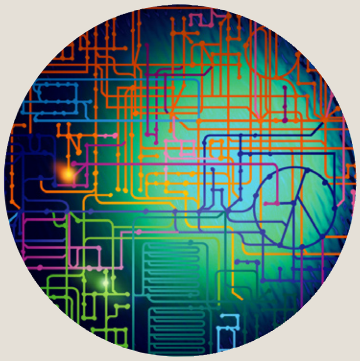
3.0. GTL Research Program

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



GTL Research Program

The mission challenges that GTL must meet are global in scale and among the most complex in biology today. The scientific knowledge required for finding solutions establishes the need for speed and performance in our capabilities for biological research and forces the development of new approaches and technologies. This chapter outlines GTL's mission science goals and the research strategy to develop the capabilities needed for achieving them. Highlights of ongoing research demonstrate progress in developing and using advanced technologies, computing, and the richness of science.

GTL's Ultimate Scientific Goal

Achieve a predictive, systems-level understanding of microbes to help enable biobased solutions to DOE missions

Science and Technology Milestones

1: Develop techniques to determine the genome structure and functional potential of microbes and microbial communities.

2: Develop methods and concepts needed to achieve a systems-level understanding of microbial cell and community function, regulation, and dynamics.

3: Develop the knowledgebase, computational methods, and capabilities to advance understanding and prediction of complex biological systems.

4: Design and build user facilities to accelerate GTL microbial systems biology.

3.1. Background and Approach

We ultimately seek to understand microbial systems at a level sufficient for predicting and confidently manipulating biological function. Building on a global perspective provided by whole-genome sequences, the GTL program provides the foundation for systems microbiology by integrating new experimental, analytical, and computational approaches toward this end. GTL analyzes critical microbial properties and processes on three fundamental systems levels.

- **Molecular.** Focusing on genes, proteins, multicomponent protein complexes, and other biomolecules that provide structure and perform the cell's functions.
- **Whole cell.** Investigating how molecular processes, networks, and subsystems are controlled and coordinated to enable such complex cellular processes as growth and metabolism.
- **Microbial community.** Understanding the ecophysiology of diverse microbes and how they interact to carry out coordinated complex ecosystem processes, enabling them to both respond to and alter their environments.

Microbes function as part of structured communities that give them enormous biochemical diversity and allow them to adapt to extremes of environmental conditions. While individual microbes are among the simplest of organisms, their species diversity and community functionality are complex, and our research capabilities must analyze those many intricacies. We must be able to assess microbes in a variety of environments and develop methods to deal with vast numbers, small size, extreme genetic diversity and dynamics, wide-ranging biochemical and physiological properties, and complex interactions and community structures (see sidebar, The Microbial World, p. 13). Meeting DOE's mission challenges often will force us to analyze microbes from diverse environments such as anaerobic conditions. Instruments we develop must operate in or maintain conditions of these special and extreme environments required by microbes.

Systems biology and the study of microbes on the scale proposed for the GTL research program are making the following demands on technology capabilities:

- Automation and parallel processing to increase throughput
- Reduced sample sizes to increase speed and system capacity and lower costs of reagents
- Improved resolution and sensitivity to accommodate the small sizes of microbes and the fine structure in microbial communities
- Integration and analysis of very large data sets
- Innovations in measurement modalities across critical variables

Hence the GTL strategy rests on DOE's hallmark capabilities.

- **Advanced technologies.** GTL will scale up technologies in high-throughput production user facilities to comprehensively analyze the makeup and functions of living systems.
- **Computing and information technologies.** GTL will operate within an infrastructure containing data, tools, models, and communication resources for systems biology.
- **Multidisciplinary teams focused on strategic science goals and managed for results.** GTL will make its resources available to all scientists, enabling them to practice systems microbiology and thus involving the whole scientific community in important national problems (see 1.3.4. Bridging the Gap Between Big and Small Science—The Need for a Third Model, p. 9). Proof-of-principle experiments in systems biology, technology prototyping, and piloting are in progress, and a community of scientists is becoming conversant with DOE mission challenges, microbes, and systems biology.

3.1.1. Phase I Implementation: Current GTL and Related Projects

The GTL program, begun in 2002, is in its initial phase, making the transition from genomics to systems biology (see 1.3.6.1. Three-Phase Implementation of the GTL Program, p. 10). GTL-funded research projects collectively have set out to decipher, on a global scale, the molecular biochemistry and mechanisms for regulation of microbial processes (for more information, visit the GTL web site, www.doegenomestolife.org).

The GTL program currently funds projects in academia, national laboratories, and the private sector. Contributions to the program are from experts in the life sciences, computing, mathematics, physics, chemistry, geology, oceanography, engineering, project management, and communications (see Appendix E. GTL-Funded Projects, p. 245).

3.2. Scientific Goals and Milestones

GTL's ultimate scientific goal is to achieve a predictive, systems-level understanding of microbes to help enable biobased solutions to DOE mission challenges.

Key Questions in Biology

An understanding of molecular mechanistic processes of natural systems will provide needed insights to support policy and the design of systems to support applications. This will allow us to begin answering some of the most challenging questions in biology, including:

- How is the information contained within genomes and metagenomes translated into living, functioning, and self-perpetuating life forms and systems of life forms? What are the principles and details?
- How does a microbe or microbial community sense and respond to its dynamic environment?
- What are the life strategies of microbes, from the molecular to community levels?
- How are information, energy, and material managed and manipulated within biological systems?
- What are the principles and details of biological molecular, pathway, and system functionality, structure, and control?

3.2.1. Missions Science Goals

As described in Missions Overview and related appendices, each mission has a distinct endpoint and overall set of subsidiary science goals.

