useful starting point for mining these data for information relevant to carbon cycling research. The Department of Energy's Joint Genome Institute is a valuable resource in this regard and already has embarked on the sequencing of numerous ecologically relevant organisms and communities, including those inhabiting soils and plant biomes (see sidebar, Analysis and Annotation at DOE's Joint Genome Institute, p. 23).

Data Integration and Linking Analysis and Experimentation

Once data are generated, researchers face the challenging task of integrating metagenomic, metatranscriptomic, and metaproteomic data with physical and biogeochemical data and ultimately relating them to carbon cycling or subsurface biogeochemical processes. Tools must be developed that can correlate biogeochemical parameters with genomic information and generate metabolic predictions based on incomplete genomic, transcriptomic, and proteomic data. Databases such as IMG (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi) and CAMERA (http://camera.calit2.net) exist for comparative analyses of metagenomic data, and initial efforts

should be actively encouraged to integrate metadata into these frameworks. Metatranscriptomic and metaproteomic projects, on the other hand, are still very much individual laboratory efforts in need of more advanced tools to facilitate comparisons with metagenomic data and with data from other environments and to make the data and information available to the broader research community.

A potentially successful approach for connecting gene sequence to function centers on the concept of intensively characterizing keystone genes and organisms. This approach, for example, could involve bringing relevant, experimentally tractable organisms into the laboratory for genomics and systems biology–type investigations. Eventually, research would move into the field to address fundamental questions such as, "Which genes function in the environment?" Needed are high-throughput methods that are sensitive but do not require high concentrations of biomass. Genomic and functional genomic approaches also can be used to gather information about which organismal processes are important in the environment and which data should be incorporated into models. Arguably, the number of keystone genes and organisms involved in carbon cycling and environmental remediation may be immense. For this approach to be successful, high-throughput analyses will need to be coupled with robust systems for data capture and data analysis that can be used to develop models of metabolism and regulation. How can massive volumes of high-throughput experimental systems biology data from ecological observations be automated to convert the data into a model that can be tested dynamically? This is a mathematics and computing problem—not about getting omic data simply because it can be obtained, but about using larger-scale models to drive development of data needed to populate models and increase their ability to predict.

Another approach for linking genomes to function would be to foster communication and data and information sharing among researchers in the metagenomics and general metaomics realms. A specific initial concept for advancing the dialog is mutual list building with intercomparison. Cultural exercises could be supported in which biologists itemize the metabolic- and biogeochemical-level information they can provide currently or will be able to provide in the near term to large-scale modelers. Scientists on the computational side of the carbon cycling or environmental remediation issues would construct lists of their own, reflecting their metabiogeochemical information needs. Overlap would be identified and concepts developed for iteration plus expansion of the intersection zone. This process can be viewed as a simple Venn diagram with growing disciplinary area coverage and increasing conjunction.

A next level of interaction could then be attained by leveraging gene expression. This process can be considered as classical annotation run both forward and backward for modeling purposes. Within the available environmental sequences, mapping of genes to enzymes remains largely incomplete. Laboratory experiments, however, with relatively simple, defined model systems can demonstrate at the metabolic level that certain key processes are active. Marine organisms that have been studied in this manner include cyanobacteria, diatoms, and other eukaryotes along with certain classes of heterotrophs. Metabolic pathways can be mapped in reverse to the active genes if they are not apparent from analysis of the sequences themselves. A subgenome is thus identified as containing an initial kernel of critical biogeochemical information. The means for accelerating this process are in fact related to the above discussion; simple list comparisons will pinpoint processes in which the required laboratory and field work can be performed quickly.

At a more challenging level, the entire sequence of data processing from genome to biogeochemical function may be viewed as a unified or potentially unifiable information sciences problem. Many individual steps already have been automated. Examples include genome reads leading to library development on the biological end of the spectrum and modular additions at the field, ecosystem, and global scales. In the near term, only automating the gaps in between will remain. The genome can be viewed as a vector of the most fundamental biogeochemical data, the transcriptome likewise, the proteome as an amino acid matrix, and the metabolome as a multidimensional space containing stoichiometries and rates. Integrating model assembly upward then becomes a matter of mathematically manipulating the resulting datasets from each stage. They may be configured in a relational manner. Standard matrix algebra is then applied to yield biogeochemical source-sink relationships. In fact, data arrays and their mathematical relationships constitute the most concise theoretical representation possible for global biotic systems.

Appendix 5
Summary List of Findings from Introduction

Finding 1: The emergence of systems biology as a research paradigm and approach to DOE missions is founded on the dramatic increase in the volume of data from a new generation of genomics-based technologies. Data management and analysis are critical to the viability of this approach.

Finding 2: The GTL program has several large and highly focused research efforts in, for example, systems biology, bioenergy, and genomics. Each area is investing in and dependent on rapidly growing capabilities for data resources and management, making the associated needs of each an ideal initial focus for GTL Knowledgebase development.

Finding 3: Development and use of the GTL Knowledgebase require a comprehensive, flexible policy and supporting programs that will meet GTL's current and emerging research needs.

Finding 4: Researchers require the integration of a wide range of high-volume data and a computational environment designed to support modeling, derivation of predictions, and exchange of data.

Finding 5: Systems biology is contingent on the ability to integrate and utilize a wide variety of types of data and computational technologies to systematically address a progression of problems leading to effective modeling of organisms.

Finding 6: The GTL Knowledgebase should lead to the creation of abstract models that demonstrate increasing correspondence with the underlying physical reality. These models would play increasingly important roles in addressing major applications of interest to DOE.

Finding 7: Other agencies and groups, most notably the National Institutes of Health, have developed integrated databases for studying organisms related to human diseases. These community-driven efforts have dramatically impacted biomedical research. A similar effort in systems biology for bioenergy, carbon cycling and biosequestration, and environmental remediation will significantly aid these DOE missions.

Finding 8: DOE's national laboratory enterprise, collective and individually, has developed much of the necessary infrastructure to rapidly deploy components of the GTL Knowledgebase. A concerted effort would be needed to integrate these elements.

Appendix 6
Bibliography

American Academy of Microbiology. 2004. *Systems Microbiology: Beyond Microbial Genomics.*

Arkin, A. 2008. "Setting the Standard in Synthetic Biology," *Nature Biotechnology* **26**(7), 771–74.

Azam, F., and F. Malfatti. 2007. "Microbial Structuring of Marine Ecosystems," *Nature Reviews Microbiology* **5**(10), 782–91.

Azam, F., and A. Z. Worden. 2004. "Microbes, Molecules, and Marine Ecosystems," *Science* **303**(5664), 1622–24.

Bansal, M., et al. 2007. "How to Infer Gene Networks from Expression Profiles," *Molecular Systems Biology* **3**, 78.

Bare, J. C., et al. 2007. "The Firegoose: Two-Way Integration of Diverse Data from Different Bioinformatics Web Resources with Desktop Applications," *BMC Bioinformatics*, **8**(456).

Bonneau, R., et al. 2004. "Comprehensive De Novo Structure Prediction in a Systems-Biology Context for the Archaea *Halobacterium* Sp. *NRC-1*," *Genome Biology* **5**(8), R52.

Bonneau, R., et al. 2006. "The Inferelator: An Algorithm for Learning Parsimonious Regulatory Networks from Systems-Biology Data Sets De Novo," *Genome Biology* **7**(5), R36.

Bonneau, R., N. Baliga, et al. 2007. "A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell," *Cell* **131**(7), 1354–65.

Canton, B., A. Labno, and D. Endy. 2008. "Refinement and Standardization of Synthetic Biological Parts and Devices," *Nature Biotechnology* **26**(7), 787–93.

Cassman, M. 2005. "Barriers to Progress in Systems Biology," *Nature* **438**(7071), 1079.

Chiou, T. J., and D. R. Bush. 1998. "Sucrose is a Signal Molecule in Assimilate Partitioning," *Proceedings of the National Academy of Sciences of the United States of America* **95**(8), 4784–88.

DeJongh, M., et al. 2007. "Toward the Automated Generation of Genome-Scale Metabolic Networks in the SEED," *BMC Bioinformatics* **8**(139).

DeRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale," *Science* **278,** 680–686.

Dowell, R. D. 2001. "The Distributed Annotation System," *BMC Bioinformatics* **2**(7).

Eichenberger, P., et al. 2004. "The Program of Gene Transcription for a Single Differentiating Cell Type During Sporulation in *Bacillus subtilis*," *PLoS Biology* **2**(10), e328.

Facciotti, M. T., et al. 2004. "Systems Biology Experimental Design—Considerations for Building Predictive Gene Regulatory Network Models for Prokaryotic Systems," *Current Genomics* **5**(1), 527–44.

Faith, J. J., et al. 2007. "Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles," *PLoS Biology* **5,** e8.

Falkowski, P. G., T. Fenchel, and E. F. Delong. 2008. "The Microbial Engines That Drive Earth's Biogeochemical Cycles," *Science* **320**(5879), 1034–39.

Field, D., B. Tiwari, and J. Snape. 2005. "Bioinformatics and Data Management Support for Environmental Genomics," *PLoS Biology* **3**(8), 1352–53.

Follows, M. J., et al. 2007. "Emergent Biogeography of Microbial Communities in a Model Ocean," *Science* **315**(5820), 1843–46.

Fredrickson, J. K, et al. 2008. "Towards Environmental Systems Biology of *Shewanella*," *Nature Reviews Microbiology* **6**(8), 592–603.

Gerdes, S. Y., et al. 2006. "Comparative Genomics of NAD Biosynthesis in Cyanobacteria," *Journal of Bacteriology* **188**(8), 3012–23.

Goble, C., and R. Stevens. 2008. "State of the Nation in Data Integration for Bioinformatics," *Journal of Biomedical Informatics* **41**(5), 687–93.

Gorton, I., et al. 2008. "The MeDICi Integration Framework: A Platform for High Performance Data Streaming Applications," *Seventh Working IEEE/IFIP Conference on Software Architecture*, 95–104.

Gupta, N., et al. 2008. "Comparative Proteogenomics: Combining Mass Spectrometry and Comparative Genomics to Analyze Multiple Genomes," *Genome Research* **18**(7), 1133–42.

Hlavacek, W. S., et al. 2006. "Rules for Modeling Signal-Transduction Systems," *Science STKE* **2006**(334), re6.

Howe, D., and S. Y. Rhee. 2008. "The Future of Biocuration," *Nature* **455**, 47–50.

Jamshidi, N., and B. O. Palsson. 2008. "Formulating Genome-Scale Kinetic Models in the Post-Genome Era," *Molecular Systems Biology* **4(171)**.

Johnson, Z. I., E. R. Zinser, et al. 2006. "Niche Partitioning Among *Prochlorococcus* Ecotypes Along Ocean-Scale Environmental Gradients," *Science* **311**(5768), 1737–40.

Jones, M. B., et al. 2006. "The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere," *Annual Review of Ecology, Evolution, and Systematics* **37**, 519–44.

Kirschner, M. W. 2005. "The Meaning of Systems Biology," *Cell* **121**(4), 503–04.

Kitano, H. 2002. "Systems Biology: A Brief Overview," *Science* **295**(5560), 1662–64.

Klamt, S., et al. 2008. "Modeling the Electron Transport Chain of Purple Non-Sulfur Bacteria," *Molecular Systems Biology* **4(156)**.

Laub, M. T., et al. 2000. "Global Analysis of the Genetic Network Controlling a Bacterial Cell Cycle," *Science* **290**(5499), 2144–48.

Laws, E. A., et al. 2000. "Temperature Effects on Export Production in the Open Ocean," *Global Biogeochemical Cycles* **14**(4), 1231–46.

Liu, Y., J. Zhou, et al. 2003. "Transcriptome Dynamics of *Deinococcus radiodurans* Recovering from Ionizing Radiation," *Proceedings of the National Academy of Sciences of the United States of America* **100**(7), 4191–96.

Lynch, C. 2008. "Big Data: How Do Your Data Grow?" *Nature* **455**, 28–29.

Marx, J. 2004. "The Roots of Plant-Microbe Collaborations," *Science* **304**(5668), 234–36.

Masuda, N., and G. M. Church. 2003. "Regulatory Network of Acid Resistance Genes in *Escherichia coli*," *Molecular Microbiology* **48**, 699–712.

Nature. 2008a. "A Place for Everything," *Nature* **453,** 2.

Nature. 2008b. "Community Cleverness Required," *Nature* **455,** 1.

Osterman, A. L. 2006. "A Hidden Metabolic Pathway Exposed," *Proceedings of the National Academy of Sciences of the United States of America* **103**(15), 5637–38.

Osterman, A. L., and R. Overbeek. 2003. "Missing Genes in Metabolic Pathways: A Comparative Genomics Approach," *Current Opinion in Chemical Biology* **7**(2), 238–51.

Osterman, A. L., and T. P. Begley. 2007. "A Subsystems-Based Approach to the Identification of Drug Targets in Bacterial Pathogens." In *Systems Biological Approaches in Infectious Diseases,* 131–70. From the *Progress in Drug Research* series **64,** Birkhäuser Basel.

Overbeek, R., et al. 2005. "The Subsystems Approach to Genome Annotation and Its Use in the Project to Annotate 1000 Genomes," *Nucleic Acids Research* **33**(17), 5691–702.

Pinchuk, G., D. Rodionov, C. Yang, et al. 2009. "Genomic Reconstruction of *Shewanella oneidensis* MR-1 Metabolism Reveals a Previously Uncharacterized Machinery for Lactate Utilization," *Proceedings of the National Academy of Sciences of the United States of America,* published online before print February 5. DOI: 10.1073/pnas.0806798106.

Reiss, D. J., N. S. Baliga, and R. Bonneau. 2006. "Integrated Biclustering of Heterogeneous Genome-Wide Datasets for the Inference of Global Regulatory Networks," *BMC Bioinformatics* **7**(280).

Rodionov, D. A. 2007. "Comparative Genomic Reconstruction of Transcriptional Regulatory Networks in Bacteria," *Chemical Reviews* **107**(8), 3467–97.

Rodionov, D. A., et al. 2007. "Genomic Identification and In Vitro Reconstitution of a Complete Biosynthetic Pathway for the Osmolyte Di-Myo-Inositol-Phosphate," *Proceedings of the National Academy of Sciences of the United States of America* **104**(11), 4279–84.

Rodionov, D. A., et al. 2008a. "Transcriptional Regulation of NAD Metabolism in Bacteria: Genomic Reconstruction of NiaR (YrxA) Regulon," *Nucleic Acids Research* **36**(6), 2032–46.

Rodionov, D. A., et al. 2008b. "Transcriptional Regulation of NAD Metabolism in Bacteria: NrtR Family of Nudix-Related Regulators," *Nucleic Acids Research* **36**(6), 2047–59.

Rodionov, D. A., et al. 2009. "A Novel Class of Modular Transporters for Vitamins in Prokaryotes," *Journal of Bacteriology* **191**(1), 42–51.

Rohde, A., et al. 2007. "Gene Expression During the Induction, Maintenance, and Release of Dormancy in Apical Buds of Poplar," *Journal of Experimental Botany* **58**(15/16), 4047–60.

Shannon, P. T., et al. 2006. "The Gaggle: An Open-Source Software System for Integrating Bioinformatics Software and Data Sources," *BMC Bioinformatics* **7**(176).

Shetty, R. P., D. Endy, and T. F. Knight, Jr. 2008. "Engineering BioBrick Vectors from BioBrick Parts," *Journal of Biological Engineering* **2(5)**.

Slepchenko, B. M., et al. 2003. "Quantitative Cell Biology with the Virtual Cell," *Trends in Cell Biology* **13**(11), 570–76.

