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.

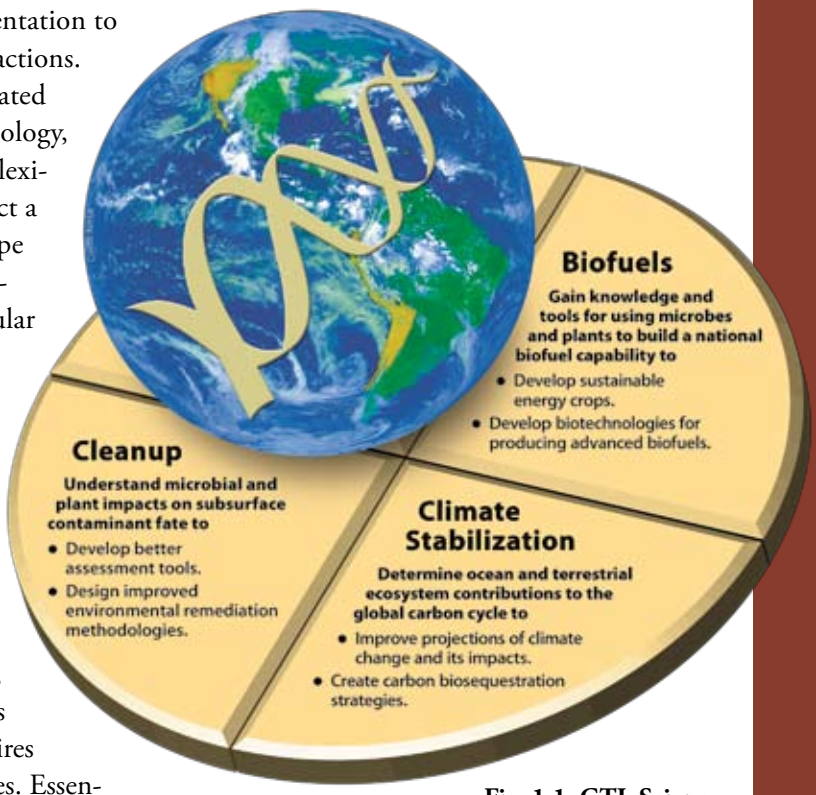


Fig. 1.1. GTL Science for DOE Missions.