- **Energy.** Understand the principles underlying the structural and functional designs of microbial and molecular systems, and develop the capability to model, predict, and engineer optimized enzymes and microorganisms for the production of such biofuels as ethanol and hydrogen.
- **Environmental Remediation.** Understand the processes by which microbes function in the earth's subsurface, mechanisms by which they impact the fate and transport of contaminants, and the scientific principles of bioremediation based on native microbial populations and their interactions with the environment. Develop methods to relate genome-based understanding of molecular processes to long-term conceptual and predictive models for simulating contaminant fate and transport and development of remediation strategies.
- **Carbon Cycling and Sequestration.** Understand the microbial mechanisms of carbon cycling in the earth's ocean and terrestrial ecosystems, the roles they play in carbon sequestration, and how these processes respond to and impact climate change. Develop methods to relate genome-based microbial ecophysiology (functionality) to the assessment of global carbon-sequestration strategies and climate impacts.

3.2.2. Science and Technology Milestones

The technical strategy for timely achievement of these systems-level goals rests on four major complementary milestones. Pursued simultaneously in a coordinated way, the first three establish capabilities and concepts that are a starting point and should evolve along with technical progress. These milestones can be scaled up in the facilities discussed in the fourth milestone.

The functional properties of all biological systems ultimately are specifically encoded in their genomes and are of two general aspects:

- Potential functions represented in the set of genes that each genome contains and
- Control apparatus required to program the regulated expression of those genes.

Milestone 1 deals with the makeup and characteristics of the gene-encoded parts of microbes and communities, while Milestone 2 deals with function; we must have both in context to derive a systems understanding. Milestones 3 and 4 describe computing and facilities, which will provide the engine to attack these large problems. The first milestone underlies the rationale for the Protein Production and Characterization Facility and the Molecular Machines Facility. The second milestone discusses the Proteomics and Cellular Systems facilities.

3.2.2.1. Milestone 1: Develop Techniques to Determine the Genome Structure and Functional Potential of Microbes and Microbial Communities

3.2.2.1.1. Component A. Microbial Sequences and Protein Characteristics

3.2.2.1.1.1. Background and Science Needs

Proteins are the chemically and physically active products of virtually all genes. Highly dynamic and shifting in amount, modification state, higher-order association, and subcellular localization, proteins carry out the primary functions of a cell in response to intracellular and extracellular signals.

For a systems understanding of microbes, we first must understand the panoply of proteins the genome is capable of producing. GTL's first challenge in studying mission-relevant microbes and microbial communities is to determine the system's genetic makeup and the extent and patterns of genetic diversity. This is especially true when many identified coding genes are unknown, microbes are unculturable, or only gene sequence is in hand (e.g., metagenomic experiments involve determining the genetic sequence of a whole community of microbes).

Unknown genes are the first target. With a mature database of thousands of microbes available within a decade, comparative genomics, phylogenetic analysis, and sophisticated computational annotation will provide an increasingly complete set of gene functional assignments. In the interim and to reach that end state, we must be able to perform functional annotations based on information from proteins produced from sequence and analyzed biophysically and biochemically in vitro. GTL's ultimate goal, however, goes beyond simple assignment to achieving a mechanistic structural and functional understanding of proteins and molecular machines that can form the basis for comprehensive and predictive systems models.

The availability of gene sequence and proteins allows the generation of various affinity reagents. Development of affinity methods and reagents from produced proteins will open the door to identifying and tracking microbes and specific proteins in complex and dynamic microbial systems. Affinity reagents also can be used to manipulate (activate or inactivate) proteins, capture and track them, and determine their relative locations through a variety of sensitive analytical methods for understanding and visualizing protein structure, function, and behavior. (An extension of this discussion is in the Protein Production and Characterization Facility chapter, section 5.1.3. Development of Methods for Protein Production, p. 118.) Specific milestone objectives are set forth below.

- **Genome Sequences.** Develop methods for sequencing uncultivated microbes and microbial communities and identifying the extent and patterns of genetic diversity and evolution, including:
  - Sequence-assembly methods.
  - Single-cell in situ sequencing for verification and environmental experimentation.
- **Protein Characteristics.** Develop methods and concepts to understand the range and characteristics of proteins encoded in genomes, including:
  - Refined computational annotation for primary gene assignment and putative protein function.
  - Advanced comparative analysis and methods based on evolutionary relationships to understand the functions of newly discovered genes and proteins using the comprehensive GTL Knowledgebase.
  - Biophysical and biochemical analyses of proteins produced directly from genome sequences for more rigorous assignment of gene function and as a starting point for mechanistic understanding of microbial capabilities, function, and control. This capability provides a cost-effective and rapid alternative to culturing for the determination of hypotheticals and unknowns.
  - Application of these analytical capabilities to genetically modified proteins to assist in derivation of design principles and optimization of microbial and protein function.
  - Affinity methods and reagents for locating and analyzing proteins and complexes outside living cells and dynamically inside living cells and for identifying and tracking microbes and specific proteins in complex microbial systems.

3.2.2.1.1.2. Computation Needs

Computational challenges in characterizing the composition and functional capability of microorganisms range from "simple" data management to complex data analysis, integration, and use. New algorithms for DNA sequence assembly, as well as better use of current state-of-the-art methods and annotation, will be required to analyze multiorganism sequence data; new modeling methods will be needed to predict the behavior of microbial communities. Computational research must develop methods to

- Deconvolute mixtures of genomes sampled in the environment and identify individual microbial genomes.
- Facilitate multiple-organism, shotgun-sequence assembly.