Smith, A. K., et al. 2007. "LinkHub: A Semantic Web System That Facilitates Cross-Database Queries and Information Retrieval in Proteomics," *BMC Bioinformatics* **8**(Suppl 3), S5.

Smith, B., et al. 2007. "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration," *Nature Biotechnology* **25**(11), 1251–55.

Stein, L. D. 2003. "Integrating Biological Databases," *Nature Reviews Genetics* **4**(5), 337–45.

Stein, L. D. 2008. "Towards a Cyberinfrastructure for the Biological Sciences: Progress, Visions and Challenges," *Nature* **9**, 678–88.

Taylor, C. F., et al. 2007. "The Minimum Information About a Proteomics Experiment (MIAPE)," *Nature Biotechnology* **25**(8), 887–93.

U.S. DOE. 2006. *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/b2bworkshop.shtml).

U.S. DOE 2008. *Carbon Cycling and Biosequestration: Integrating Biology and Climate Through Systems Science,* DOE/SC-108, U.S. Department of Energy Office of Science (http://genomicsgtl.energy.gov/carboncycle/).

Vaughn, M. W., G. N. Harrington, and D. R. Bush. 2002. "Sucrose-Mediated Transcriptional Regulation of Sucrose Symporter Activity in the Phloem," *Proceedings of the National Academy of Sciences of the United States of America* **99**(16), 10,876–80.

Yang, C., et al. 2006. "Comparative Genomics and Experimental Characterization of N-Acetylglucosamine Utilization Pathway of *Shewanella oneidensis*," *Journal of Biological Chemistry* **281**(40), 29,872–85.

Yang, C., et al. 2008. "Glycerate 2-Kinase of *Thermotoga maritima* and Genomic Reconstruction of Related Metabolic Pathways," *Journal of Bacteriology* **190**(5), 1773–82.

Ye, Y. Z., et al. 2005. "Automatic Detection of Subsystem/Pathway Variants in Genome Analysis," *Bioinformatics* **21**, i478–86.

Yun, W., et al. 1998. "X-Ray Imaging and Microspectroscopy of Plants and Fungi," *Journal of Synchrotron Radiation* **5**, 1390–95.

Descriptions of a Selected Sampling of Databases Having Relevance to the GTL Knowledgebase

Several existing databases have created effective systems for storing and analyzing genomic, metagenomic, proteomic, and other data. Having implemented successful data analysis tools, information management strategies, user capabilities, and architectures, these systems can provide viable examples of components envisioned for the GTL Knowledgebase. Moreover, many of these databases will provide important supplements to and links with other GKB capabilities. Descriptions of several such systems and their features follow.

Integrated Microbial Genomes (IMG, http://img.jgi.doe.gov). Developed through collaboration between the Department of Energy's (DOE) Joint Genome Institute (JGI) and the Biological Data Management and Technology Center at DOE Lawrence Berkeley National Laboratory, DOE's IMG is a data management, analysis, and annotation platform that enables the efficient comparative analysis of all complete public microbial genomes, draft or finished, produced at JGI and throughout the world. IMG currently integrates data from 4570 genomes (1155 bacteria, 56 archaea, 40 eukaryotes, 932 plasmids, and 2387 viruses), consisting of more than 4.9 million genes, with publicly available metabolic pathway collections and protein family information. IMG offer various tools for comparing genes, pathways, and functions across genomes; visualizing the physical distribution of genes within genomes; investigating the evolutionary history of genes; and developing user-defined pathways and functional categories to aid the analysis of poorly characterized genomes.

In addition to supporting the analysis of complete genomic sequences data from microbial isolates, the **Integrated Microbial Genomes with Microbiome Sampling (DOE IMG/M, http://img.jgi.doe.gov/m)** portal supports comparative analyses of more than 40 community sequences generated with various metagenomic sequencing technologies and data processing methods. IMG/M allows examination of profiles of functional annotations across microbial communities and isolate organisms of interest as well as analysis of strain-level heterogeneity within a species population in metagenomic data.

Phytozome (http://www.phytozome.net). This tool for green plant comparative genomics is a joint project of the DOE Joint Genome Institute and the Center for Integrative Genomics at the University of California, Berkeley. Phytozome provides access to nine sequenced and annotated green plant genomes, including poplar, grape, sweet sorghum, rice, soybean, green algae, moss, spikemoss, and the small flowering plant *Arabidopsis*. Clusters of orthologous and paralogous genes that represent the modern descendents of ancestral gene sets can be analyzed to explore genes associated with significant evolutionary events related to the development of green plants.

SEED (http://www.theseed.org). The SEED Project is a cooperative effort to annotate ever-expanding genomic data so researchers can conduct effective comparative analyses of genomes. Launched in 2003 by the Fellowship for Interpretation of Genomes (FIG), the project is one of several initiatives in ongoing development of data curation systems. SEED is designed to be used by scientists from numerous centers and with varied research objectives. As such, several institutions have since joined FIG in a consortium, including the University of Chicago, DOE's Argonne National Laboratory (ANL), the University of Illinois at Urbana-Champaign, and others. As one example, ANL has used SEED to develop the National Microbial Pathogen Data Resource. Other agencies and institutions have used the project to discover genome components and clarify gene functions such as metabolism. SEED also has enabled researchers to conduct comparative analyses of closely related genomes and has supported derivation of stoichiometric models to understand metabolic processes.

The SEED Project has been extended to support metagenomic samples and concomitant analytical tools. Moreover, the number of genomes being introduced into SEED is growing very rapidly. Building a framework to support this growth while providing highly accurate annotations is centrally important to SEED. The project's subsystem-based annotation strategy has become the technological foundation for addressing these challenges.

Proteomics Research Information Storage and Management (DOE PRISM, http://ncrr.pnl.gov/about/process.stm). The PRISM system enables the management and analysis of high-throughput proteomic data at DOE's Pacific Northwest National Laboratory (PNNL). The laboratory's proteomics capability, based on high-performance mass spectrometry, combines commercially available informatics systems with those developed at PNNL. Streams of raw data from 15 mass spectrometers are joined and analyzed via an automated workflow designed to transform raw instrument data into useful biological information, such as which proteins an organism produces under different growth and stress conditions.

PRISM—which contains more than 100,000 datasets and 1 billion mass spectra—manages data, metadata, and analysis workflows; maintains a sample request and queuing system, reports research results; and provides a user interface for data searches and queries. Since its initial development in 2000, DOE's PRISM has undergone several advancements, including (1) addition of a mass tag system, high-capacity storage servers, and a plug-in architecture for automated analysis tools; (2) improved peptide and protein identification; (3) use of archived data from DOE's Environmental Molecular Sciences Laboratory; and (4) automated interaction between computers for PRISM's data management system and liquid chromatography cart.

MicrobesOnline (http://www.microbesonline.org). Since 2003, this integrated Web portal has enabled scientists to conduct comparative genomic analyses. MicrobesOnline currently contains more than 700 complete microbial genome sequences and offers a suite of analysis tools, including (1) a multispecies genome browser, (2) operon and regulon prediction methods, (3) a combined gene and species phylogeny browser, (4) a gene expression data browser with expression profile searches, (5) a Gene Ontology browser, and (5) sequence analysis capabilities (e.g., sequence motif detection, motif searches, sequence alignment, and phylogeny reconstruction).

Funded by DOE, the database also offers resources for community annotation, integrates functional genomic data, and provides novel Web-based viewing and analysis tools for proteomic, gene expression microarray, and phenotype microarray data. Interactive heat maps allow users to compare microarray data for microbes under multiple stress conditions. Users also can analyze correlations between gene expressions from different experiments. Among the major new features of MicrobesOnline is the ability to search the data compendium for genes with expression profiles similar to those resulting from query profiles.

Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA, http://camera.calit2.net). The ability to explore large metagenomic datasets can enhance the research of microbial ecologists. Until recently, large-scale metagenomic analysis has been limited by the availability of computational resources that provide scientists with easy-to-use, scalable, and fully integrated Web frameworks. One such resource—CAMERA—features a rich data repository, associated bioinformatic tools, and cyberinfrastructure for conducting such analyses.

CAMERA was launched in 2007 as a collaboration between the University of California at San Diego's Calit2 division and the J. Craig Venter Institute (JCVI). With funding from the Gordon and Betty Moore Foundation, CAMERA contains 12 metagenomic datasets consisting of 14 million genomic fragment sequences. Genomic data are layered with associated geographical, temporal, and physicochemical metadata to assist in metagenomic analyses. Additional capabilities enable homology identifications using Basic Local Alignment Search Tool (BLAST), an algorithmic resource for sequence comparisons. Furthermore, CAMERA provides graphical tools for viewing sequence regions that indicate genomic conservation and divergence and for correlating such regions with environmental factors. This new platform allows microbial researchers to begin to analyze large-scale sampling and sequencing endeavors such as JCVI's Global Ocean Sampling expedition.

Comprehensive Microbial Resource (CMR, http://cmr.jcvi.org). Containing more than 600 sequenced prokaryotic genomes, the CMR database provides researchers with information on inter- and intragenomic relationships for comparative genomics, genome diversity, and evolutionary studies. CMR—which is operated by JCVI—enables a wide variety of data retrievals and offers scientists numerous analytical tools for exploring the system's prokaryotic genomes. These data retrievals can be based on different gene properties that include molecular weight, hydrophobicity, guanine-cytosine (GC) content, functional-role assignments, and taxonomy. The system also has special Web-based analysis tools for precomputed homology searches, whole-genome dot plots, batch downloads, and searches across genomes using various data types.

Since consistent annotation is essential for robust genomic comparisons, CMR features primary annotations, as assigned by GenBank, and secondary annotations provided by JCVI. In addition, CMR provides comprehensive views of genes and gene annotations, genome-level structures, pathway maps, codon usage tables, GC plots, the ability to generate and visualize whole-genome alignments between two bacteria, and tabulated summary data from both individual genomes and CMR's entire genome collection.

Pathema (http://pathema.jcvi.org). As the Web resource for JCVI's Bioinformatics Resource Center, Pathema provides detailed curation of six target pathogens: *Bacillus anthracis*, *Clostridium botulinum*, *Burkholderia mallei*, *Burkholderia pseudomallei*, *Clostridium perfringens*, and *Entamoeba histolytica*. Funded by the National Institute of Allergy and Infectious Disease (NIAID), the center is one of eight designed to support biodefense and infectious disease research. Initially developed at JCVI, Pathema is co-maintained by the Institute of Genome Sciences (IGS) at the University of Maryland School of Medicine.

The Pathema website is separated into four main taxonomic clades: *Bacillus*, *Burkholderia*, *Clostridium*, and *Entamoeba*, allowing developers to customize clade-specific sites to each research community's needs. Pathema's dataset includes Gene Ontology assignments, metabolic pathway identification, transporter characterization, and predicted ortholog analysis and identification. The center's overarching goal is to provide a core online resource to accelerate scientific progress in understanding, detecting, diagnosing, and treating several categories of NIAID priority pathogens and other agents involved in new and re-emerging infectious diseases. Bioinformatics software with significant new capabilities, novel data types, Web resources, and analysis tools specifically geared toward biodefense are available on Pathema. Such capabilities—including intergenomic comparisons—help identify potential targets for vaccine development, therapeutics, and diagnostics. The site also serves as a focal point for the biodefense research community by disseminating data from bacterial genome-sequencing projects and by providing access to results of intergenomic comparisons.

Appendix 8

Genomics:GTL Systems Biology Knowledgebase Workshop:

Agenda, Participant List, and Biosketches

Agenda

Preworkshop Assignments		
Participants will come with a 1- to 2-page position paper addressing particular charge questions.		
Participants will come with a 5- to 7-minute presentation.		
Tuesday, May 27, 2008		
Time	**Presentation**	**Speakers**
7:30 – 9:00 p.m.	Registration and Refreshments	
Wednesday, May 28, 2008		
8:00 – 8:05 a.m.	Welcome and Introductions	Susan K. Gregurick, Program Manager, Biological Systems Science Division, DOE Office of Biological and Environmental Research
8:05 – 8:20 a.m.	Introduction and OBER Overview	Anna Palmisano, Associate Director, DOE Office of Biological and Environmental Research
8:20 – 8:30 a.m.	Introduction and GTL Overview	David Thomassen, Acting Director, DOE Life and Medical Sciences Division, Office of Biological and Environmental Research
8:30 – 9:00 a.m.	Overview of Knowledgebase Workshop	Rick Stevens, Associate Laboratory Director for Computing and Life Sciences, Argonne National Laboratory, and University of Chicago
9:00 – 9:30 a.m.	Systems Analysis of Microbial Solar Energy Capture and Utilization	Tim Donohue, Great Lakes Bioenergy Research Center
9:30 – 10:00 a.m.	Nitrogen Regulatory Networks and Plant Systems Biology	Manpreet Katari, New York University
10:00 – 10:30 a.m.	Systems Biology Overview	Andrei Osterman, Burnham Institute
10:30 – 10:45 a.m.	Break	
11:00 – 2:30 p.m.	Science Drivers Breakouts: Basic Research Needs and Use Cases	
	• Systems Biology for Bioenergy Solutions	Brian Davison, Paul Adams, and Tim Donohue, DOE Bioenergy Research Centers
	• Systems Biology for Carbon Cycle Understanding	Mick Follows, Massachusetts Institute of Technology Grant Heffelfinger, Sandia National Laboratories
	• Systems Biology Core	Nitin Baliga, Institute for Systems Biology Andrei Osterman, Burnham Institute
2:30 – 3:00 p.m.	Reconvene as Larger Group to Report on Basic Research Needs and Use Cases	

3:30 – 3:45 p.m.	Break	
Time	**Presentation**	**Speakers**
3:45 – 6:00 p.m.	Science Drivers Breakouts: Path Forward	
	• Systems Biology for Bioenergy Solutions	Brian Davison, Paul Adams, and Tim Donohue, DOE Bioenergy Research Centers
	• Systems Biology for Carbon Cycle Understanding	Mick Follows, Massachusetts Institute of Technology
	• Systems Biology Core	Nitin Baliga, Institute for Systems Biology Andrei Osterman, Burnham Institute
6:00 p.m.	Group Dinner and Writing	
Thursday, May 29, 2008		
8:30 – 9:00 a.m.	Introduction and ASCR Overview	Michael Strayer, Associate Director, DOE Office of Advanced Scientific Computing Research
9:00 – 9:30 a.m.	CaBIG Overview	Cheryl Marks, National Institutes of Health
9:30 – 10:00 a.m.	European Efforts in Knowledgebase Development	Dawn Field, Natural Environment Research Council, Oxford, U.K.
10:00 – 10:15 a.m.	Break	
10:15 – 11:00 a.m.	Report out by Chairs on Day 1 Science Drivers	Chairs of Science Driver Breakouts
11:00 – 2:30 p.m.	Technology Driver Breakouts: Charge Questions	
	• Knowledgebase Community and User Issues	Owen White, University of Maryland Jennifer Reed, University of Wisconsin
	• Data, Metadata, and Information	Denise Schmoyer, Oak Ridge National Laboratory Gordon Anderson, Pacific Northwest National Laboratory
	• Architecture and Technology	Folker Meyer, Argonne National Laboratory Ed Uberbacher, Oak Ridge National Laboratory
	• Data Integration	Ross Overbeek, FIG, Argonne National Laboratory Nagiza Samatova, Oak Ridge National Laboratory and North Carolina State University
2:30 – 3:00 p.m.	Reconvene as larger Group to Report on Charge Questions	
3:30 – 3:45 p.m.	Break	