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

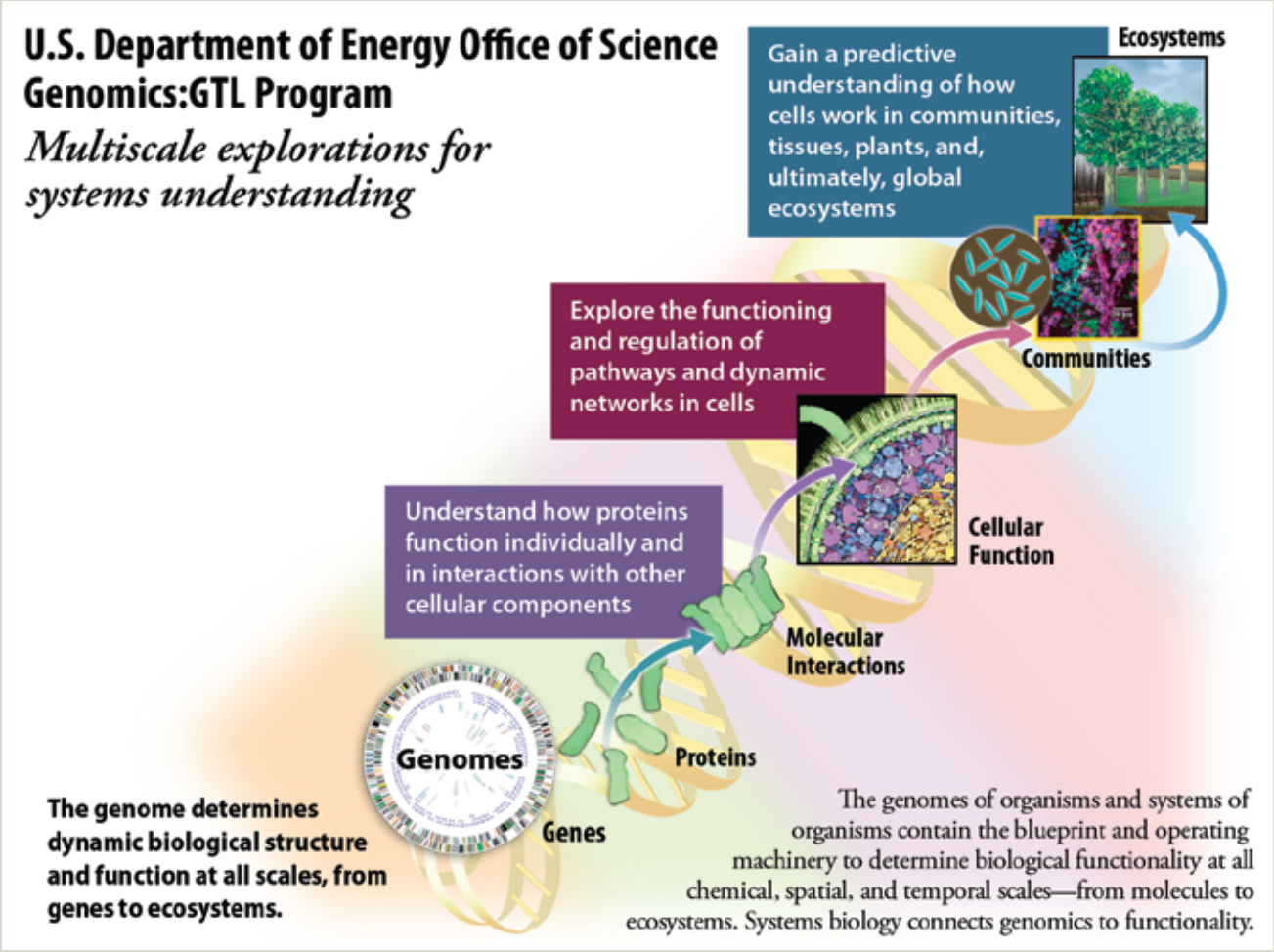


Fig. 1.2. Multiscale Explorations for Systems Understanding.

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.

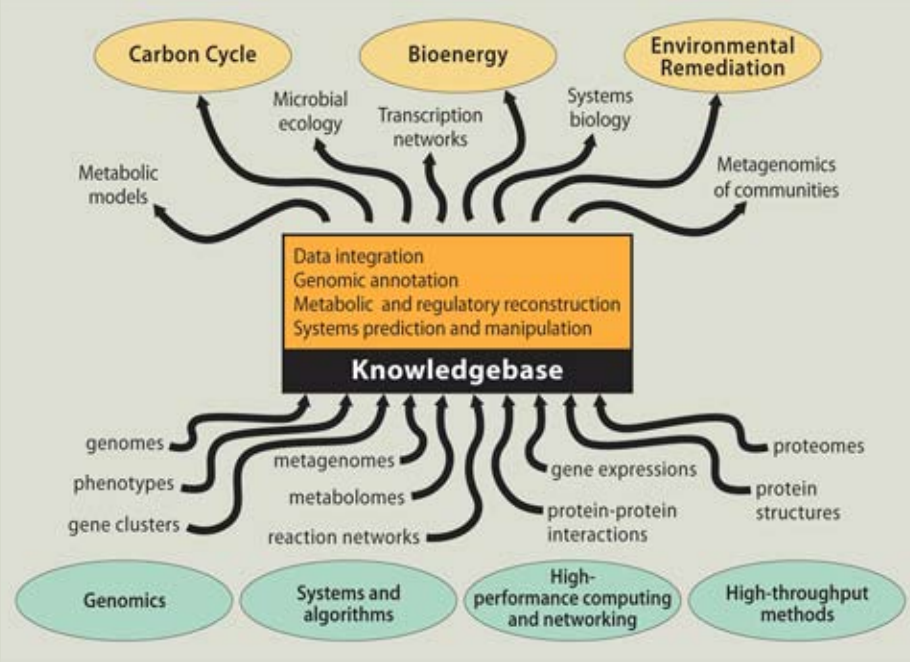
Capabilities for integrating and synthesizing various classes of existing data and data that will be acquired by current and developing technologies are recognized as major unmet needs and thus impediments to the advancement of systems biology (American Academy of Microbiology 2004). Some capabilities 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)