- Improve comparative approaches to microbial-sequence annotation and use them in conjunction with data generated by high-throughput experimentation to more accurately assign functions to genes and proteins.
- Accomplish pathway reconstruction from sequenced or partially sequenced genomes to evaluate the combined metabolic capabilities of heterogeneous microbial populations.

3.2.2.1.2. Component B. Molecular Complexes

3.2.2.1.2.1. Background and Science Needs

Most proteins do not act alone but instead are organized into molecular complexes (machines) that carry out activities needed for metabolism, communication, growth, and structure. GTL's first milestone includes the creation of capabilities for comprehensively identifying, characterizing, and beginning to understand multiprotein complexes. These studies will help build the essential knowledgebase, and the stage will be set for linking proteome dynamics and architecture to cellular and community functions.

Identifying and characterizing multiprotein complexes on a genome-wide scale will require new tools and research strategies designed to increase throughput, reliability, accuracy, and sensitivity. While RNA measurements, such as microarrays, can give us a notion of which machines might form, the importance of understanding post-transcriptional and post-translational regulation requires direct knowledge of proteins and their interactions. Also, new tools for characterizing these complexes must bridge current size and resolution gaps between the high-resolution technologies for studying single proteins and those suitable for very large protein assemblies and cellular ultrastructures that are more amenable to just-emerging nanoscale structural techniques (see Table 3. Technology Development Roadmap for Complex Identification and Characterization, p. 146).

An initial target for GTL is to develop a suite of methods to isolate, identify, and characterize all essential protein complexes in a microbial system. Currently, only a few of the most stable and common protein complexes are well characterized, but data suggest that hundreds, if not thousands, of other complexes operate together to carry out cellular functions. Many important associations may be less stable, less abundant, and more dynamic. The near-term challenge is to develop methods to analyze the difficult ones. These most demanding protocols can be supported in a comprehensive way only with a technically and scientifically robust infrastructure. Providing the necessary infrastructure and scaling up these capabilities in facilities will enable scientists to rapidly generate a draft protein-machinery map of a typical microbe of interest to DOE.

An important aspect of understanding the assembly, stability, and function of protein complexes is the high-throughput characterization of protein-protein proximity and interfaces within complexes and between interacting complexes. When coupled with other information about structure and interrelationships among proteins, this characterization will provide a comprehensive database for understanding spatial and temporal hierarchies in the assembly of protein complexes. Ultimately, this analysis will reveal the internal, transmembrane, and extracellular structure of cells and bring understanding of how assembly and disassembly of these complexes are organized and controlled. Data on coincident expression and cellular or subcellular localization can powerfully constrain possible functions for a given multiprotein complex. By coupling localization and colocalization information with genetic and biochemical data from diverse sources, scientists can postulate and then test the contributions of specific complexes to a cell's survival and behavior. High-throughput implementation of new and existing technologies will be needed to achieve these goals (Appella and Anderson 2005, in press; Pennisi 2003).

3.2.2.1.2.2. Molecular Complexes. Develop Capabilities for a Predictive Understanding of Protein Interactions and the Resulting Structure and Properties of Molecular Complexes

- Discover and define the repertoire of molecular interactions and multimolecular complexes. New methods will be required to isolate and analyze transient and rare complexes.

- Develop predictive methods to define experimental conditions favoring the occurrence of condition-dependent and transitory complexes to assist in their capture.
- Determine the structure of complexes, with localization of components and characterization of their reaction interfaces. Establish high-throughput methods to define the protein-protein interfaces within and between complexes.
- Determine the cellular and subcellular localization and colocalization of protein complexes, including their conditional and temporal variations. Define physical relationships among protein complexes and integrate this information with candidate functions.
- Develop principles, theory, and predictive models for the structure, function, assembly, and disassembly of multiprotein complexes. Test predictions of these models in experimental systems and apply them to optimization of functions for applications.
- Correlate information about multiprotein complexes with relevant structural-fold data generated in the NIH Protein Structure Initiative to better understand the geometry, organization, and function of these protein machines.

3.2.2.1.2.3. Computation Needs

- Identify and characterize life's multiprotein complexes, involving substantial computational demands and ranging from sophisticated data analysis to atomic-scale simulations of protein interactions. Meeting these needs will require the development of new algorithms and databases and the use of high-performance computers.
- Adapt and develop databases and analysis tools for integrating experimental data on protein complexes measured with different methods under varied conditions.
- Develop novel approaches and methods for automatically identifying protein functional modules based on high-throughput genomic, proteomic, and metabolomic data.
- Develop algorithms for integration of diverse biological databases including transcriptome and proteome measurements, as well as functional and structural annotations of protein-sequence data to infer complex formation and function.
- Develop modeling capabilities for simulating multiprotein complexes and for predicting the behavior of protein complexes in cell networks and pathways.

See section 3.3. Highlights of Research in Progress to Accomplish Milestones, p. 55.

3.2.2.2. Milestone 2: Develop Methods and Concepts Needed to Achieve a Systems-Level Understanding of Microbial Cell and Community Function, Regulation, and Dynamics

3.2.2.2.1. Component A. Systems Analytical Measurements (Omics) of Microbes and Microbial Communities

3.2.2.2.1.1. Background and Science Needs

Of all the molecular components, the proteome is the most critical to measure comprehensively. The end result of genome transcription and expression, the proteome comprises the cell's working parts. Understanding its dynamic nature calls for methods to accurately, sensitively, and temporally monitor the conditional state of any organism's entire proteome, correlated to other cellular molecular species. This task will require greater completeness, resolution, and sensitivity than has been possible with conventional imaging and gel-based technologies. Providing a comprehensive view of proteome organization and dynamics promises to be a singularly important watershed of whole-genome biology for the coming decade because it will enable, inform, and enhance virtually all other molecular and cellular investigations (Falkowski and de Vargas 2004). As a starting point for studying regulatory networks, cell pathways, and metabolic interactions in microbial communities, such a comprehensive view would provide basic understanding of how an entire cell and community work.