Participant List

***Paul Adams**
Lawrence Berkeley National Laboratory

****Gordon Anderson**
Pacific Northwest National Laboratory

Rolf Apweiler
European Bioinformatics Institute

***Nitin Baliga**
Institute for Systems Biology

Jacek Becla
Stanford Linear Accelerator Center

Richard Bonneau
New York University, Courant

Olga Brazhnik
National Institutes of Health

Thomas Brettin
Los Alamos National Laboratory

C. Robin Buell
Michigan State University

Dylan Chivian
Lawrence Berkeley National Laboratory

Bob Cottingham
Oak Ridge National Laboratory

Terence Critchlow
Pacific Northwest National Laboratory

***Brian Davison**
Oak Ridge National Laboratory

Matt DeJongh
Hope College

Patrik D'haeseleer
Lawrence Livermore National Laboratory

***Tim Donohue**
University of Wisconsin, Madison

Scott Elliott
Los Alamos National Laboratory

Dawn Field
Centre of Ecology and Hydrology

***Michael (Mick) Follows**
Massachusetts Institute of Technology

Peg Folta
Lawrence Livermore National Laboratory

James K. Fredrickson
Pacific Northwest National Laboratory

Damian Gessler
National Center for Genome Resources

Stephen Goff
University of Arizona

Yakov Golder
Lawrence Berkeley National Laboratory

Ian Gorton
Pacific Northwest National Laboratory

Grant Heffelfinger
Sandia National Laboratories

Tatiana Karpinets
Oak Ridge National Laboratory

Manpreet Katari
New York University

Ken Kemner
Argonne National Laboratory

Les Klimczak
Great Lakes Bioenergy Research Center

Cheryl Kuske
Los Alamos National Laboratory

Mary Lipton
Pacific Northwest National Laboratory

Michael Lomas
Bermuda Institute of Ocean Sciences

John L. Markley
University of Wisconsin, Madison

Cheryl Marks
National Cancer Institute

Celeste Matarazzo
Lawrence Livermore National Laboratory

Raymond McCord
Oak Ridge National Laboratory

Lee Ann McCue
Pacific Northwest National Laboratory

Peter McGarvey
Georgetown University Medical Center

****Folker Meyer**
Argonne National Laboratory

Bob Morris
University of Washington

Sean Murphy
J. Craig Venter Institute

Ilya Nemenman
Los Alamos National Laboratory

Gary J. Olsen
University of Illinois

***Andrei Osterman**
Burnham Institute for Medical Research

****Ross Overbeek**
Argonne National Laboratory

George N. Phillips, Jr.
University of Wisconsin, Madison

David Pletcher
Joint BioEnergy Institute

****Jennifer Reed**
University of Wisconsin, Madison

****Nagiza F. Samatova**
Oak Ridge National Laboratory

* DOE Mission Session Cochair.
** Knowledgebase Session Cochair.

Denise Schmoyer
Oak Ridge National Laboratory

Blake A. Simmons
Joint BioEnergy Institute

Tom Slezak
Lawrence Livermore National Laboratory

Rick Stevens
Argonne National Laboratory

Michael R. Sussman
University of Wisconsin, Madison

Ronald Taylor
Pacific Northwest National Laboratory

Gerald A. Tuskan
Oak Ridge National Laboratory

Edward C. Uberbacher
Oak Ridge National Laboratory

Owen White
University of Maryland, College Park

John Wooley
University of California, San Diego

Alex Worden
Monterey Bay Aquarium Research Institute

Cathy Wu
Georgetown University

Liming Yang
National Center for Research Resources

William S. York
University of Georgia

Biosketches

Paul D. Adams
Lawrence Berkeley National Laboratory
Paul Adams studied biochemistry at Edinburgh University where he received a doctorate in structural biology in 1992. Adams is a senior scientist and deputy director of the Physical Biosciences Division at Lawrence Berkeley National Laboratory, head of the Berkeley Center for Structural Biology, vice president for technology at the Joint BioEnergy Institute, and an adjunct professor in the bioengineering department at the University of California, Berkeley. His current research interests span computation, structural biology, and biofuels. Much of Adams's research is focused on developing new algorithms and computational methods for addressing problems in structural biology. He also leads development of the technology portal for the Protein Structure Initiative Knowledge Base.

Gordon Anderson
Pacific Northwest National Laboratory
Gordon Anderson has more than 30 years' experience in developing systems for instrument control, high-performance data acquisition, and data management. His experience and skills have been applied to high-throughput proteomic research at Pacific Northwest National Laboratory (PNNL). Proteomics produces large volumes of multidimensional data that must be organized and processed using a combination of commercial software and custom-designed tools. Anderson has assembled a multidisciplinary team at PNNL where he has led development of proteomic data management and analysis. The development of hardware and software has enabled advanced instrument control schemes for state-of-the-art, high-performance mass spectrometers at the Environmental Molecular Sciences Laboratory located at PNNL. Anderson's efforts in software development have enabled proteomics capabilities in the areas of complex spectral analysis and feature detection.

The informatics group at PNNL consists of 12 staff members responsible for data management and knowledge extraction from the raw data resulting from analysis of biological samples. Anderson holds 2 R&D 100 awards and 7 patents and has authored or coauthored more than 100 journal articles. He received his bachelor's degree in electrical engineering from Washington State University in 1985.

Rolf Apweiler
European Bioinformatics Institute
Rolf Apweiler studied biology in Heidelberg, Germany, and Bath, United Kingdom. He worked 3 years in drug discovery in the pharmaceutical industry and has been involved in bioinformatics since 1987. Apweiler started his bioinformatics career working on Swiss-Prot at the European Molecular Biology Laboratory (EMBL) in Heidelberg. He joined EMBL's European Bioinformatics Institute (EBI) in Hinxton, United Kingdom, in 1994 and is now joint head of the Protein and Nucleotide Data Group (PANDA) at EBI (http://www.ebi.ac.uk/panda/). This group coordinates UniProt activities, InterPro, GOA, Reactome, PRIDE, IntAct, Ensembl, the EMBL nucleotide sequence database, and other projects at EBI.

Nitin Baliga
Institute for Systems Biology
Nitin Baliga received a master's degree in marine biotechnology from Goa University, India, and has a doctorate in microbiology from the University of Massachusetts, Amherst. He is an associate professor at the Institute for Systems Biology where he leads a multidisciplinary team in deciphering quantitative systems-scale models for complete gene regulatory circuits of diverse prokaryotic organisms. With a special focus on organisms such as *Halobacterium salinarum* NRC-1, Baliga's long-term goal is to tap into the extraordinary biological potential of extremophiles.

Jacek Becla
Stanford Linear Accelerator Center
Jacek Becla earned a master's degree in electronics engineering from the University of Science and Technology in Poland in 1995. He joined the Stanford Linear Accelerator Center in 1997 as an information systems specialist. Becla's primary expertise is developing systems for managing very large datasets, and he leads efforts related to architecting the petabyte database for the Large Synoptic Survey Telescope astronomical survey. Becla also organizes XLDB (Extremely Large Databases) workshops and helps coordinate the open-source science database, SciDB. In the past, he was one of the main designers of the BaBar database and was the manager of the BaBar database group.

Richard Bonneau
New York University
Richard Bonneau is a joint assistant professor in both the New York University biology department and the computer science department at the Courant Institute for Mathematical Sciences. He also serves as an affiliate faculty member at the Institute for Systems Biology in Seattle, Washington. Bonneau is the technical lead on two grid-computing collaborations with IBM—the first and second phases of

the Human Proteome Folding Project. Rich also oversees TACITUS's (http://www.tacitus.com) approach to data gaming for all applications that focus on genomics, computational biology, and cell biology and is part of multiple groups developing open-source tools for data visualization. His research focuses on three main topics: (1) learning dynamical regulatory and signaling networks from functional genomic data, (2) using state-of-the-art structure prediction and design methodologies (e.g., Rosetta) to predict protein function and to design new functions, and (3) conducting multiple data and multiple species biclustering (data integration). All these activities are united by a common motivation: developing novel computational tools that extract genome-wide mechanistic models from large functional genomic datasets.

Olga Brazhnik
National Institutes of Health
Olga Brazhnik is a program manager at the National Center for Research Resources within the National Institutes of Health. She started her career as a physicist applying theoretical and computational methods in biology and medicine. In 1993, she was awarded a research grant by the U.S. National Research Council and Academy of Sciences and joined the James Franck Institute at the University of Chicago. Brazhnik then took a position at Virginia Tech and in 1998 transitioned into information technology, searching for capabilities to enable effective transformation of abundant scientific data into knowledge. In 2000, she joined the Virginia Bioinformatics Institute where her work resulted in the creation of several bioinformatics databases (e.g., ESTAP, DOME, and SeedGenes). In 2002, Brazhnik became the chief database architect for the Epidemic Outbreak Surveillance Project and later for the COHORT project on real-time integration of clinical systems with the U.S. Air Force Surgeon General Office. Her work involved integrating clinical and biological data; designing and developing database systems in Oracle, SQL-2000, PostgreSQL, and Access; and participating in development of protocols for study design and data collection, analysis of microarray data, and implementation of MIAME and HL7 standards. Brazhnik joined the National Institutes of Health in 2004, and she is an affiliate associate professor at George Mason University.

Thomas Brettin
Los Alamos National Laboratory
Thomas Brettin is the bioinformatics team leader in the genome sciences group at Los Alamos National Laboratory (LANL). He currently is serving on a change of station to the Department of Energy (DOE) Joint Genome Institute's Production Genomics Facility where he works as a software systems architect. Brettin has a master's degree in genetics; his more than 15 years' experience in genomics includes hands-on work and leadership roles in high-throughput sequencing laboratory automation, sequence annotation and analysis, software engineering, and information technology. Brettin is the principal investigator for a 5-year pathogen sequencing project funded by the Office of the Chief Scientist (formerly the Intelligence Technology Innovation Center). He also is principal investigator of the oral pathogens database project, now in its eighth year of funding as an interagency agreement among the National Institute for Dental and Craniofacial Research, the National Institutes of Health (NIH), and DOE. Brettin is a member of the Information Science and Technology Center's science council at LANL; serves on the scientific advisory board for the viral bioinformatics resource funded by NIH's National Institute of Allergy and Infectious Diseases; and is a longtime member of the scientific leadership team within LANL's Bioscience Division. He came to LANL and the DOE Joint Genome Institute from the Whitehead Institute/MIT Center for Genome Research (now the Broad Institute) where he retrained in computer science by taking night classes at Boston University. He became a software architecture professional from the Software Engineering Institutes at Carnegie Mellon University in 2005 and has more than 10 years of experience in software engineering. Brettin has taught computer science at the University of New Mexico, Los Alamos, since fall 2000 and has received several distinguished performance awards at LANL.

C. Robin Buell
Michigan State University
C. Robin Buell is an associate professor of plant biology at Michigan State University in East Lansing, Michigan. Buell joined Michigan State from the Institute for Genomic Research in Rockville, Maryland, where she was on the faculty for nearly 9 years. She has been involved in the genome sequencing of *Arabidopsis*, rice, and potato and led the sequencing effort for *Pseudomonas syringae*. Her current research focuses on plant and plant-pathogen genomics. Research projects in her group include annotation of the rice genome, potato sequencing and annotation, comparative sequencing of *Pythium ultimum,* and development of a comprehensive database for plant-pathogen genome sequences. Components of these projects—which are funded through several federal grants—include the generation of public resources such as large-scale sequence and annotation data, as well as bioinformatics resources like databases and data-mining websites for the greater scientific community.

Dylan Chivian
Lawrence Berkeley National Laboratory
Dylan Chivian received his doctorate from the University of Washington where he worked with David Baker on methods for protein structure prediction, including the creation of the Robetta server for protein structure prediction. He conducted his postdoctoral work with Adam Arkin at Lawrence Berkeley National Laboratory (LBNL) where he studied environmental and comparative genomics of bacteria and archaea, discovering the first single-species ecosystem deep within the Earth. Chivian is currently a scientist at LBNL where he leads the bioinformatics team for the Department of Energy Joint BioEnergy Institute, studying ways of engineering microbes to adopt capabilities ordinarily accomplished by communities in nature.

Bob Cottingham
Oak Ridge National Laboratory
Bob Cottingham is one of the pioneers of bioinformatics. In the 1970s, he began his career as a software developer on some of the first programs for genetic linkage analysis applied to mapping human disease traits. In 1989, Cottingham became directeur informatique at the Centre d'Etude du Polymorphisme Humain (CEPH) in Paris. There he oversaw the database of CEPH family genotypes, a resource ultimately used by more than 1000 laboratories in an international consortium to construct the first genetic maps of the human genome. Cottingham then joined the U.S. Human Genome Project, first as codirector of the Informatics Core in the Baylor College of Medicine Human Genome Center, then as operations director of the Genome Database at Johns Hopkins University School of Medicine. Subsequently, he became vice president of computing at Celltech Chiroscience, a biopharmaceutical company in the United Kingdom that develops drugs based on gene targets. In 2000, he cofounded Vizx Labs, a bioinformatics company that developed GeneSifter, the first Web-based gene expression microarray analysis service now used worldwide by hundreds of laboratories. In 2008, Cottingham joined Oak Ridge National Laboratory where he leads the computational biology and bioinformatics group currently working on projects for the Department of Energy's BioEnergy Science Center and Genomics:GTL program.

Terence Critchlow
Pacific Northwest National Laboratory
Terence Critchlow is the chief scientist and associate division director for scientific data management in the Computational Sciences and Mathematics Division at Pacific Northwest National Laboratory (PNNL). Critchlow earned his bachelor of science degree from the University of Alberta in 1990; he received his master's and doctorate in computer science from the University of Utah in 1992 and 1997, respectively. Critchlow worked at Lawrence Livermore National Laboratory (LLNL) from 1997 to 2007, spending time as a postdoc, individual contributor, and principal investigator. He led several projects while at LLNL, such as data management efforts supporting the Advanced Simulation and Computing program and several Department of Homeland Security (DHS) programs, including the Biodefense Knowledge Center. Critchlow joined PNNL in April 2007. He currently is the technical group manager for the Scientific Data Management group, is thrust area lead for the Scientific Process Automation area within the Department of Energy's SciDAC Scientific Data Management Center, and is a principal investigator for a DHS S&T data management and analysis project. Critchlow's current research interests are data analysis, data integration, metadata, and large-scale data management.

Brian Davison
Oak Ridge National Laboratory
Brian Davison is chief scientist for systems biology and biotechnology at Oak Ridge National Laboratory (ORNL), and in fall 2009, he will begin serving as chief scientist for the Department of Energy's (DOE) Genomics:GTL program. Davison is a deputy lead in the recently awarded DOE BioEnergy Science Center (http://www.bioenergycenter.org). He also is an adjunct professor of chemical engineering at the University of Tennessee. Davison recently served 2 years as director of ORNL's Life Sciences Division, and he previously was a Distinguished Researcher and BioChemical Engineering Research group leader. In his 24 years at ORNL, Davison has performed biotechnology research in a variety of areas, including bioconversion of renewable resources (e.g., ethanol, organic acids, and solvents); nonaqueous biocatalysis; systems analysis of microbes (cultivation and proteomics); biofiltration of volatile organic compounds; mixed cultures; immobilization of microbes and enzymes; metal biosorption; and extractive fermentations. The theme connecting his work is life at the interface of solid, liquid, and gas phases between biocatalysts and their environments, and this research has resulted in 100 publications and 6 patents. Davison received his doctorate in chemical engineering from the California Institute of Technology and his bachelor's degree in chemical engineering from the University of Rochester.

He led a multilaboratory team that in 1997 received an R&D 100 Award for "Production of Chemicals from Biologically Derived Succinic Acid." Davison also cochaired

the 15th to 26th Symposia on Biotechnology for Fuels and Chemicals, served as editor of *Proceedings in Applied Biochemistry and Biotechnology* from 1994 to 2005, and received the 2006 C.D. Scott Award by the Society of Industrial Microbiology. Davison has served as chairman of the ORNL Institutional Biosafety Committee from 2001 to present, and he was named a fellow in the American Institute for Medical and Biological Engineering in 2006.