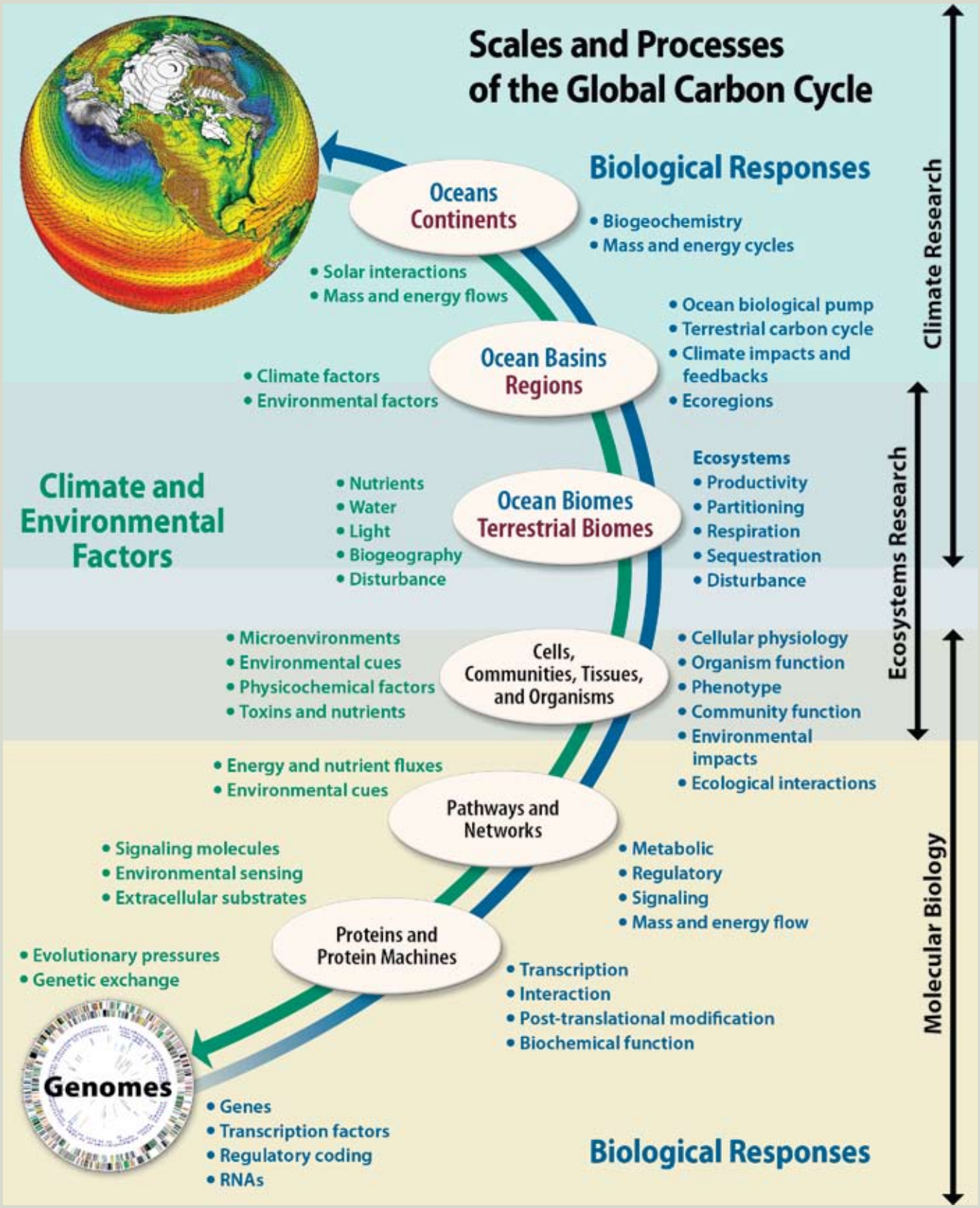
Box 1.1

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 Basins Regions
• Ocean biological pump
• Terrestrial carbon cycle
• Climate impacts and feedbacks
• Ecoregions
• Climate factors
• Environmental factors
Climate and Environmental Factors
Ocean Biomes Terrestrial Biomes
• Nutrients
• Water
• Light
• Biogeography
• Disturbance
Ecosystems
• Productivity
• Partitioning
• Respiration
• Sequestration
• Disturbance
Cells, Communities, Tissues, and Organisms
• Microenvironments
• Environmental cues
• Physicochemical factors
• Toxins and nutrients
• Cellular physiology
• Organism function
• Phenotype
• Community function
• Environmental impacts
• Ecological interactions
Pathways and Networks
• Energy and nutrient fluxes
• Environmental cues
• 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