This information would complement that derived by using capabilities developed under Milestone 1. Progress already has been achieved in the development of technologies with the resolving power, dynamic range, and sensitivity to rapidly measure a cell's proteome.

To develop ultimately a predictive understanding of these systems, the proteome must be analyzed in conjunction with the intracellular mix of RNAs, metabolites, and signaling molecules. Also requiring analysis is the extracellular and intercellular mix of environmental physicochemical variables; signaling molecules; metabolites and their metabolic intermediates (e.g., in syntrophy); genetic materials; and other microbial species and their genetic, phenotypical, and physiological makeup.

3.2.2.2.1.2. Community Structures and Processes: Science Needs

The core of systems biology is the ability to measure, in a coordinated way, all the cell's responses and functions as referenced to the genome sequence. Microbes have many mechanisms to position themselves relative to their environment's physicochemical variables and to each other to optimize microbial-community function. The dynamic and intimate nature of interactions in microbial communities is such a dominant phenomenology that community behavior, not just microbes acting alone, must be deciphered to develop a predictive understanding of microbial systems, even at the cellular level. These structured communities live in ocean water columns, on particulates or plant roots in soils, and on minerals in the deep subsurface of the earth and ocean. While initial analytical attempts necessarily will be global measurements of ensemble samples, the nature of interactions and behaviors in local niches will require the ability to make measurements that can spatially resolve (image) such variables in a single cell, within a community, and in a well-defined environment. A citation by the American Academy of Microbiology reads:

> There is a need to "develop technology and analysis capability to study microbial communities and symbioses holistically, measuring system-wide expression patterns (mRNA and protein) and activity measurements at the level of populations and single cells." (Stahl and Tiedje 2002)

Microbial communities essentially act as a multicellular organism, utilizing the function of individual components for the benefit of the whole, including functional flexibility and diversity (see sidebars, Quorum Sensing, p. 19, and Life in a Biofilm, p. 18). Microniches, in which microbes exhibit unique phenotypes, are formed within communities. In these communities, microbes find protection from the environment and communicate within and between populations, exchanging nutrients, regulatory and sensing molecules, metabolites, and genetic materials. They exhibit a wide variety of ecosystem interactions including syntrophy, commensalism, amensalism, predation, parasitism, mutualism, competition, and warfare. These complex functions and relationships can be analyzed only in a community context (Winzer, Hardie, and Williams 2002).

Success in achieving this milestone will set the stage for causally linking gene regulation, proteome composition, architecture, and dynamics with cellular and community function. The ultimate test for an accurate and useful understanding of causality in any system is the capacity to predict how the system will change when perturbed by new external or internal stimuli, in this instance including genetic changes. A long-term aim of GTL is to develop the theoretical infrastructure and knowledgebase for understanding the microbe and community at the proteome level, in multiprotein complexes and the pathways and structures they comprise, and in intermicrobe interactions and processes. (An extension of this discussion is in 5.3.1. Scientific and Technological Rationale, p. 156.)

This understanding will require the coupling of increasingly sophisticated models with experimental tests of predictions from models. The following are specific milestone objectives under Component A:

- Develop methods and concepts to understand microbial and microbial-community responses, interactions, and processes including:
  - Genomic basis and mechanisms underlying microbial-community structure, activities, functions, stability, adaptation, and succession.

- Dynamics of microbial populations identified via metagenomic analysis through time and under various perturbations.
- Presence and fluxes of key molecular species in cells and communities—proteins, RNA, metabolites, and signaling molecules.
- Global net microbial-community function via markers of its physicochemical presence—export of biomass, transformation of waste, and creation of energy.
- Ecophysiology—community structure; relationship to environment; members and their phenotypes, locations, and contributions to function.
- Key physicochemical environmental variables and mechanisms for microbial and community sensing and response.
- Effects of viruses and plasmids on community structure and dynamics and as members of the community as a whole.
- Extracellular processes and phenomenology [e.g., quorum sensing, positioning, biofilms, electron transfer, depolymerization, nutrient gathering (siderophores), and complexation].
- Lateral gene-transfer factors driving genomic plasticity and conditional expression (e.g., conjugation, transformation, and transfection).
- Identification of signal-transduction pathways.
- Microbial ecological interactions (e.g., syntrophy, commensalism, amensalism, predation, parasitism, mutualism, competition, and warfare). Understanding these processes will help us to understand the evolutionary dynamics of these systems.

3.2.2.2.2. Component B. Networks and Regulatory Processes

3.2.2.2.2.1. Background and Science Needs

Understanding gene regulatory networks is prerequisite for redesigning biological control systems required to solve a wide range of problems we can barely fathom today. Gene regulatory networks explicitly represent the causality of life systems. They explain exactly how genomic sequence encodes the regulation of expression of the large sets of genes that create the biological processes we observe, measure, and utilize to practical ends. It is at the system level of gene regulatory networks that we can address biological causality and provide a complete answer to why biological systems function as they do.