Matt DeJongh
Hope College
Matt DeJongh received his doctorate in artificial intelligence from Ohio State University. He worked as a senior software engineer in the bioinformatics software industry before joining the faculty of Hope College in Holland, Michigan, where he is an associate professor of computer science. DeJongh is active in bioinformatic research with undergraduate students at Hope College in reconstructing and modeling cellular metabolic systems.

Patrik D'haeseleer
Lawrence Livermore National Laboratory
Patrik D'haeseleer received a master's degree in electrical engineering from Ghent University in Belgium, a master's in computer science from Stanford University, and a doctorate in computer science from the University of New Mexico. His research includes metabolic and regulatory networks, large-scale comparative genomics, and metagenomics. D'haeseleer currently is a research scientist in the Microbial Systems Biology Group at Lawrence Livermore National Laboratory. He also is part of the microbial communities team at the Department of Energy Joint BioEnergy Institute where he studies metabolic processes in natural biomass-degrading microbial organisms and communities.

Tim Donohue
University of Wisconsin, Madison
Tim Donohue has a bachelor of science degree from Polytechnic Institute of Brooklyn and earned a master's degree and doctorate from Pennsylvania State University in1977 and 1980, respectively. Donohue has been a member of the bacteriology department at the University of Wisconsin, Madison, for more than 20 years. In 2007, he was named director of the Department of Energy's Great Lakes Bioenergy Research Center.

Scott Elliott
Los Alamos National Laboratory
Scott Elliott began his career as a laboratory marine chemist then shifted to atmospheric photochemistry and aerosol microphysics modeling in the 1990s. In this role, he participated in regional simulations of megacity and Asian air pollution and was involved in elucidation of heterogeneous reactions within the Antarctic ozone hole. After joining Los Alamos National Laboratory, Elliott worked on various defense- and security-oriented environmental chemistry projects, including studies of plume composition of boost phase missiles, degradation of nerve agents in urban atmospheres, and hyperspectral infrared imaging for remote identification.

In the late 1990s, the opportunity arose for Elliott to contribute his modeling skills to a team developing an ultrafast, fine-resolution marine general circulation model—the Parallel Ocean Program (POP). Elliott introduced global biogeochemistry modules into the code and now specializes in simulation of geocycling for dissolved, climate-relevant trace gases. Demonstrations have included the computation of total marine distributions for methane, nitrous oxide, nonmethane hydrocarbons, and organohalogens. Development has progressed farthest with mechanisms for dimethyl sulfide and carbon monoxide, which influence tropospheric cloud nucleus and ozone fields, respectively. Over the past decade, POP has evolved into the core ocean model in the primary U.S. Earth System simulator—the Department of Energy and National Science Foundation's Community Climate System Model (CCSM). Elliott's trace gas mechanisms are now running within CCSM in a coupled surface ocean–to–atmosphere mode, both for preindustrial and contemporary situations. Simulations of the upcoming period of global warming are now under way, and projects planned for the medium term involve incorporating polar ice algal biogeochemistry and global bacterial population dynamics into the CCSM framework.

Dawn Field
Natural Environment Research Council's Centre for Ecology and Hydrology
Dawn Field is head of the Molecular Evolution and Bioinformatics section of the Natural Environment Research Council's (NERC) Centre for Ecology and Hydrology. She is principal investigator on a NERC project to develop a new genomic data standard to capture a richer set of information about genome sequences. She also is principal investigator on a NERC-funded effort to understand the evolution and function of microsatellites in microbial species. One outcome of this project thus far is Msatfinder, a Perl script designed to allow the identification and characterization of microsatellites in a comparative genomic context. In addition, Field is coinvestigator on the Marine Metagenomics project, which is being undertaken by an integrated consortium of United Kingdom microbiologists who will use postgenomics to investigate aquatic microbial

assemblages that control biogeochemical cycles. She also is participating in the Floral Genome Project, which aims to investigate the origin, conservation, and diversification of the genetic architecture of the flower and to develop conceptual and real tools for evolutionary functional genomics in plants. Field is director and founding member of the NERC Environmental Bioinformatics Centre, which provides bioinformatic and data management solutions for environmental genomic research.

Michael (Mick) Follows
Massachusetts Institute of Technology
Michael (Mick) Follows received his doctorate from the University of East Anglia, United Kingdom, in 1991 and is a senior research scientist in the Department of Earth, Atmospheric, and Planetary Sciences at the Massachusetts Institute of Technology. He uses idealized and numerical models to explore and better understand the interactions of ocean circulation, chemistry, and biology that regulate the productivity of the oceans and marine biogeochemical cycles of key elements, including carbon. His recent work focuses on the relationship of marine microbial communities and their environment.

Peg Folta
Lawrence Livermore National Laboratory
Peg Folta is the associate department head for Computing in Biology at Lawrence Livermore National Laboratory. Bioinformatics and computational biology research within the department are focused primarily on bioenergy and biodefense. Large-scale genomic and proteomic analyses are designed to predict function, identify and characterize unique regions, and determine metabolic pathways. Large-scale data integration and automated high-throughput sequencing also are emphasized. In recent years, Folta was the interim department head at the Department of Energy's Joint Genome Institute and was leader of the computational biology thrust area within the Chemical and Biological Countermeasures Program. She received her master's degree in applied mathematics from the University of Missouri, Rolla, and her bachelor of science in mathematics degree at Truman State University.

James K. Fredrickson
Pacific Northwest National Laboratory
James K. Fredrickson specializes in microbial ecology and environmental microbiology. He received a master's degree in soil chemistry in 1982 and a doctorate in soil microbiology in 1984 from Washington State University. Since joining Pacific Northwest National Laboratory (PNNL) in 1985, he has focused his research efforts in subsurface microbiology and biogeochemistry. Fredrickson has been responsible for laboratory and field research programs investigating the microbial ecology and biogeochemistry of geologically diverse subsurface environments and is recognized nationally and internationally for these efforts. He also has served as subprogram coordinator for the Department of Energy's (DOE) Subsurface Science Program from 1991 to the present. In this role, Fredrickson coordinated the technical aspects of DOE's deep subsurface microbiology subprogram at the national level and assisted DOE program managers in setting programmatic research directions. This subprogram involved more than 15 projects at universities and national laboratories nationwide and focused on multidisciplinary, field-scale research. At the request of DOE, he currently serves as the national coordinator for the *Shewanella* Microbial Cell Project, part of DOE's Genomics:GTL program. Additionally, Fredrickson was appointed chief scientist for GTL in 2005 and serves as a spokesman for the program to the scientific community. He is a senior chief scientist (laboratory fellow, Level VI) within the Biological Sciences Division, Fundamental and Computational Sciences Directorate, at PNNL.

Damian Gessler
National Center for Genome Resources
Damian Gessler earned degrees in biology and mathematics at Beloit College, Wisconsin, and received his doctorate in population genetics from the University of California, Santa Cruz. Gessler's biological expertise is in evolution and population genetics, as studied via computational techniques. He has used these skills to delineate conditions favorable for the evolution of recombination and meiosis and to quantify the rate of Muller's ratchet in populations unable to achieve mutation-selection balance. Gessler continues research in the evolution of recombination. His informatics expertise is in simulation, modeling, and data integration, and he has more than 20 year of experience in computer programming and systems operations. Gessler's recent work focuses on the challenges of integrating data and services from across the Web in a semantic Web architecture. This complements ongoing work to build a new class of data-driven simulation designs aimed at constructing better predictive models.

Stephen Goff
University of Arizona
Stephen Goff received his bachelor's degree in biology from the University of California, Santa Cruz, in 1978; he earned a doctorate from Harvard University in 1985. His graduate research focused on cell and molecular physiology, and his research training involved molecular genetics of bacteria and bacteriophage, molecular biology, and mammalian cellular

physiology. Goff continued research at Biogen Inc. in Cambridge, Massachusetts, and Geneva, Switzerland, and then joined Tuft's Medical School as a research associate where he focused on transcriptional control mechanisms in mammalian cells. In 1997, he shifted his research focus to gene expression in plants at the Plant Gene Expression Center, a collaboration between the U.S. Department of Agriculture and the University of California, Berkeley. Goff continued this research at the Institute for Molecular Biology at the University of Oregon; he then joined Ciba Biotechnology in Research Triangle Park, North Carolina, in 1992 as a senior scientist. After building up a group involved in gene discovery in plant and animal systems, Goff continued his research in gene discovery and function as director of genome technology at the Torrey Mesa Research Institute, a subsidiary of Novartis/Syngenta in San Diego, California, originally funded by the Novartis Foundation. He initiated and led a large effort to improve genomics technologies to better understand both model and crop plants, especially *Arabidopsis* and rice. This effort resulted in Goff being awarded Research Leader of the Year by *Scientific American* magazine in 2002. From 2003 to 2007, he worked with corporate business development at Syngenta as a senior Syngenta Fellow and senior technical analyst. Goff then became science advisor for Syngenta's corn and soybean business and focused on molecular approaches to enhancing yield and understanding hybrid vigor. Goff also advised Syngenta's vegetable business on appropriate scientific targets for vegetable improvement. At the end of April 2008, Goff left Syngenta and joined the iPlant Collaborative (where he currently is project director) at the University of Arizona's BIO5 Institute.

Yakov Golder

Lawrence Berkeley National Laboratory

Yakov Golder joined the Department of Energy Joint Genome Institute (JGI) in April 2007 and oversaw the Informatics Department until fall 2008. Golder received his bachelor's degree in computer science from City College of New York and a master's degree in computer science from the New York Institute of Technology. He has more than 20 years of technical leadership experience at both established and startup companies in the delivery of complex, high-performance software applications for social networking, workflow management, investment management, customer relationship management; and health care. Prior to joining JGI, Golder served as vice president of technology at CNET Networks where he oversaw the engineering organization in the Online Community Division. There he was responsible for the high-performance photosharing website (http://www.webshots.com), which boasted more than 1 billion monthly page views. Prior to his work at CNET, Golder was responsible for two critical areas of eBay's complex Web infrastructure: application data persistence and messaging. Golder's experience in designing enterprise-class software systems for the private and public sector builds upon previous efforts in creating both software-as-a-service websites and more traditional software product development in the engineering, industrial automation, and online communities markets.

Ian Gorton

Pacific Northwest National Laboratory

Ian Gorton is the associate division director in the Computational Sciences and Mathematics Division at Pacific Northwest National Laboratory (PNNL). Gorton also serves as the chief architect for PNNL's Data Intensive Computing Initiative. Gorton received his doctorate in computer science from Sheffield Hallam University, United Kingdom, in 1988. Before coming to PNNL, from March 2004 to July 2006 he led software architecture research and development at National Information and Communications Technology Australia in Sydney. Gorton was PNNL's chief architect in Information Sciences and Engineering, a group of more than 200 software developers who created applications that ranged from full-production deployments to advanced research prototypes and demonstrators. Gorton was responsible for infusing component-based development approaches into Information Sciences and Engineering projects, promoting best-practice architecture designs and review processes, acting as technical lead on several key client projects, and pursuing an R&D agenda to develop new infrastructure technology for data integration and content-based messaging. In addition, he holds the position of honorary associate at the School of Information Technologies at the University of Sydney in Australia.

Grant S. Heffelfinger

Sandia National Laboratories

Grant S. Heffelfinger is deputy director for Materials Science and Technology in the Materials and Process Sciences Center at Sandia National Laboratories in Albuquerque, New Mexico. His graduate research in molecular physics led to a doctorate in chemical engineering from Cornell University in 1988. Since that time, Heffelfinger has held various staff and management positions at Sandia. His research achievements include coinventing the dual control volume molecular dynamics simulation method for modeling diffusion in molecular systems with chemical potential gradients, such as diffusion through membranes. Heffelfinger was the principal author and technical leader for "Accelerating Biology with Advanced Algorithms and Massively Parallel Computing,"

a cooperative research and development agreement between Sandia National Laboratories and Celera Genomics that was signed in January 2001. He also is the principal investigator in the Department of Energy Genomics:GTL project, Carbon Fixation in *Synechococcus* sp.: From Molecular Machines to Hierarchical Modeling, which is developing advanced computational biology tools and prototyping these tools to understand how marine cyanobacteria fix carbon.

Tatiana Karpinets
Oak Ridge National Laboratory
Tatiana Karpinets is a research scientist in the Computer Science and Mathematics Division at Oak Ridge National Laboratory. She also is an adjunct professor in the plant sciences department the University of Tennessee. Karpinets received a master's degree in biophysics from Kharkov State University in Ukraine. From 1991 to 2001, she worked as a research scientist and then as chief scientist on computational and mathematical modeling of biological systems in the All-Russian Scientific Research Institute of Agriculture. In 2002, Karpinets joined the physics department at Wright State University in Dayton, Ohio, as a postdoctoral research scientist. There she worked on two projects: bioinformatics support for toxicogenomics and simulating the interactions of genes, proteins, and metabolites in cell-like entities. Karpinets specializes in bioinformatics, computational biology, biostatistics, and mathematical modeling of biological systems. She has dozens of publications in these areas of research.

Manpreet Katari
New York University
Manpreet Katari is a postdoctoral fellow and manager of bioinformatics at New York University's plant systems biology laboratory. He received his bachelor's degree in biochemistry from State University of New York, Buffalo, in 1996 and his doctorate in genetics from State University of New York, Stony Brook, in 2004. Katari's research interests include systems biology, comparative genomics, and software and database development. Specifically, his research focuses on identifying networks of genes involved in regulating different metabolic pathways and development stages in *Arabidopsis*. Katari uses both computational and experimental methods to solve biological questions. He is participating in several software projects, including VirtualPlant (http://www.virtualplant.org), a system containing a set of data integration, analysis, and visualization tools to support plant systems biology investigations, and Vicogenta (VIewer for COmparing GENomes to *Arabidopsis*, http://www.vicogenta.org), a data-mining tool that allows users to simultaneously search sequence databases for multiple taxa to find closest matches to the *Arabidopsis* genome based on sequence similarity.

Ken Kemner
Argonne National Laboratory
Ken Kemner is leader of the Molecular Environmental Science (MES) Group at Argonne National Laboratory. He received his doctorate in physics from the University of Notre Dame in 1993. A main emphasis during creation and growth of the MES Group has been development of an internationally recognized and integrated multidisciplinary scientific team focused on investigating fundamental biogeochemical questions. Members of the group have expertise in areas such as high-energy X-ray physics, environmental chemistry, environmental microbiology, and radiolimnology. Additional expertise in geomicrobiology, electron microscopy, and X-ray microscopy often is provided by collaborations with scientists outside the MES group.

Kemner's group uses numerous analytical techniques (e.g., inductively coupled plasma atomic emission spectroscopy, high-performance liquid chromatography, ion chromatography, kinetic phosphorescence analysis, X-ray diffraction, and electron microscopy) to better understand the role of minerals, microbes, and microbial exudates in determining carbon and contaminant mobility and fate in the environment. The group also uses and develops several synchrotron-based X-ray techniques to advance scientists' understanding of processes occurring at physical, geological, chemical, and biological interfaces that determine fate and transport. Kemner's group has begun integrating metagenomic sequencing and bioinformatic approaches to understand microbial community evolution during biostimulation of terrestrial environments.