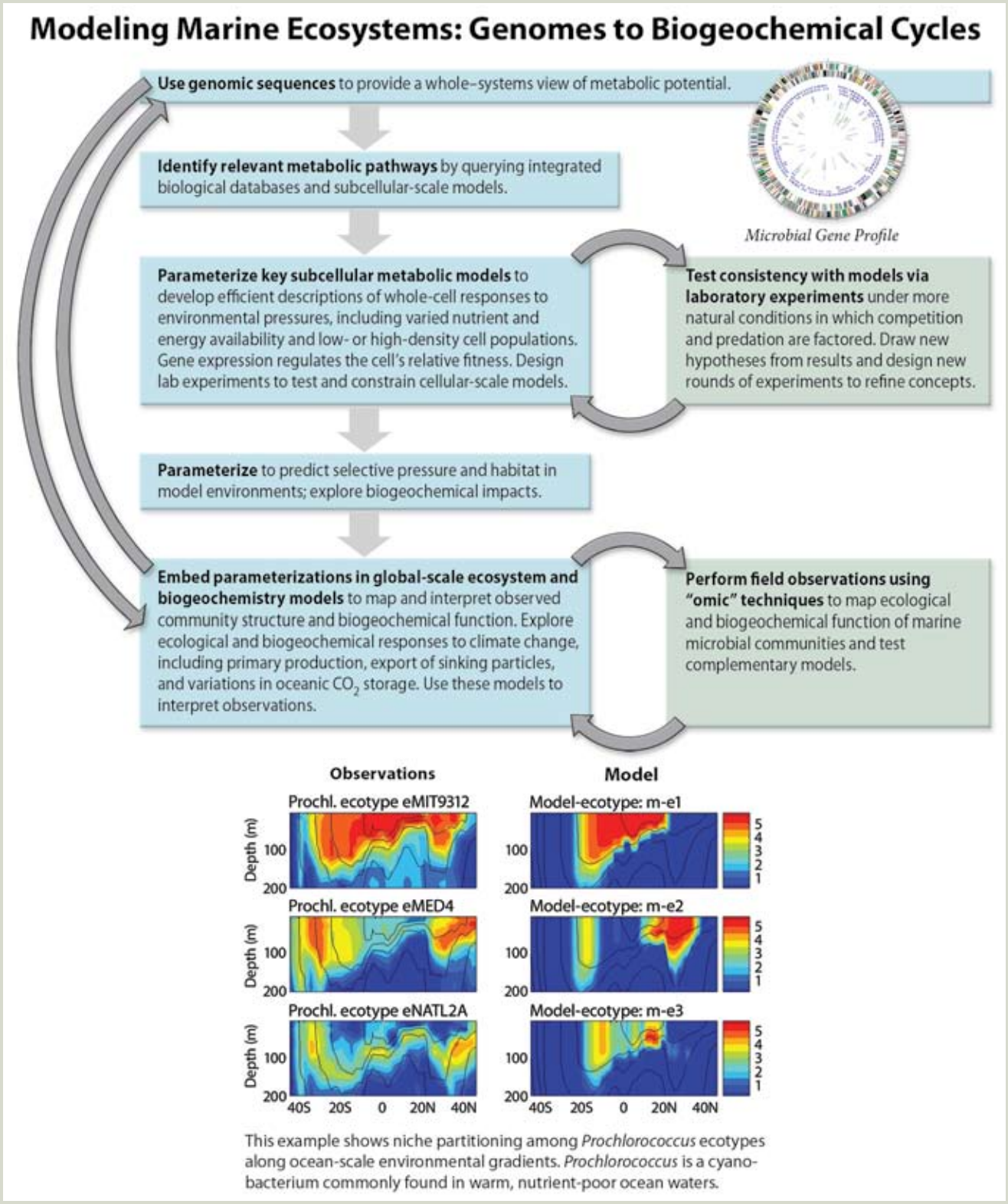


Fig. 1.3. Modeling Marine Ecosystems: Genomes to Biogeochemical Cycles. Depicted are observed and modeled distributions of ecotypes of *Prochlorococcus* [log (cells ml^{-1})]—one of the important marine phytoplankton groups—along a meridional transect in the Atlantic Ocean. Black lines indicate isotherms. 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/