Regulatory processes govern which genes are expressed in a cell at any given time, the level of that expression, the resultant biochemical activities, and the cell's responses to diverse environmental cues and intracellular signals. This most fundamental domain of life—genomic control systems—is now within reach of the biosciences. Flexible and responsive, these genomic control systems consist essentially of hardwired regulatory codes that specify the sets of genes that must be expressed in specific spatial and temporal patterns in response to internal or external inputs. In physical terms, the control systems consist of thousands of modular DNA sequences, which receive and integrate multiple regulatory inputs in the form of proteins. These proteins recognize and bind to them, resulting in transfer of specific transcriptional instructions to the protein-coding genes they direct. The most important of all classes of such regulatory modules are those that control the activity of genes encoding the DNA-recognizing regulatory proteins themselves. These genes, and the control sequences of the genes to which their protein products bind, can be treated literally as networks of functional regulatory linkages. Each such linkage joins a regulatory gene to its target DNA regulatory sequence modules. For microbial systems, GTL will encompass comprehensive mapping of all these regulatory processes, including the cytoplasmic regulation that operates following gene expression of the functioning networks.

The regulatory genome is a logic-processing system. Every regulatory module encoded in the genome—that is, every node of every gene regulatory network—receives multiple disparate inputs and processes them in ways that can be represented mathematically as combinations of logic functions (e.g., "and" functions, "switch" functions, and "or" functions). At the system level, a gene regulatory network consists of assemblages of these information-processing units. Thus it is essentially a network of analogue computational devices, the functions of which are conditional on their inputs.

Major objectives for this milestone are to develop methods to discover the architecture, dynamics, and function of regulation; make useful computational models; and learn how to adapt and design them. To redesign these most potent of all biological control systems to produce desired functions, first we must be able to insert regulatory subcircuits—far beyond any simple gene insertions—into the target biology; second, we must understand the flow of causality in a genomically encoded gene regulatory network to design an effective means of altering it.

Gaining a comprehensive view of the architecture of microbial regulatory networks will not necessarily reveal how such networks really work, nor will it provide a solid basis for employing or modifying them in useful ways or designing new ones. Mastering the complexities of regulatory switches, oscillators, and more complex functions will require a predictive theoretical framework and computational horsepower, coupled with experimental resources to test and validate models. To meet this challenge, GTL will seek to nurture and accelerate emerging capabilities that include new concepts combined with relevant ideas from engineering, applied mathematics, and other disciplines.

Within this network-discovery portion of the milestone, one activity is to map related networks at multiple nodes across phylogeny based on comparison of genome sequences. Knowledge of comparative network structure and function is likely to produce insights into fundamental issues in biology, in addition to providing essential information for GTL's later phases. Initial tasks will be to identify and map core regulatory network components (e.g., regulons, operons, and sRNAs). Integral to this effort is the task of relating the regulatory apparatus to the groups of target genes they regulate and to whatever is known about the function of those target genes.

To map regulatory networks, several core technologies and approaches will be needed. Pilot studies will further define the best approach to use in genomes of varying sizes and structures. One such promising strategy is to use comparative genomics to initiate large-scale component identification, focusing on candidate regulatory sequences and their interacting regulatory proteins. Results from comparative sequence analysis would then be integrated with data from such other key technologies as large-scale gene-expression analysis, comprehensive loss-of-function and gain-of-function genetic analyses, and measures of in vivo protein-DNA interactions and proteome status, among others.

Other critical elements in network mapping will come from, for example, proteomic and metabolomic activities encompassed by Component A of this milestone or by specific adaptation of those technologies to regulatory network components. These elements include learning the composition of multiprotein complexes that assemble on DNA to regulate gene expression; learning the composition and regulatory actions of protein machinery that govern post-transcriptional and post-translational regulation; and determining subcellular localization of regulatory proteins and how localization changes as a function of circuit dynamics.

Vigorous application of a comprehensive genome-wide approach to network mapping in selected microbes has the potential to yield the first complete dissection of the regulatory networks that run a living cell. Regulatory networks in microbes employ many mechanisms distinct from both transcription and translation. Examples include active control of protein turnover, dynamic localization of regulatory and structural proteins, cell membrane processes, and complex phosphor-transfer pathways. Studying nontranscriptional systems, therefore, is critical for fully understanding regulatory mechanisms. The following are specific milestone objectives under Component B:

- Develop methods to define cellular networks and the molecular interactions and mechanisms of their regulation.
  - Comprehensive mapping of microbial regulatory processes, including
    - Develop the capability to construct detailed regulatory maps for specific subgenomic networks positioned across multiple species.

    - » Build comprehensive regulatory-circuitry maps.
    - » Connect regulatory properties (including operons and regulons) and their repressors and inducers with cellular functions and phenotypes.
  - – Elucidation and correlation of links between intracellular regulatory processes and extracellular cues from the environment and from microbial-community members.
  - – Support for a theoretical framework based on evolutionary biology and associated set of computational modeling tools to predict the dynamic behavior of natural or designed regulatory mechanisms. This will provide a solid basis for understanding how regulatory mechanisms work so we can use them, modify them in useful ways, and design new ones.

3.2.2.2.2.2. Computation Needs

Computational capabilities must be developed for the following:

- Extraction of regulatory elements using sequence-level comparative genomics.
- Inference of regulatory processes and networks from microbe and community functional data.
- Simulation of regulatory networks using both nondynamic models of regulatory capabilities and dynamic models of regulatory kinetics.
- Prediction of modified or redesigned gene regulatory system behavior.
- Integration of regulatory-network, pathway, and expression data into integrated models of microbial function.

See section 3.3. Highlights of Research in Progress to Accomplish Milestones, p. 55.