Les Klimczak
Great Lakes Bioenergy Research Center
Les Klimczak is the chief informatics officer at the Department of Energy's Great Lakes Bioenergy Research Center (GLBRC). Prior to his work at GLBRC, Klimczak was a research informatics consultant at several biotechnology and research organizations. He served as senior director of bioinformatics and information technologies at Psychiatric Genomics Inc., a genomics-based drug discovery company that develops and creates small-molecule drugs for the treatment of psychiatric diseases. Klimczak also was program coleader of bioinformatics at Akkadix Corporation, an agricultural biotechnology company that uses functional genomics, bioinformatics, and other approaches to discover novel plant genes and agrochemicals. In addition to bioinformatics and information technology, Klimczak's

expertise includes chemoinformatics, genomics, data mining, databases, statistics, biotechnology startups, knowledge management, laboratory information management systems, biomedical research, and biofuels. He was educated at the University of Würzburg in Germany.

Cheryl Kuske
Los Alamos National Laboratory
Cheryl Kuske has a doctorate in plant pathology and molecular biology and 27 years of research experience in microbial ecology, plant-microbe interactions, and pathogen epidemiology. Her professional experience has included positions in academic, industrial, and national laboratory settings. Over the past 15 years, Kuske has developed and applied molecular methods to study microbial communities and their functions in the environment. Her research portfolio while at Los Alamos National Laboratory (LANL) has focused on two goals: (1) understanding the diversity, structure, and functions of soil microbial communities with applications to Department of Energy missions in climate change, carbon cycling, and environmental remediation and (2) developing technology for rapid, accurate detection of pathogens in the environment and understanding their ecology when not associated with a host. Kuske has published about 50 peer-reviewed manuscripts and 14 LANL unclassified reports and holds 4 patents. She has mentored 9 postdoctoral scientists and more than 30 undergraduate and graduate students.

Mary Lipton
Pacific Northwest National Laboratory
Mary Lipton is a senior scientist in systems biology at Pacific Northwest National Laboratory where she specializes in mass spectrometry and ultrasensitive approaches for globally and quantitatively monitoring gene product expression at the protein level. She received her bachelor's degree in chemistry from Juniata College in 1988 and her doctorate in biochemistry from the University of South Carolina in 1993. She has additional research expertise in Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) for biological research; proteomics of *Yersina pestis*; *Rhodopseudomonas palustris* microbial cell; comprehensive analysis of the proteome of *Deinococcus radiodurans*; determination of metal reduction by *Shewanella onedensis*; and direct characterization of DNA damage from ionizing radiation.

Michael Lomas
Bermuda Institute of Ocean Sciences
Michael Lomas received his doctorate in biological oceanography in 1999 from the University of Maryland where he studied the nitrogen metabolism of marine phytoplankton in response to variable light, and therefore cellular energy, environments. He was a postdoctoral scholar at Horn Point Laboratory in the Harmful Algal Research Group before joining in 2001 the Bermuda Institute of Ocean Sciences' Bermuda Atlantic Time-Series Study (BATS). His primary interest is studying the ecological linkages between phytoplankton functional diversity and nutrient biogeochemical cycling. Lomas currently is involved in several projects, including examining long-term patterns in phytoplankton diversity at BATS and relationships to ocean carbon cycling; investigating linkages among interannual variability in sea ice, phytoplankton diversity, and primary production in the eastern Bering Sea; studying dissolved organic phosphorus utilization by phytoplankton taxonomic groups; and linking phytoplankton diversity to variability in carbon export in the Sargasso Sea and the subarctic North Pacific.

John L. Markley
University of Wisconsin, Madison
John L. Markley is the Steenbock professor of biomolecular structure in the biochemistry department at the University of Wisconsin, Madison. Markley, who received his doctorate from Harvard University, uses biophysical and biochemical approaches, principally nuclear magnetic resonance (NMR) spectroscopy, to investigate the structure and function of proteins. He also is active in the field of metabolomics. Markley is director of both the BioMagResBank, which is the NMR component of the Worldwide Protein Data Bank, and the National Magnetic Resonance Facility at Madison. He is the principal investigator for the Center for Eukaryotic Structural Genomics and is a fellow of both the American Association for the Advancement of Science and the Biophysical Society. Markley is an honorary member (and silver medal recipient) of the Nuclear Magnetic Resonance Society of Japan and has authored more than 400 research publications, mainly in the field of structural biology.

Cheryl Marks
National Cancer Institute
Cheryl Marks is associate director of the Division of Cancer Biology at the National Cancer Institute where she also serves as director of the Mouse Models of Human Cancers Consortium Program.

Celeste Matarazzo
Lawrence Livermore National Laboratory
Celeste Matarazzo is a computer scientist at Lawrence Livermore National Laboratory where she is participating in the the Advanced Simulation and Computing (ASCI) program. Matarazzo has more than 15 years of experience

in software development and is a research program manager and leader of the data science research group in the Center for Applied Scientific Computing. Matarazzo also leads the ASCI Scientific Data Management project, which aims to provide intelligent assistance in managing terabytes of complex scientific data through development of data models and tools and integration of databases, storage, networks, and other computing resources. Her previous work experience includes developing software for climate modeling simulations, output devices, and defense applications. Matarazzo has a bachelor's degree in mathematics and computer science from Adelphi University.

Raymond McCord
Oak Ridge National Laboratory
Raymond McCord has been an environmental information manager in the Environmental Sciences Division at Oak Ridge National Laboratory for 21 years. He has managed the development and operation of three major information systems supporting environmental assessment, research, and restoration. McCord also was responsible for establishing a geographic information system within the division. Currently, he is manager of the data archive for the Atmospheric Radiation Measurements Program that supports climate change research. This archive contains 8 million files (~140 TB of storage) about meteorology, solar radiation, and cloud formation. McCord received his doctorate in ecology from the University of Tennessee in 1980.

Lee Ann McCue
Pacific Northwest National Laboratory
Lee Ann McCue received a doctorate in microbiology from Ohio State University. Her research interests focus on comparative genomics, transcription regulation, and the inference of regulatory networks in prokaryotic systems. McCue is a senior research scientist in the Computational Biology and Bioinformatics Group at Pacific Northwest National Laboratory.

Peter McGarvey
Georgetown University Medical Center
Peter McGarvey has 20 years of academic and commercial experience in molecular biology, biotechnology, bioinformatics, and software development. He is interested in genomic and proteomic analysis, biological databases, data integration, and visualization. McGarvey currently is managing the data integration and website for the Biodefense Proteomics Resource, a project of the National Institute of Allergy and Infectious Diseases. He also is a funded participant in the caBIG VCDE (vocabulary and common data elements) workspace and has served as project manager for several caBIG adopter projects. In addition, McGarvey is active in UniProt consortium activities and databases as a member of the Protein Information Resource. He received a doctorate in biological sciences from the University of Michigan in 1988 and a master's degree in technology management from the University of Maryland University College in 2004.

Folker Meyer
Argonne National Laboratory
Folker Meyer is a computational biologist at Argonne National Laboratory and a senior fellow at the Computation Institute at the University of Chicago. He was trained as a computer scientist, which led to his interest in building software systems. Meyer now focuses on building systems that advance scientists' understanding of biological datasets. In the past, he has been known best for his leadership role in developing the GenDB genome annotation system and designing and implementing a high-performance computing facility at Bielefeld University in Germany. Currently, Meyer is most interested in the comparative analysis of large numbers of microbial genomes. He received his doctorate in bioinformatics from Bielefeld University in 2001.

Bob Morris
University of Washington
Bob Morris is an assistant professor of biological oceanography at the University of Washington School of Oceanography. He received his doctorate in microbiology from Oregon State University in 2004. Morris' research interests are marine microbial ecology, bacterioplankton physiology, and microbial community interactions. His laboratory uses cultivation, genomic, and proteomic approaches to study relationships between biogeochemical cycles and microbial processes in the oceans. Morris is specifically interested in exploring the diversity and metabolism of dominant, uncultured bacterioplankton.

Sean Murphy
J. Craig Venter Institute
Sean Murphy is a software engineer at the J. Craig Venter Institute (JCVI) where he develops enterprise software applications to support bioinformatic research. He currently is a member of the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (termed CAMERA) team. With an extensive software background, including database design, middleware architecture, asynchronous messaging systems, grid-computing interfaces, model-view controllers, and website design, Murphy is experienced in developing software systems end to end. Before joining JCVI, he worked at Celera

Genomics and Intelligent Medical Imaging. Murphy has developed several products, including the Moore Microbial Genome website, the JCVI Blast Server, the JCVI resequencing primer designer, the Celera Gene Index pipeline, and machine-vision algorithms for automated pathology applications. He has bachelor's degrees in electrical engineering and biology from the Massachusetts Institute of Technology, and he earned a doctorate in neuroscience from Yale University.

Ilya Nemenman
Los Alamos National Laboratory
Ilya Nemenman received his doctorate in theoretical physics, specializing in biophysics, from Princeton University. He completed extra postdoctoral training at the NEC Research Institute and Kavli Institute for Theoretical Physics at the University of California, Santa Barbara. Nemenman was a member of the research faculty at Columbia University Medical School. In 2005, he joined the Computer, Computational, and Statistical Sciences Division at Los Alamos National Laboratory. His research interests focus on information processing in biological systems, from neural assemblies to molecular signaling and regulatory pathways.

Gary J. Olsen
University of Illinois
Gary J. Olsen is a microbiology professor at the University of Illinois, Urbana-Champaign. He received a bachelor's degree in physics at the University of California, Los Angeles, in 1975; a master's in physics from UCLA in 1976; and a doctorate in biophysics from the University of Colorado Health Sciences Center in 1983. Olsen conducted postdoctoral work in molecular and cellular biology at the National Jewish Hospital and Research Center (1983–84) and in biology at Indiana University (1984–1985). He also was an assistant scientist in biology at Indiana University from 1985 to 1988. Most of Olsen's current research focuses on two areas: (1) gene expression in archaea and its relation to corresponding systems in eucarya and bacteria and (2) genomics, with an emphasis on comparative genomics and genome evolution. His approach combines experimental work and computational analyses of genomes and proteins.

Andrei Osterman
Burnham Institute for Medical Research
Andrei Osterman is an associate professor in the Bioinformatics and Systems Biology Program at the Burnham Institute for Medical Research (BIMR). He received his doctorate in biochemistry at Moscow University in Russia. In 1993, Osterman joined the laboratory of Meg Phillips at the University of Texas in Dallas to perform structure–functional studies of the ornithine decarboxylase enzyme family. Recognizing the new frontiers of metabolic biochemistry and enzymology enabled by the genomics revolution, Osterman joined Integrated Genomics, a startup biotechnology firm in Chicago in 1999. As a director and vice president of research at Integrated Genomics, he pioneered integration of comparative genomics with biochemical and genetic experiments for gene and pathway discovery. His research team published the first genome-scale study of gene essentiality in *Escherichia coli* by genetic footprinting. Osterman is one of the founders of the Fellowship for Interpretation of Genomes (FIG), a nonprofit research organization that launched the Project to Annotate 1000 Genomes in 2003. FIG provides the open-source integration of all publicly available genomes and tools for their comparative analysis, annotation, and metabolic reconstruction. Osterman's laboratory at BIMR focuses on fundamental and applied aspects of key metabolic subsystems in a variety of species, from bacteria to human. His group applies bioinformatic techniques followed by experimental validation to reconstruct metabolic pathways from genomic data, reveal gaps in current knowledge, and identify previously uncharacterized (missing) genes. The power of this integrative approach is illustrated by the discovery and characterization of more than 20 enzyme families in the metabolism of cofactors, carbohydrates, and amino acids. Most applications pursued by this group are related to pathogenic and environmental bacteria. New research directions include the analysis of regulatory networks and the application of proteomics and metabolomics technology for identification of novel diagnostic and therapeutic targets in cancer.

Ross Overbeek
Argonne National Laboratory
Ross Overbeek received a doctorate in computer science in 1972 from Penn State University. For 11 years, Overbeek taught mathematics and computer science at Northern Illinois University where his research focused on computational logic and database systems. From 1983 to 1998, Overbeek worked at Argonne National laboratory (ANL), focusing on parallel computation and logic programming. Overbeek collaborated with Carl Woese and helped in the founding of the Ribosomal Database Project. Overbeek participated in analysis of *Methanococcus jannaschii*, the first archaeal genome to be completely sequenced. He was the lead architect of the PUMA and WIT genomics database systems at ANL before becoming a founder of Integrated Genomics (IG) in 1998. While at IG, Overbeek participated in the sequencing and analysis of more than 50

genomes and led the company's bioinformatics effort. The most significant product was ERGO, a system to support comparative genomic analyses. In mid- 2003, Overbeek left IG to become a founding fellow of the Fellowship for Interpretation of Genomes (FIG). His efforts at FIG centered on building the SEED, an open-resource data curation system to facilitate comparative analyses of genomic data. Since 2004, Overbeek has been a coprincipal investigator of the National Microbial Pathogen Data Resource, a framework to support comparative analysis of pathogen genomes.

George N. Phillips, Jr.
University of Wisconsin, Madison
George N. Phillips, Jr. received his doctorate in biochemistry at Rice University in Houston, Texas, in 1976. He currently is professor of biochemistry and computer sciences at the University of Wisconsin, Madison. Phillips leads the Computation and Informatics in Biology and Medicine training program supported by the National Library of Medicine. He also is coinvestigator at the Center for Eukaryotic Structural Genomics and serves as the informatics and information technology manager at the Department of Energy's Great Lakes Bioenergy Research Center. Phillips' research interests are computational and structural biology.

David Pletcher
Joint BioEnergy Institute
David Pletcher is the director of informatics at the Department of Energy's (DOE) Joint BioEnergy Institute (JBEI). Prior to joining JBEI, Pletcher worked for nearly 7 years as a computer scientist at Lawrence Livermore National Laboratory. He also served as group leader of production informatics at the DOE Joint Genome Institute from June 2004 to August 2008. Pletcher began his career in the private sector, working as a programmer and software engineer and developer for several companies, including Rockwell Scientific, Lumisys, Visual Edge Technology, and idrive.com. He graduated from Harvey Mudd College in 1992.

Jennifer Reed
University of Wisconsin, Madison
Jennifer Reed is an assistant professor in the chemical and biological engineering department at the University of Wisconsin, Madison. She received her bachelor's and master's degrees as well as her doctorate from the University of California, San Diego. Most of Reed's research interests involve studying metabolism and regulation through the generation and subsequent analysis of metabolic models and reconciling the models with experimental data. Overall, her research group uses computational models and develops methods to study biological systems, engineer cells, and expand scientific knowledge of the mechanisms underlying observed cellular behavior. The group specifically is interested in building, analyzing, and using metabolic and regulatory models of organisms involved in environmental remediation, biofuels, and pharmaceutical applications. Reed's laboratory also uses the developed models to identify novel gene functions or regulatory interactions. In addition to model building, her research involves computational methods for designing strains or cell lines with enhanced production yields of desired products.

Nagiza F. Samatova
Oak Ridge National Laboratory
Nagiza F. Samatova is a senior research scientist in the Computational Biology Institute, Computer Science and Mathematics Division, at Oak Ridge National Laboratory. She received her bachelor's degree in applied mathematics from Tashkent State University in Uzbekistan in 1991 and her doctorate in mathematics from the Russian Academy of Sciences in Moscow in 1993. Samatova also obtained a master's in computer science in 1998 from the University of Tennessee. She specializes in computational biology and high-performance data mining, knowledge discovery, and statistical data analysis. Samatova is the author of more than 50 publications, including 1 book, and she holds 2 patents.

Denise Schmoyer
Oak Ridge National Laboratory
Denise Schmoyer is a research staff member in the Computer Science and Mathematics Division at Oak Ridge National Laboratory. Schmoyer has worked on the design and development of several large-scale database systems for human, model organism, and microbial sequence annotation. She is the primary developer of a laboratory information management system and database for protein complexes in microbial organisms.