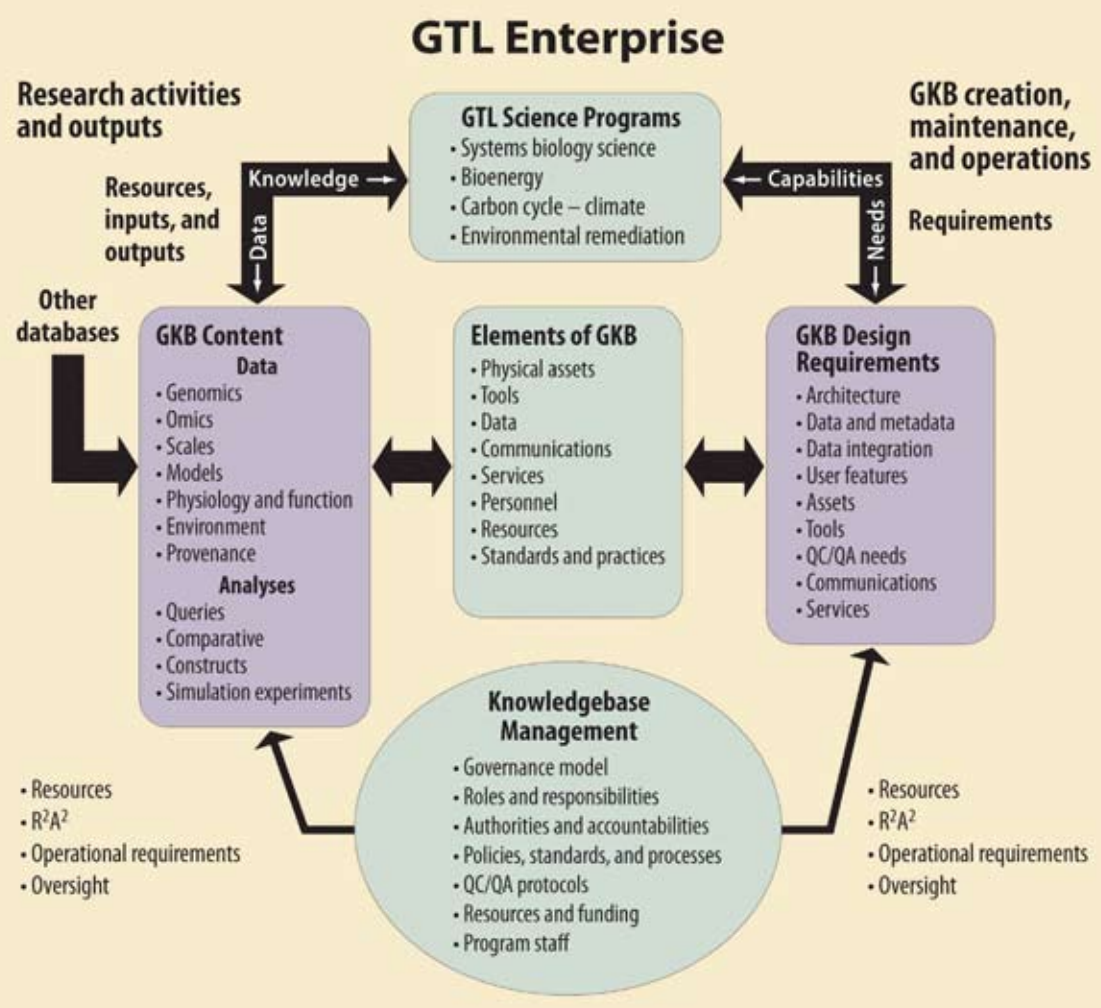
Box 1.3

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