3.2.2.3. Milestone 3: Develop the Knowledgebase, Computational Methods, and Capabilities to Advance Understanding and Prediction of Complex Biological Systems

3.2.2.3.1. Background and Strategy

GTL's central goal is to provide the technologies, computing infrastructure, and comprehensive knowledgebase to surmount the barrier of complexity that prevents the translation of genome sequence directly into predictive understanding of function. *Genome sequences furnish the blueprint, technologies can produce the data, and computing can relate enormous data sets to models of process and function.*

The ultimate goal of every science is to achieve such a complete understanding of a phenomenon that a set of mathematical laws or models can be developed to predict accurately all its relevant properties. Although such capabilities now exist for certain areas of physics, chemistry, and engineering, virtually no biological systems are understood at this level of detail and accuracy. Because theory and computation are limited by the lack of experimental data and the means to verify models quantitatively, their application has had relatively little impact on biology. With the developments described in this plan, the biosciences are poised for rapid progress toward becoming the quantitative and predictive science known as systems biology.

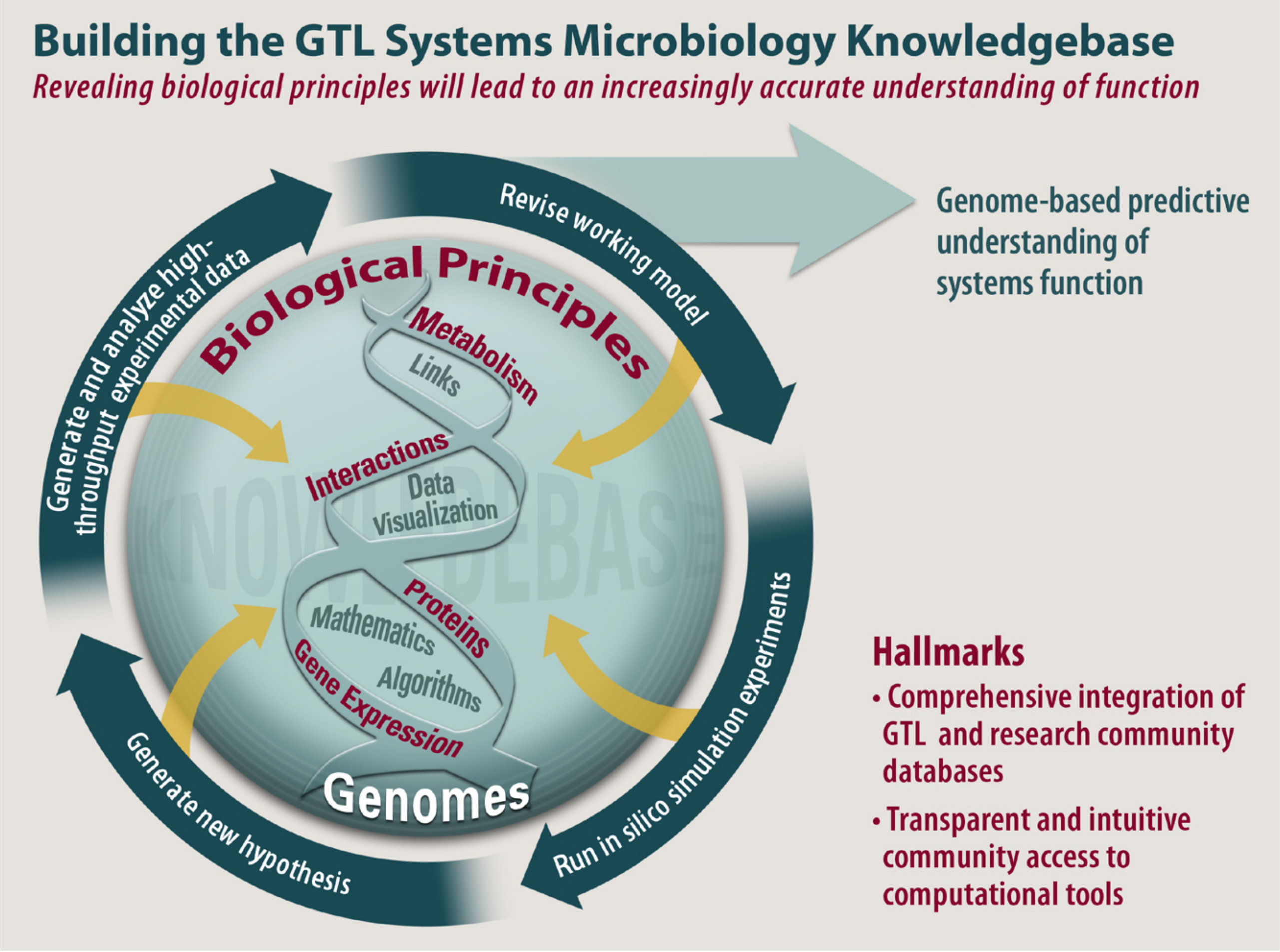
Models can form the foundation for understanding complex systems. They can be applied to such useful ends as developing biological sources of clean energy, cleaning up toxic wastes, and understanding the roles of microbial communities in ocean and terrestrial carbon cycling (i.e., how they sequester carbon and how the processes involved respond to and impact climate change). The key challenge to achieving GTL goals will be development of capabilities for modeling and simulation—capabilities that must be coupled tightly with experimental methods to identify and characterize biological components, their interactions, and the products of those interactions.

The program's computational component will require developments ranging from more-efficient modeling tools to fundamental breakthroughs in mathematics and computer science, as well as algorithms that efficiently use platforms ranging from workstations to the fastest available computers.

3.2.2.3.2. GTL Knowledgebase

Analysis of each new microbe benefits from insights gained through knowledge about all other microbes and life forms. To achieve our goal of understanding life's full complexity, we can take advantage of the unity of life and its evolutionary history brought about by the common hereditary molecule DNA and the underlying principles and instructions it encodes. Nature's simplifying principles make this a powerful strategy. Just as a finite number of rules determine the structure and function of proteins, so the higher-order functions of cells seem to emanate from another finite set of principles and interactions. Once successful machines, pathways, and networks arise by evolution, they tend to be preserved, subtly modified and optimized, and then reused as variations on enduring themes throughout many organisms and species. Thus, accumulating detailed information on numerous microbes across a wide range of functionality ultimately will provide the insight needed to interpret these principles. Comparative genomics, founded on these principles, will allow us to predict the functions of unknown microbes by deriving a working model of a cell from its genetic code.