Blake A. Simmons
Joint BioEnergy Institute
Blake A. Simmons received a bachelor's degree in chemical engineering in 1997 from the University of Washington and attended graduate school at Tulane University where the focus of his thesis work was the synthesis and characterization of templated nanomaterials. He earned a doctorate in chemical engineering from Tulane in 2001. Simmons then joined Sandia National Laboratories in Livermore, California, as a senior member of the technical staff, working in the Materials Chemistry Department. He participated in and led various projects, including the development of cleavable surfactants, enzyme engineering for biofuel cells, microfluidics, and the synthesis of

silicate nanomaterials. In 2004, Simmons was promoted to principal member of the technical staff. He expanded his research portfolio to include the design, fabrication, integration, and testing of polymeric microfluidic devices for several lab-on-a-chip and homeland security applications. He also continued to pursue opportunities in renewable energy. In 2006, Simmons was promoted to manager of the Energy Systems Department, which focuses on developing novel, materials-based solutions to meet the United States' growing energy demands. In 2007, Simmons was named one of the principal coinvestigators of the Joint BioEnergy Institute (JBEI, http://www.jbei.org), a $135 million project funded by the Department of Energy and tasked with developing next-generation biofuels produced from nonfood crops. He currently is serving as vice president of the Deconstruction Division at JBEI where he leads a team of 35 researchers working on advanced methods of liberating fermentable sugars from lignocellulosic biomass. He also manages the Biomass Science and Conversion Technology Department at Sandia. Simmons has authored more than 70 publications, book chapters, and patents.

Tom Slezak

Lawrence Livermore National Laboratory

Tom Slezak has been involved with bioinformatics at Lawrence Livermore National Laboratory (LLNL) for more than 28 years. He received his bachelor's degree in computer science from the University of San Francisco and his master's in computer science from the University of California, Davis. Slezak participated in the Human Genome Project from its inception and led the informatics efforts at LLNL and then the Department of Energy's Joint Genome Institute from 1987 to 2000. In 2000, Slezak began to assemble a pathogen bioinformatics team at LLNL, pioneering a whole genome analysis approach to DNA signature design. His team developed signature targets to identify multiple human pathogens, and these targets were used as part of the biodefense measures at the 2002 Winter Olympic Games under the BASIS program. They later were adapted for use nationwide as part of the Centers for Disease Control and Prevention's (CDC) BioWatch program. Slezak's bioinformatics team has developed DNA-based signatures of virtually every biothreat pathogen (the organisms identified by CDC as high-priority threat agents) for which adequate genomic sequences are available, as well as of several other human and livestock pathogens. LLNL signatures are part of the nation's public health system and have been in use for homeland defense since fall 2001.

Rick Stevens

Argonne National Laboratory

Rick Stevens is associate laboratory director for Computing, Environment, and Life Sciences at Argonne National Laboratory and professor of computer science at the University of Chicago. His research interests are high-performance computer architectures and computational science, especially challenges in the life sciences. Stevens leads Argonne's efforts in advanced computing that target the development of exascale computing technology and applications in systems and computational biology and environmental modeling and simulation. He is a fellow of the American Association for the Advancement of Science and is also a cofounder and senior fellow of the Argonne and University of Chicago Computation Institute, a multidisciplinary institute aimed at connecting computing to all areas of inquiry at the university and laboratory.

Michael R. Sussman

University of Wisconsin, Madison

Michael R. Sussman has been a faculty member at the University of Wisconsin, Madison, for the past two decades. During that time, he has become recognized as a leading expert on signal transduction and genomics in plants. Sussman's research interests have focused on using the model higher-plant *Arabidopsis thaliana* for understanding the role of plasma membrane proteins in signal transduction and solute transport. His laboratory was the first to report on unique protein kinases found only in plants and protists and on the plasma membrane proton pump, which provides the driving force for the uptake of all nutrients. To help understand the in situ role played by these important proteins, Sussman's laboratory pioneered the development of genome-wide reverse genetics techniques. Specifically, the lab used an insertional mutagenesis scheme to isolate "knockout" plants, starting with the sequence for any one of the roughly 30,000 genes in *Arabidopsis*. For example, Sussman's laboratory was the first to demonstrate that the plant homologue for a brain potassium channel performs a nutritional role in plants (i.e., is responsible for the uptake of potassium from soil). Similar studies have been performed to identify the in planta roles of several plasma membrane hormone receptors.

In 1999, Sussman, together with colleague Franco Cerrina, a professor in the College of Engineering, developed a new instrument known as a MAS (Maskless Array Synthesizer), which makes "gene chips" that can analyze hundreds of thousands of genes at once. MAS is unique because it eliminates the requirement for expensive masks used in traditional DNA

chip technology, thus making MAS accessible to all scientists. Based on the MAS technology, Sussman and Cerrina founded in 2000 a biotechnology startup company, NimbleGen Systems Inc., which after 8 years, was sold to Roche Inc.

Sussman's awards have included a Fulbright research fellowship for a sabbatical in Belgium; a McKnight Foundation award; a University of Wisconsin, Madison, WARF Kellett Mid-Career Award; and selection as a fellow to the American Association for the Advancement of Science. In 1996, Sussman was appointed director of the UW Biotechnology Center (UWBC), a campus-wide facility devoted to research, outreach, and service in the area of biotechnology and genomic science and instrumentation.

Ronald Taylor
Pacific Northwest National Laboratory
Ronald Taylor earned a doctorate in bioinformatics from George Mason University in Fairfax, Virginia. He received his bachelor's degree in physics, master's degree in computer science, and master's in biology from Case Western Reserve University in Cleveland, Ohio. Taylor is a research scientist at Pacific Northwest National Laboratory (PNNL) where he develops algorithms and software for inference of biological networks. He also is involved in the development of biological databases, leading one such project at PNNL.

Gerald A. Tuskan
Oak Ridge National Laboratory
Gerald A. Tuskan is a distinguished scientist in the Plant Genomics Group within the Environmental Sciences Division at Oak Ridge National Laboratory (ORNL) where he coordinates the Department of Energy's (DOE) effort to sequence the *Populus* genome. He received a master's degree in forest genetics from Mississippi State University in 1980 and a doctorate in genetics from Texas A&M University in 1984. In addition to his work at ORNL, Tuskan is involved in the laboratory science program for the DOE Joint Genome Institute (JGI). In this role, he coordinates the solicitation and review of principal investigator–led sequencing proposals submitted through the DOE laboratory system; helps establish multiple large-genome sequencing projects that address DOE missions in biofuel development, carbon biosequestration, and global climate change; and facilitates DOE, laboratory, and JGI interactions. Tuskan also is a research professor in the entomology, plant pathology, and plant sciences departments at the University of Tennessee. His research interests include understanding the genetic basis of tree growth and development with emphasis on biomass accumulations, carbon allocation, and cell-wall chemistry; *Populus* genomics, including assembly of the draft sequence, comparative genomics, and functional gene identification; and short-rotation woody crop silvicultural systems.

Edward C. Uberbacher
Oak Ridge National Laboratory
Edward C. Uberbacher received his bachelor's degree in chemistry from Johns Hopkins University in 1974 and a doctorate in physical chemistry from the University of Pennsylvania in 1979. Beginning in 1980, he conducted postdoctoral studies in the University of Pennsylvania's biophysics department and in the Biology Division of Oak Ridge National Laboratory (ORNL) and the University of Tennessee's Graduate School of Biomedical Sciences, investigating the structure and function of genetic materials using crystallography, electron microscopy, and computational modeling. In 1985, Uberbacher became an investigator at the Center for Small-Angle Scattering Research at ORNL, pursuing structural and dynamic studies of macromolecules in solution using techniques involving neutron and X-ray scattering and molecular modeling. In 1987, he also became a research assistant professor at UT's Graduate School of Biomedical Sciences and an investigator in the ORNL Biology Division, focusing on X-ray and neutron crystallography, scattering, and other biophysical methods. Uberbacher became a consultant in the ORNL Engineering Physics and Mathematics Division in 1988 to develop artificial intelligence and high-performance computing methods for genomic DNA sequence analysis; in 1991, he joined the staff of the Computer Science and Mathematics Division as the informatics group leader. In this role, he received an R&D 100 Award for developing the GRAIL DNA sequence analysis system. In 1997, Uberbacher became the head of ORNL's Computational Biology Section in Life Sciences and a codeveloper of the PROSPECT computational protein fold prediction system, which received an R&D 100 Award in 1998. Uberbacher performed part-time duties as an IPA in 2003–04 for the Department of Energy's Office of Biological and Environmental Research, contributing extensively to the Genomics:GTL computing roadmap. He is currently the program leader for Computational Biology at ORNL and is an adjunct professor in the Genome Science and Technology Program at the University of Tennessee. His scientific interests include the application of pattern recognition; artificial intelligence; concurrent processing techniques and algorithm development for computational biology; computational genome sequence analysis; mass spectrometry analysis; and macromolecular structure, dynamics, and docking.

Owen White
University of Maryland, College Park
Owen White, professor of epidemiology and preventive medicine, is the director of bioinformatics at the University of Maryland School of Medicine. He received his doctorate in molecular biology from New Mexico State University in 1992 and is an internationally recognized expert in bioinformatics. He is the principal investigator of the Data Analysis and Coordination Center (funded by the National Human Genome Research Institute) of the Human Microbiome Project, a National Institutes of Health Roadmap Initiative. In this capacity, White is responsible for coordinating the collection, integration, standardization, analysis, and distribution of all genomic and metagenomic data related to the Human Microbiome Project. At the Institute for Genome Sciences (IGS), he leads a group of more than 20 scientists and engineers who collectively are responsible for developing nearly all IGS production-level annotation pipelines, database systems, and tools for automated and manual annotation of genomes and metagenomic datasets. White has experience in DNA sequence generation and genomic analysis of human expressed sequence tags, other eukaryotes, and prokaryotes as well as in comparative analyses.

John Wooley
University of California, San Diego
John Wooley is associate vice chancellor for research; professor of pharmacology, chemistry, and biochemistry; and director of digitally enabled genomic medicine at the University of California, San Diego. He also is chief scientific officer of the metagenomics cyber-resource project termed CAMERA at the university. This infrastructure project focuses on linking environmental metadata to molecular data and on the development and provision of software tools in a rich computing environment to probe metagenomic data and advance microbial ecology. Wooley's current research involves bioinformatics and structural biology focused on protein structure-function relationships. He is co-principal investigator of the Joint Center for Structural Genomics, a high-throughput structural pipeline. For nearly three decades, Wooley has been focused on nurturing the interface between computing and biology. He received his doctorate in 1975 at the University of Chicago.

Alex Worden
Monterey Bay Aquarium Research Institute
Alex Worden is a microbiologist at the Monterey Bay Aquarium Research Institute (MBARI). She earned a bachelor's degree in history from Wellesley College, with a concentration in Earth, atmospheric, and planetary sciences at the Massachusetts Institute of Technology (MIT). Worden remained at MIT for 2 years as a research technician and then joined the University of Georgia where she received a NASA Earth systems science fellowship and completed her doctorate in ecology in 2000. Worden spent 3.5 years conducting postdoctoral research on microbial interactions at the Scripps Institution of Oceanography. She then accepted an assistant professorship at the Rosenstiel School of Marine and Atmospheric Science at the University of Miami. In 2007, Worden joined MBARI where she leads a microbial ecology research group. Her research interests include population regulation of photoautotrophic microbes, with an emphasis on carbon cycling in marine systems. Worden's laboratory uses a range of methods and technologies, from seagoing oceanography to genomics and metagenomics.

Cathy Wu
Georgetown University
Cathy Wu is a professor in the biochemistry and molecular and cellular biology department at Georgetown University's School of Medicine. She also is a professor in the oncology department and is director of the Protein Information Resource (PIR) at Georgetown University Medical Center. Wu has master's degrees in plant pathology and computer science and received her doctorate in molecular plant pathology from Purdue University in 1984. She has conducted bioinformatic research since 1990 and has developed several protein classification systems and databases. Wu has managed large software and database projects and has led the bioinformatics effort of PIR since 1999, becoming PIR director in 2001. Her research interests include protein family classification and functional annotation, biological data integration, and literature mining.

Liming Yang
National Center for Research Resources
Liming Yang is a health scientist administrator in the Biomedical Technology Division of the National Center for Research Resources (NCRR) within the National Institutes of Health (NIH). He manages a portfolio of grants on computational biology, software development, and genetic studies. Before joining NCRR, Yang was associate director of biomedical informatics from 2005 to 2008 at the Center for Bioinformatics within the National Cancer Institute. He led several projects to build bioinformatics infrastructures to support large genomics and proteomics initiatives. Prior to that position, Liming was an intramural scientist at NIH where he played an important role in data analysis and management for the multi-institute Lymphoma and Leukemia Molecular Profiling Project. Yang received his doctorate in pathology from the University of Utah School

of Medicine. After that, he spent 2 years as a postdoctorate fellow at NIH. Yang is from Beijing, China, where he attended Peking University for undergraduate studies and Peking Union Medical College for medical school.

William S. York
University of Georgia
William S. York received his bachelor's degree in molecular, cellular, and developmental biology in 1978 from the University of Colorado and his doctorate in biochemistry and molecular biology in 1996 from the University of Georgia. York was senior research chemist at the Complex Carbohydrate Research Center from 1985 to 1996 before beginning his faculty career at the University of Georgia. His diverse research interests include the development and application of spectroscopic and computational methods for structural and conformational analysis of complex carbohydrates, development of bioinformatic tools to study the roles of carbohydrates in living systems, and the use of these tools to create realistic models describing the assembly and morphogenesis of the walls surrounding the cells of higher plants. York's current research includes the application of these techniques to understand the recalcitrance of biomass to saccharification. Results of this research may lead to improvement of feedstocks for the biofuel industry. His research is supported by the Department of Energy, the National Science Foundation, the National Institutes of Health, and the University of Georgia Research Foundation.

Appendix 9
Glossary

algae: Photosynthetic, aquatic, eukaryotic organisms that contain chlorophyll but lack terrestrial plant structures (e.g., roots, stems, and leaves). Algae can exist in many sizes ranging from single cells to giant kelps several feet long.

algorithm: Formal set of instructions that tells a computer how to solve a problem or execute a task. A computer program typically consists of several algorithms.

annotation: Addition of biologically meaningful descriptions to data (e.g., by labeling regions of sequence data that encode a gene or regulatory region or by identifying the active site of a protein structure).

application programming interface (API): A set of standardized messages or protocols that a program can use to communicate with and request services from another program.

archaea: One of the three domains of life (along with bacteria and eukarya) distinguished through DNA sequence analysis. Archaea are structurally and metabolically similar to bacteria but share some features of their molecular biology with eukaryotes.

architecture: Operational structure of a computer system.

bacteria: One of the three domains of life (along with archaea and eukarya) distinguished through DNA sequence analysis. Also a general term referring to prokaryotic organisms that do not belong to the archaea domain (singular: bacterium).

Bayesian approach: Use of statistical methods that assign probabilities or distributions to future events based on knowledge of prior events.

bioenergy: Energy-related product (e.g., solid, liquid, or gaseous fuels; electricity; and heat) derived from renewable biobased materials (e.g., plant matter and organic waste) or biological processes (e.g., biochemical activities of microbes or plants).

biofilm: Community of microorganisms living together on a surface and embedded in extracellular polymers they create.

biogeochemistry: Study of how interactions among biological and geochemical processes influence the global cycling of such essential elements as carbon, nitrogen, phosphorus, and sulfur.

biogeography: Study of the physical distribution of organisms.

bioinformatics: Science of managing and analyzing biological data using advanced computing techniques.

biomass: Organic material from living organisms, typically plant matter such as trees, grasses, and agricultural crops, that can be burned or converted to liquid or gaseous fuels for energy.

biome: Large geographic region defined by environmental conditions and biological communities found in the area.

bioreactor: Vessel in which biocatalysts or microorganisms involved in the production of a desired biological product are maintained. In industry, bioreactors typically house fermentation reactions and are called fermenters.

biosequestration: Biologically mediated uptake and conversion of carbon dioxide to inert, long-lived, carbon-containing materials.

biota: Living organisms.