Comparative genomics has been a powerful computational technique in the genome-sequencing era, yielding insights into gene function that provide discoveries and allowing prediction and hypothesis development. In this new era of systems biology, all-against-all comparisons of much more extensive microbial data amassed in the GTL Knowledgebase (see figure below) will accelerate and sharpen our research strategies. Foundation of the knowledgebase, the DNA sequence *code* will relate the many data sets emanating from microbial systems biology research and discovery. Over time, an intensely detailed description of each gene's function and regulatory elements will be created from which networks and subsystems—and eventually cellular and community structure and functionality—can be derived. As these capabilities improve, focus of the experimental



Building the GTL Systems Microbiology Knowledgebase

Revealing biological principles will lead to an increasingly accurate understanding of function

Hallmarks

- Comprehensive integration of GTL and research community databases
- Transparent and intuitive community access to computational tools

process and research resources can shift from the study of unknown components and functions to development of a new generation of capabilities for probing system functions by such methods as predicting, testing, and manipulating the role of microbes in ecosystems or designing systems for biofuel production. Along with high-throughput facilities and computing, this strategy is a key element of our approach to reducing a microbial system's analysis time from many years to months.

Given a knowledgebase with many genes from organisms highly annotated with functional data (cross-referenced to each other), much information about a newly sequenced genome will be at scientists' fingertips.

A History of Genomics at DOE

In 1986, DOE's Office of Health and Environmental Research, precursor to the Office of Biological and Environmental Research (BER), initiated the Human Genome Project (HGP), becoming the first federal agency to provide directed funding for the HGP. In 1994, BER launched the Microbial Genome Program, which was responsible for determining genomic sequences of most of the first microbes sequenced. One of these organisms was *Methanococcus jannaschii*, which established the existence of the Archaea domain; another was *Mycoplasma genitalium*, a free-living cellular organism with the smallest genome yet discovered. Continuing to support microbial genomic sequencing, DOE has completed more than 200 genomes of microbes relevant to DOE missions, demonstrating the metabolic importance of nonpathogenic microbes for processes like terrestrial nitrification, ocean carbon sequestration, and potentially for bioenergy production and bioremediation of contaminated sites. In 2000, BER started the Microbial Cell Program to take advantage of the growing wealth of microbial DNA sequence data for understanding microbes as complete biological systems. The following year, this program became Genomics:GTL.

Microbial genome sequencing is changing the way we do science. In the early 1990s, years were required to sequence a microbe, but now a high-throughput facility can sequence two microbial genomes in less than a day. These sequences provide the basis for comparative genomic analyses, including gene annotations, and are the foundation for GTL systems biology studies. Although the reductionist approach to studying isolated components of cells has been highly productive, scientists now can proceed reconstructively with complete microbial sequences. This complete (and finite) "parts list" for a microbe is the starting point in exploring the capabilities of microbes and how their molecular machines (protein complexes) are built and function in vast interconnected networks.

Joint Genome Institute

The sequences of DOE-relevant microbes have been provided largely by BER's Joint Genome Institute (JGI), an important resource that is producing microbial, microbial-community, and other genome sequences. Formed in 1997 to make DOE's contribution to human sequencing in the Human Genome Project, JGI devotes 40% of its capacity to sequencing organisms relevant to DOE missions. Current JGI capacity is more than 30 billion base pairs of raw sequence per year.

Additionally, JGI is a user facility that accepts annual proposals from the broader research community as part of its Community Sequencing Program (CSP); these projects use about 60% of JGI's sequencing capacity. CSP's primary goal is to provide a world-class sequencing resource for the expanding diversity of disciplines—geology, oceanography, and ecology, among others—that can benefit from the application of genomics. JGI recently brought online its new clearinghouse web site—Integrated Microbial Genomes (IMG, http://img.jgi.doe.gov/ v1.1/main.cgi/). IMG's aim is to help researchers analyze the deluge of DNA data on microorganisms. Nearly 300 draft or completed genome sequences are now available from archaea, bacteria, and other microbes, along with tools for sifting through the data. Included is basic information about genes, proteins, and their functions. Diagrams illustrate which biochemical pathway is influenced by a given gene, and browsing tools can be used to pinpoint similar genes in different organisms and compare them side by side. See the JGI web site for more information and sequence data (www.jgi.doe.gov).

Researchers will use the knowledgebase with computation and modeling to drive hypothesis formulation, experiment design, and data collection. The interoperable, open-access knowledgebase will enable quick deduction of any gene's function or complex biochemical pathways of interest. Insights gained in these studies will help transform biology into a more quantitative and predictive science based on models that synthesize observations, theory, and experimental results. This paradigm combines both discovery and computationally driven hypothesis science as we navigate massive data sets to reveal unforeseen properties and phenomena and derive insights from previously unfathomable complexity.

Building over time to an intensely detailed and annotated description of microbial functional capabilities, the GTL Knowledgebase will assimilate a vast range of microbial data as they are generated. The knowledgebase will grow to encompass program and facility data and information, metadata, experimental simulation results, and links to relevant external data. It also will incorporate existing microbial data, including model microbial systems such as *Escherichia coli* and *Bacilus subtilis,* to take advantage of extensive understanding. The power of conservation in biology will be used to leverage and extend our partial knowledge about a few organisms to a more complete understanding of many microbes and their communities. Underlying the GTL Knowledgebase will be an array of databases, bioinformatics and analysis tools, modeling programs, and other transparent resources.