C_3 plant: Plants (e.g., soybean, wheat, and cotton) whose carbon-fixation products have three carbon atoms per molecule. Compared with C_4 plants, C_3 plants show a greater increase in photosynthesis with a doubling of CO_2 concentration and less decrease in stomatal conductance, which results in an increase in leaf-level water use efficiency

carbon cycle: The complex carbon flows and transformations among major Earth system components (atmosphere, oceans, and terrestrial systems). The global flow of carbon from one reservoir (carbon sink) to another. Each carbon exchange among reservoirs is mediated by a variety of physical, biogeochemical, and human activities.

carbon dioxide (CO_2): Colorless, odorless gas that absorbs infrared radiation and traps heat in the Earth's atmosphere. CO_2, which is important to the global carbon cycle, is emitted from a variety of processes (e.g., cellular respiration, biomass decomposition, fossil-fuel use) and taken up primarily by photosynthesis and the oceans via air-sea gas exchange.

carbon fixation: Conversion of inorganic carbon dioxide to organic compounds by photosynthesis.

carbon flux: Rate of carbon movement as it flows from one carbon reservoir to another within an organism, ecosystem, or the global carbon cycle.

carbon partitioning: Partitioning to different parts of a plant (e.g., leaf, stem, root, and seed) versus carbon allocation (partitioning between biomass and respiration).

carbon sequestration: Biological or physical process that captures carbon dioxide and converts it into inert, long-lived, carbon-containing materials.

carbon sink: A pool (reservoir) that absorbs or takes up released carbon from another part of the carbon cycle.

carbon source: A pool (reservoir) that releases carbon to another part of the carbon cycle.

cellulose: Large, complex polysaccharide that is a major component of plant cell walls. Each cellulose molecule is a linear chain of thousands of glucose subunits; multiple cellulose chains form cable-like structures that stabilize the matrix of plant cell-wall materials.

chromatin immunoprecipitation (ChIP): In vivo method that uses antibodies targeted to specific DNA-binding proteins to analyze protein-DNA interactions and determine which sequences these proteins bind and where these proteins bind the genome.

climate model: Mathematical model used to understand, simulate, and predict climate trends by quantitatively analyzing interactions among Earth system components (e.g., land, ocean, atmosphere, and biosphere).

cofactor: Inorganic substance (e.g., metal ion) that is a component of an enzyme complex and required for enzyme activity.

co-immunoprecipitation (Co-IP): Technique that uses antibodies to detect interacting proteins. An antibody that specifically binds a target protein is added to a sample of cellular material. The antibody forms a complex with its target and any protein or molecule bound to the target. Then an antibody-binding protein immobilized on a tiny bead is added and used to pull the antibody-protein complex out of solution.

community: All the different species of organisms living together and interacting in a particular environment.

complexes: Aggregates of multiple, interrelated molecular parts.

contig: Group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome.

curation: A process in which experts manually review, validate, update, and add value to biologically meaningful representations of data, information, and knowledge in computer systems.

cyanobacteria: Division of bacteria capable of oxygen-producing photosynthesis and found in many environments, including oceans, fresh water, and soils. Cyanobacteria contain chlorophyll a and other photosynthetic pigments in an intracellular system of membranes called thylakoids. Many cyanobacterial species also are capable of nitrogen fixation.

data mining: Data analysis techniques used to sift through large amounts of data and identify hidden patterns and relationships.

data model: Logical structure for representing data associated with a particular concept and relating it to other data in a database.

data standard: Set of specifications, established by community consensus or authorized by an official standards organization, for representing and organizing data in ways that promote the exchange, comparison, and integration of different datasets.

DNA (deoxyribonucleic acid): Molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases adenine (A), guanine (G), cytosine (C), and thymine (T).

ecosystem: Set of organisms (plants, animals, fungi, and microorganisms) and the physical and chemical factors that make up a particular environment.

electrophoretic mobility shift assay (EMSA): In vitro method for characterizing the interactions between a protein and DNA or RNA. When a protein binds a labeled piece of DNA or RNA, it forms a large molecular complex that moves more slowly down through a gel than free DNA or RNA molecules. Variations of this basic method can identify the specific DNA or RNA sequence that the protein binds, determine the affinity of the protein for a specific sequence, and reveal which proteins in a mixture bind a particular sequence.

eukaryote: A single-celled or multicellular organism (e.g., plant, animal, or fungi) with a cellular structure that includes a membrane-bound, structurally discrete nucleus and other well-developed subcellular compartments. *See also prokaryote.*

expressed sequence tag (EST): A short segment of DNA sequence derived from the mRNA of transcribed (expressed) genes that can be used to uniquely identify and locate full-length, protein-coding genes within a genome.

expression: *See gene expression.*

fluorescence in situ hybridization (FISH): Technique for microscopic imaging that uses fluorescent probes targeted to signature DNA or RNA sequences to identify and locate different populations in a microbial community without having to grow microbes in culture.

fungi: Eukaryotic, heterotrophic organisms—ranging from single-celled yeasts to multicellular molds and mushrooms—that lack chlorophyll, have rigid cell walls, and absorb nourishment from living or dead organic matter.

gene: Fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides, located in a particular position on a particular chromosome, that encodes a specific functional product (i.e., a protein or RNA molecule).

gene calling: Computational process for identifying where genes begin and end within a genome and for assigning meaningful descriptions to DNA segments recognized as genes.

gene expression: Process by which a gene's coded information is converted into structures present and operating in the cell. Expressed genes include those transcribed into mRNA and then translated into proteins, as well as those transcribed into RNA but not translated into proteins [e.g., transfer (tRNA) and ribosomal RNA (rRNA)].

gene family: Group of closely related genes that make similar products.

gene prediction: Computational method for identifying the locations and sequences of possible genes within a genome. Several gene prediction approaches are based on how well an unknown stretch of DNA sequence matches known gene sequences.

gene product: Protein or RNA molecules resulting from the expression of a gene's DNA sequence. The amount of gene product is used to measure a gene's level of expression.

gene regulatory network: Intracellular network of regulatory proteins that control the expression of gene subsets involved in particular cellular functions. A simple network would consist of one or more input signaling pathways, regulatory proteins that integrate the input signals, several target genes (in bacteria, a target operon), and the RNA and proteins produced from those target genes.

genetic code: Nucleotide sequence, coded in triplets along the mRNA, that determines the sequence of amino acids in a protein product. Each set of three nucleotides (codon) in a gene specifies a particular amino acid or signals the start or stop of protein synthesis.

genome: All the genetic material in the chromosomes of a particular organism. Most prokaryotes package their entire genome into a single chromosome, while eukaryotes have different numbers of chromosomes. Genome size generally is given as total number of base pairs.

genome sequence: Order of nucleotides within DNA molecules that make up an organism's entire genome.

genomics: Study of genes and their function.

genotype: An organism's genetic constitution, as distinguished from its physical characteristics (phenotype).

gross primary production (GPP): Total amount of organic matter created by photosynthesis.

hemicellulose: Any of several polysaccharides (e.g., xylans, mannans, and galactans) that cross-link and surround cellulose fibers in plant cell walls.

heterotroph: Organism that obtains organic carbon by consuming other organisms or the products of other organisms.

high throughput: Analytical or computational analysis done on a massive, automated scale.

horizontal gene transfer (or lateral gene transfer): Exchange of genetic material between two different organisms (typically different species of prokaryotes). This process gives prokaryotes the ability to obtain novel functionalities or cause dramatic changes in community structure over relatively short periods of time. *See also vertical gene transfer.*

horizontal queries: Queries that associate equivalent data entities across species, samples, or habitats (e.g., homologous genes between species, community composition across samples, and abundance or enrichment of metabolic pathways across habitats).

informatics: Science of managing and analyzing data using advanced computing techniques.

interoperability: Ability of two or more computer systems to work together by exchanging services or communicating, sharing, and interpreting data using common protocols.

isobaric tag for relative and absolute quantitation (iTRAQ): Chemical probe that labels the N-terminus of all peptides in up to eight different biological samples, thus enabling the identification and quantitation of corresponding proteins using mass spectrometry–based proteomic approaches. Samples subjected to different experimental

conditions can be tagged with different iTRAQ labels and then mixed together (multiplexed) to enable simultaneous quantitative analysis.

isotherm: Line on a map indicating points of equal temperature.

isotope-coded affinity tag (ICAT): Chemical probe that labels cysteine residues in proteins, thus enabling the selective isolation and quantitation of particular subsets of proteins using mass spectrometry–based proteomic approaches. Samples subjected to different experimental conditions can be tagged with ICAT labels of different molecular mass and then mixed together to enable comparisons of protein abundance levels in a single analysis step.

knowledgebase: Comprehensive collection of knowledge stored in databases and used to solve problems in a particular subject area such as biology.

latency: Delays that can affect system response time.

lateral gene transfer: *See horizontal gene transfer.*

life-cycle management: End-to-end management of a project.

lignin: Complex, insoluble polymer whose structure surrounds and gives strength and rigidity to cellulose fibers in the cell walls of woody plants. Lignin makes up a significant portion of the mass of dry wood and, after cellulose, is the second most abundant form of organic carbon in the biosphere.

lignocellulose: Refers to plant cell-wall materials primarily made up of lignin, cellulose, and hemicellulose.

LIMS: Acronym for laboratory information management system, which is a computerized system used by laboratories to track samples; automate data capture from laboratory instruments; and facilitate storage, presentation, and sharing of data among collaborating researchers.

machine reasoning: Ability of a computer to make selections or solve problems using approaches that model human reasoning and learning.

mass spectrometry: Method involving specialized instruments for measuring the mass and abundance of molecules in a mixture and identifying mixture components by mass and charge.

messenger RNA (mRNA): RNA that serves as a template for protein synthesis. *See also transcription and translation.*

metabolic flux analysis: System-level understanding and quantitation of the flow of molecules through metabolic networks.

metabolism: Collection of all biochemical reactions that an organism uses to obtain the energy and materials it needs to sustain life. An organism uses energy and common biochemical intermediates released from the breakdown of nutrients to drive the synthesis of biological molecules.

metabolites: Small molecules (<500 Da) that are the substrates, intermediates, and products of enzyme-catalyzed metabolic reactions.

metabolomics: Type of global molecular analysis that involves identifying and quantifying the metabolome—all metabolites present in a cell at a given time.

metadata: Data that describe specific characteristics and usage aspects of other data (e.g., what data are about, when and how data were created, who can access the data, and the formats available).

metagenomics: Study of the collective DNA isolated directly from a community of organisms living in a particular environment.

metaomics: High-throughput, global analysis of DNA, RNA, proteins, or metabolites isolated directly from a community of organisms living in a particular environment.

metaproteomics: High-throughput, global analysis of proteins isolated directly from a community of organisms living in a particular environment. Metaproteomics can reveal which genes are actively translated into functional proteins by a community.

metatranscriptomics: High-throughput, global analysis of RNA isolated directly from a community of organisms living in a particular environment. Metatranscriptomics can reveal which genes are actively expressed by a community.

microalgae: Microscopic, unicellular aquatic plants.

microarray: Analytical technique used to measure the mRNA abundance (gene expression) of thousands of genes in one experiment. The most common type of microarray is a glass slide onto which DNA fragments are chemically attached in an ordered pattern. As fluorescently labeled nucleic acids from a sample are applied to the microarray, they bind the immobilized DNA fragments and generate a fluorescent signal indicating the relative abundance of each nucleic acid in the sample.

microbiome: A community of microorganisms that inhabits a particular environment. For example, a plant microbiome includes all the microorganisms that colonize a plant's surfaces and internal passages.

microorganism: Sometimes called a microbe, this is any microscopic prokaryotic or eukaryotic organism, including bacteria, archaea, and protists.

model: Mathematical representation used in computer simulations to calculate the evolving state of dynamic systems.

model ecosystem: A specific type of ecosystem that is widely studied in great detail by a community of researchers to provide insights into the processes controlling the behavior of other ecosystems.

modeling: Use of statistical and computational techniques to create working computer-based models of biological phenomena that can help to formulate hypotheses for experimentation and predict outcomes of research.

molecular machine: Highly organized assembly of proteins and other molecules that work together as a functional unit to carry out operational, structural, and regulatory activities in cells.

motif: A sequence motif is a short, recurring pattern of nucleotides (in DNA or RNA) or amino acids (in proteins) that can signal a particular function or molecular event (e.g., a sequence where a protein binds). A structural motif is a recurring, three-dimensional arrangement of structural elements observed in different proteins.

net primary production (NPP): Fraction of photosynthetically fixed organic matter that remains after accounting for carbon lost to cellular respiration and other biological processes.

nitrogen fixation: Process carried out by certain species of bacteria and archaea in which atmospheric nitrogen (N_2) is converted to organic nitrogen-containing compounds that can be used by other organisms.

noncoding RNA (ncRNA): Any RNA molecule that does not serve as a template for protein synthesis.

nonprocedural relational operators: Programming language constructs that are used to compare and test the relationship between two values or entities. With nonprocedural relational operators, the user specifies what output is needed but does not specify the procedure to obtain the output.

object-relational system: System that combines object-related database concepts with relational databases.

omics: Collective term for a range of new high-throughput biological research methods (e.g., transcriptomics, proteomics, and metabolomics) that systematically investigate entire networks of genes, proteins, and metabolites within cells.

ontology: Organized, hierarchical structure of concepts relevant to a particular knowledge domain. An ontology identifies which of several equivalent terms should be used to represent a concept and defines how different terms and concepts are related. Ontologies are developed to ensure the consistent use of language across multiple databases and information systems.

operon: In prokaryotic genomes, a linear group of genes transcribed together on the same mRNA molecule and controlled by the same regulatory element.

organelle: Specialized structure within a cell with a specific function.

parallelization: Simultaneous use of multiple computers to carry out a particular task or solve different parts of the same problem.

parameterization: In climate modeling, approach used to represent phenomena that are too small-scale or complex to be included in a model.

pathway: Series of molecular interactions that occur in a specific sequence to carry out a particular cellular process (e.g., sense a signal from the environment, convert sunlight to chemical energy, break down or harvest energy from a carbohydrate, synthesize ATP, or construct a molecular machine).

phenology: Study of recurring biological phenomena (e.g., seasonal leaf loss in trees) and how changes in climate or the surrounding environment can impact the timing of these periodic events.

phenomics: Collective study of multiple phenotypes (e.g., all phenotypes associated with a particular biological function).

phenotype: Physical characteristics of an organism.

photosynthate: Organic carbon produced by photosynthesis.

photosynthesis: Process by which plants, algae, and certain types of prokaryotic organisms capture light energy and use it to drive the transfer of electrons from inorganic donors (e.g., water) to carbon dioxide to produce energy-rich carbohydrates.

photosystem: Large, membrane-bound molecular complex consisting of multiple proteins containing pigment molecules (e.g., chlorophylls) that absorb light at a particular wavelength and transfer the energy from the absorbed photon to a reaction center that initiates a series of electron-transport reactions.

phototroph: Organism capable of photosynthesis.

phylogeny: Study of the relatedness and evolutionary relationships among different groups of organisms.

phytoplankton: Free-floating, microscopic photosynthetic organisms (e.g., algae, cyanobacteria, and dinoflagellates) found in the surface waters of marine and freshwater environments.

post-translational modification: Any of several chemical modifications (e.g., phosphorylation, disulfide bond formation, cleavage of inactive sequence) involved in converting a newly translated amino acid sequence into a functional protein.

post-translational regulation: Process that controls the expression of gene products in cells by influencing the conversion of a newly translated amino acid sequence into a functional protein.

***Prochlorococcus*:** Type of unicellular cyanobacterium that is an extremely abundant primary producer in the world's oceans. *Prochlorococcus* is the smallest known oxygenic phototroph. Its abundance and phototrophic metabolism make it important in global carbon cycling through CO_2 fixation.

prokaryote: Single-celled organism lacking a membrane-bound, structurally discrete nucleus and other subcellular compartments. Bacteria and archaea are prokaryotes. *See also eukaryote.*

protein: Large molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene that codes for the protein. Proteins maintain distinct cell structure, function, and regulation.

protein complex: Aggregate structure consisting of multiple protein molecules.

proteome: Collection of proteins expressed by a cell at a particular time and under specific conditions.

proteomics: Large-scale analysis of the proteome to identify which proteins are expressed by an organism under certain conditions. Proteomics provides insights into protein function, modification, regulation, and interaction.

protists: Microscopic, eukaryotic organisms that have simple cellular organization. Protists include plant-, animal-, and fungus-like organisms that range in function from photosynthetic primary producers (e.g., green algae and diatoms) to predators and parasites.

protozoa: Single-celled, eukaryotic microorganisms that use cellular appendages called flagella to propel them through their environments.

provenance data: Data describing all the details of the experiment environment (e.g., manipulation of samples; software, tools, and methods used to conduct the experiment) so that researchers can visualize the experimental process and potentially reproduce the results of a specific experiment.

quality assurance (QA): Approach used to ensure that data systems will perform to a required standard for quality.

quality control (QC): Methods used to determine if the products of a process meet or exceed a defined standard for quality.

quantitative trait loci: All the DNA regions within a genome associated with the different genes that influence a particular complex trait.

recalcitrance: Natural resistance of plant cell-wall materials to physical and biological deconstruction.

regulator: Protein (e.g., a repressor) that controls the expression or activity of other molecules in a cell.

regulatory elements: Segments of the genome (e.g., regulatory regions, genes that encode regulatory proteins or small RNAs) involved in controlling gene expression.

regulatory circuit: *See gene regulatory network*.

regulatory region or sequence: Segment of DNA sequence to which a regulatory protein binds to control the expression of a gene or group of genes that are expressed together.

regulon: Set of operons controlled by the same regulator. Operons belonging to the same regulon can be located in different regions of a genome.

respiration: Series of biochemical redox reactions in which the energy released from the oxidation of organic or inorganic compounds is used to generate cellular energy in the form of ATP.

ribosomal RNA (rRNA): Specialized RNA found in the catalytic core of the ribosome, a molecular machine that synthesizes proteins in all organisms.

RNA (ribonucleic acid): Molecule that plays an important role in protein synthesis and other chemical activities of the cell. RNA's structure is similar to that of DNA. Classes of RNA molecules include messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and other small RNAs, each serving a different purpose.

Robetta server: Source of automated tools for analyzing and predicting protein structures.

RuBisCo (Ribulose-1,5-bisphosphate carboxylase/oxygenase): Enzyme that catalyzes the first major step of photosynthetic carbon fixation by adding a molecule of carbon dioxide to a short 5-carbon sugar called ribulose bisphosphate. The resulting 6-carbon sugar is split into two 3-carbon molecules that can be used to build larger sugar molecules. RuBisCo also catalyzes photorespiration, which releases CO_2.

scalability: Ability of a computer system to respond to increased demands.

schema: Description of the structure and organization of all the elements of a database.

semantic Web technologies: Technologies based on a common set of design principles that improve the efficiency of searching and sharing information on the Web by making Web content, which is designed to be read by humans, computer readable.

shotgun sequencing: Common approach to sequencing microbial genomes that involves breaking the genome into random fragments, which are cloned into vectors and sequenced. Computational analysis is used to compare all DNA sequence reads from random fragments and assemble the entire genome by aligning overlapping sequences.

signal-transduction pathway: Series of biochemical reactions that receive extracellular chemical signals. These signals are transmitted and amplified within the cell and ultimately used to stimulate or repress a certain type of molecular activity (e.g., gene expression).

simulation: Combination of multiple models into a meaningful representation of a whole system that can be used to predict how the system will behave under various conditions. Simulations can be used to run in silico experiments to gain first insights, form hypotheses, and predict outcomes before conducting more expensive physical experiments.

single nucleotide polymorphisms (SNPs): DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered.

stimulon: Set of genes controlled by the same stimulus.

synthetic biology: Field of study that aims to build novel biological systems designed to carry out particular functions by combining different biological "parts" or molecular assemblies.

system architecture: Conceptual design depicting how data and services are partitioned and linked among the different components of interconnected database systems.

systems biology: Use of global molecular analyses (e.g., measurements of all genes and proteins expressed in a cell at a particular time) and advanced computational methods to study how networks of interacting biological components determine the properties and activities of living systems.

systems microbiology: Systems biology approach that focuses on understanding and modeling microorganisms at molecular, cellular, and community levels.

taxa: Categories (e.g., phylum, order, family, genus, or species) used to classify animals and plants (singular: taxon).

taxonomy: Hierarchical classification system for naming and grouping organisms based on evolutionary relationships.

terabyte: Unit of computer storage representing one trillion (or 10^{12}) bytes.

transcript: Messenger RNA molecule (mRNA) generated from a gene's DNA sequence during transcription.

transcription: Synthesis of an RNA copy of a gene's DNA sequence; the first step in gene expression. *See also translation.*

transcription factor: Protein that binds to regulatory regions in the genome and helps control gene expression.

transcription start site (TSS): Position within the DNA sequence of a gene where the enzyme RNA polymerase initiates synthesis of mRNA.

transcriptomics: Global analysis of expression levels of all RNA transcripts present in a cell at a given time.

transfer RNA (tRNA): A class of small RNA molecules that have triplet nucleotide sequences that are complementary to the triplet nucleotide coding sequences of mRNA. During protein synthesis, each tRNA bonds with a particular amino acid that is added to the growing amino acid chain as specified by the order of nucleotides in the mRNA.

translation: Process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids.

transporter: Protein that transports a molecule from one location to another; in most cases, transporters are membrane proteins that control the movement of molecules in and out of cells.

vertical gene transfer: Inheritance or passing of genetic material from one generation to another. *See also horizontal gene transfer.*

vertical queries: Queries that span multiple data levels (e.g., from correlating climate data and habitats to genes found in different samples).

visualization: Representation of data using images that add meaning and facilitate user access, navigation, and retrieval of data.

wiki ("what I know is"): A method for a community to collectively accumulate knowledge.

xylem: Water-carrying tissue in vascular plants that gives stalks and stems rigidity. Xylem is a major component of wood where the cells of this tissue have thick, lignin-rich walls.

yeast two-hybrid (Y2H): Method for studying and identifying novel interactions between a protein of interest and other proteins. The protein of interest is fused to one of two domains of a transcription-activating molecule. The second domain is fused to potential binding partners of the protein of interest. When the protein of interest interacts with its binding partner, the two domains of the transcription-activating molecule come together and initiate the expression of a reporter enzyme that carries out some characteristic functionality (e.g., confers antibiotic resistance, produces a blue color).

Appendix 10
List of Web Addresses

URLs of Some Research Programs, Software Tools, Databases, and Policies Relevant to the GKB

ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/)

BioCyc (http://biocyc.org)

BioEnergy Science Center (DOE BESC; http://bioenergycenter.org)

BRaunschweig ENzyme DAtabase (BRENDA; http://www.brenda-enzymes.info)

Carbohydrate-Active enZYmes database (CAZy; http://www.cazy.org)

Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA; http://camera.calit2.net)

Comprehensive Microbial Resource (CMR; http://cmr.jcvi.org)

Firegoose (http://gaggle.systemsbiology.net/docs/geese/firegoose/)

Gaggle (http://gaggle.systemsbiology.net/docs/)

GenBank (http://www.ncbi.nlm.nih.gov/Genbank)

Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo)

Gene Ontology (GO; http://www.geneontology.org)

Genomes Standard Consortium (http://gensc.org)

Global Ocean Sampling (GOS; http://collections.plos.org/plosbiology/gos-2007.php)

Great Lakes Bioenergy Research Center (DOE GLBRC; http://www.greatlakesbioenergy.org)

GTL Information and Data Sharing Policy (http://genomicsgtl.energy.gov/datasharing/GTLDataPolicy.pdf)

Human Proteome Organization (http://www.hupo.org)

Innovative and Novel Computational Impact on Theory and Experiment program (DOE INCITE; http://www.sc.doe.gov/ascr/INCITE)

International Society for Computational Biology (http://www.iscb.org)

Joint BioEnergy Institute (DOE JBEI; http://www.jbei.org)

Joint Genome Institute (DOE JGI; http://www.jgi.doe.gov)

JGI's Integrated Microbial Genome with Metagenome database (DOE IMG/M; http://img.jgi.doe.gov/cgi-bin/pub/main.cgi)

Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/)

MetaCyc (http://www.metacyc.org)

MicrobesOnline (http://www.microbesonline.org)

National Institute of Health's Human Microbiome Project (http://nihroadmap.nih.gov/hmp)

Open Biomedical Ontologies (OBO) Foundry (http://www.obofoundry.org)

Open Source Initiative (http://www.opensource.org)

Pathema (http://pathema.jcvi.org)

Phytozome (http://www.phytozome.net)

PNNL Proteomics Software Tools and Data (http://ncrr.pnl.gov, http://ober-proteomics.pnl.gov, http://omics.pnl.gov)

PromScan (http://www.promscan.uklinux.net)

Proteomics Research Information Storage and Management
(DOE PRISM; http://ncrr.pnl.gov/about/process.stm)

RCSB Protein Data Bank (PDB; http://www.rcsb.org/pdb/home/home.do)

RegTransBase (http://regtransbase.lbl.gov/cgi-bin/regtransbase?page=main)

Rfam (http://rfam.sanger.ac.uk/)

RibEx (http://132.248.32.45:8080/cgi-bin/ribex.cgi)

Robetta (http://robetta.org)

Scientific Discovery through Advanced Computing (DOE SciDAC; http://www.scidac.gov)

The SEED (http://www.theseed.org)

***Shewanella* Federation** (http://www.shewanella.org)

SourceForge (http://sourceforge.net)

Systems Biology Markup Language (http://sbml.org)

Taverna (http://www.taverna.org.uk)

Tractor_DB (http://www.ccg.unam.mx/Computational_Genomics/tractorDB/)

UniProtKB/Swiss-Prot (http://www.ebi.ac.uk/swissprot/)

University of Georgia's Complex Carbohydrate Research Center (http://www.ccrc.uga.edu)

VISTA (http://genome.lbl.gov/vista/)

Acronyms and Abbreviations

A² authorities and accountabilities

API application programming interface

BRENDA BRaunschweig ENzyme DAtabase

CAMERA Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

CARO Common Anatomy Reference Ontology

CAZy Carbohydrate-Active enZYmes database

CBoL Consortium for the Barcode of Life

CBP consolidated bioprocessing

CheBI Chemical Entities of Biological Interest

ChIP chromatin immunoprecipitation

CL cell type ontology

CMR Comprehensive Microbial Resource

CO_2 carbon dioxide

Co-IP co-immunoprecipitation

CPRO cardio pulmonary respiratory ontology

CRISPR clustered regularly interspaced short palindromic repeats

DBMS database management system

DNA deoxyribonucleic acid

DOE Department of Energy

EM electron microscopy

EMSA electrophoretic mobility shift assay

EML environmental markup language

EnvO environmental ontology

ESG Earth System Grid II

ExPAsy Expert Protein Analysis System

ESTs expressed sequence tags

FIGfams protein families generated by the Fellowship for Interpretation of Genomes

FISH fluorescence in situ hybridization

FMA Foundational Model of Anatomy

FMP Foundational Model of Physiology

FTP file transfer protocol

GEBA Genomic Encyclopedia of Bacteria and Archaea

GEO Gene Expression Omnibus

GH glycoside hydrolase

GKB GTL Knowledgebase

GO Gene Ontology

ESTs expressed sequence tags

GOS Global Ocean Sampling

GSC Genomic Standards Consortium

GT glycosyl transferase

GTL DOE's Genomics:GTL program (formerly Genomes to Life)

HUPO Human Proteome Organization

ICAT isotope-coded affinity tag

ICoMM International Census of Marine Microbes

IMG/M DOE Joint Genome Institute's Integrated Microbial Genome/ Metagenome database

INCITE DOE's Innovative and Novel Computational Impact on Theory and Experiment program

ISCB International Society for Computational Biology

ITRAQ isobaric tag for relative and absolute quantitation

JGI DOE's Joint Genome Institute

KEGG Kyoto Encyclopedia of Genes and Genomes

LANL DOE's Los Alamos National Laboratory

LBNL DOE's Lawrence Berkeley National Laboratory

LIMS Laboratory Information Management System

LLNL DOE's Lawrence Livermore National Laboratory

MALDI-TOF matrix-assisted laser desorption/ ionization–time of flight

MeDICi Middleware for Data-Intensive Computing

MGED Microarray and Gene Expression Data

MIAME Minimum Information about a Microarray Experiment

MIAPE Minimum Information about a Proteomics Experiment

MIGS Minimum Information about a Genome Sequence

MIMIx Minimum Information Required for reporting a Molecular Interaction Experiment

MIMS Minimum Information about a Metagenome Sequence

mRNA messenger RNA

MS mass spectrometry

NAD nicotinamide adenine dinucleotide

NCBI National Center for Biotechnology Information

ncRNA noncoding RNA

NDB Nucleic Acid Database

NMR nuclear magnetic resonance

NPP net primary production

OBER Office of Biological and Environmental Research

OBO Open Biomedical Ontologies

ORNL DOE's Oak Ridge National Laboratory

OWL Web ontology language

PaTO phenotypic quality ontology (formerly phenotype and trait ontology)

PDB Protein Data Bank

PEDRo Proteome Experimental Data Repository

Pfam Protein Families Database

PI principal investigator

PMBMS pyrolysis molecular beam mass spectrometry

PNNL DOE's Pacific Northwest National Laboratory

PRISM PNNL's Proteomics Research Information Storage and Management system

PRO protein ontology

PSI Proteomics Standards Initiative

QA quality assurance

QC quality control

R² roles and responsibilities

RNA ribonucleic acid

RNAO RNA ontology

rRNA ribosomal RNA

RuBisCo ribulose-1,5-bisphosphate carboxylase/ oxygenase

SBML systems biology markup language

SciDAC DOE's Scientific Discovery through Advanced Computing program

SCOP Structural Classification of Proteins

SILAC stable isotope labeling with amino acids in cell culture

SNPs single nucleotide polymorphisms

SO sequence ontology

SOA service-oriented architecture

SOAP (formerly Simple Object Access Protocol)

SOPs standard operating procedures

SSU rRNA small subunit rRNA

TF transcription factor

TFBS transcription factor–binding site

TIGR The Institute for Genomic Research

TRN transcriptional regulatory networks

tRNA transfer RNA

TSSs transcription start sites

UDDI Universal Description, Discovery, and Integration

UniProt Universal Protein Resource

UPSIDE uniform principle for sharing integral data and materials expeditiously

VIMSS Virtual Institute for Microbial Stress and Survival

XML Extensible markup language

Y2H Yeast two-hybrid

Printed with soy ink on recycled paper

U.S. DEPARTMENT OF ENERGY OFFICE OF SCIENCE 2009