3.2.2.3.3. Elements of the GTL Integrated Computational Environment for Biology

To support the achievement of GTL science and missions goals, a number of essential elements of the computational environment will be established. They will include a seamless set of foundational experimentation capabilities to support the pinnacle capability of theory, modeling, and simulation. Especially needed is a rigorous and transparent system for tracking, capturing, and analyzing data within a computing and information infrastructure accessible to the scientific community as end users of the data. The enabling environment for GTL computation consists of six complementary technical components, each with its own supporting roadmap:

- Theory, Modeling, and Simulation
- LIMS and Workflow Management
- Data Capture and Archiving
- Data Analysis and Reduction
- Computing and Information Infrastructure
- Community Access to Data and Resources

Development of these necessary capabilities is discussed in detail in 4.0. Creating an Integrated Computational Environment for Biology, p. 81. See section 3.3. Highlights of Research in Progress to Accomplish Milestones, p. 55.

3.2.2.3.4. Synergisms with Other Agencies and Industries

GTL will leverage information and methods from a variety of sources, including

- Protein structures produced in the Protein Structure Initiative of the National Institutes of Health (NIH)
- Protein Data Bank
- Databases of metabolic processes such as KEGG and WIT
- Hosts of available analytical tools in such areas as molecular dynamics, mass spectrometry (MS), and pathway modeling and simulation
- NIH National Center for Biotechnology Information data and tools
- Industrial vendors and tool developers

Successful production of new technologies and advanced tools for computational biology will require the sustained efforts of multidisciplinary teams, teraflop-scale and faster computers, and considerable user expertise.

Encompassing the entire biological community, this task will involve many institutions and federal agencies, led in many aspects by NIH and the National Science Foundation. A central component of GTL will be the establishment of effective partnerships with these and other agencies and with commercial entities to ensure the widespread adoption of computational tools and standards and to eliminate redundant work.

3.2.2.4. Milestone 4: Design and Build User Facilities to Accelerate GTL Microbial Systems Biology

3.2.2.4.1. Background and Strategy

The need for facilities is driven by the scope and scale of DOE mission challenges; the demands of systems biology for large-scale comprehensive analyses; and the scientific complexity and diversity of microbes, microbial communities, and ecosystems. Genomics provides an inherent systems perspective to biology, but to achieve the full promise of the genome revolution, we need a concomitant change in the fundamental practice of biological research. Investments in the proposed facilities will make the necessary change technically and financially tractable, provide a new engine for discovery and experimentation, and allow us to achieve timely mission impacts. Through rigorous application of high-throughput methods in a consolidated production environment and computational and information technologies, new levels of performance can be attained, productivity vastly increased, and costs greatly reduced to accomplish the type of comprehensive analyses we envision.

The facilities are the core of the DOE strategy to achieve a new third model for biology, bridging the gap between big science and small science (see Introduction, section 1.3.4, p. 9) and providing the most advanced capabilities and information to all scientists on an equitable basis. Democratization of the development and application of advanced technologies and computing has the benefit of engaging a much larger community of scientists and their skills and resources in solving national energy, environmental, and climate problems.

The four facilities are:

- Facility for Protein Production and Characterization of Proteins and Molecular Tags
- Facility for the Characterization and Imaging of Molecular Machines
- Facility for Whole Proteome Analysis
- Facility for Modeling and Analysis of Cellular Systems

The discussion of these facilities, their rationale, functions, technology challenges, and development plans begins with 5.0. Facilities Overview, p. 101.

3.2.2.4.2. Computing Needs

The GTL facilities concept is based on the integration of hundreds of high-throughput technologies to achieve new levels of performance, productivity, quality, and cost. Achieving these ends requires a rigorous computationally based system of planning, monitoring, control, and output management and dissemination. Without a comprehensive and fully integrated computing and information system, the facilities concept is not viable. The needs for this system and a roadmap for its development are discussed in the computing chapter (4.0), beginning on p. 81.

3.3. Highlights of Research in Progress to Accomplish Milestones

Highlights of research progress toward Milestones 1–3 are listed below, with references to selected sidebars that follow this section. Progress is foundational for establishing the facilities described in Milestone 4. For a comprehensive list of funded projects, see Appendix E. GTL-Funded Projects, p. 245.

3.3.1. Research Highlights for Milestone 1: Sequences, Proteins, Molecular Complexes

- Microbial genome sequencing at DOE has produced the sequences of over 200 microbes (see Appendix G. Microbial Genomes Sequenced or in Process by DOE, p. 253, and sidebar, A History of Genomics at DOE, p. 53).
- Microbial communities are being sequenced from environments as diverse as acid mine drainage sites (where the pH is less than 1.0) and the Sargasso Sea (e.g., see sidebar, Metagenomics: Opening a New Window onto Natural Microbial Communities, p. 62).
- Single-cell methods are being developed to sequence individual organisms from complex communities, and a number of laboratory culture-independent approaches are being used to investigate the composition and functionality of microbial communities.
- Improved methods for synthesizing genomes are being developed to test our understanding of gene function and regulation (see sidebar, Accurate, Low-Cost Gene Synthesis from Programmable DNA Microchips, p. 75; Smith et al. 2003).
- Several projects are developing new concepts and strategies to generate recalcitrant proteins such as those in membranes and those containing metals.
- Others are piloting high-throughput methods for turning out proteins needed now by GTL projects. The ultimate goal is creating the capabilities to produce on demand any protein potentially expressed by microbes and microbial communities (see 5.1. Facility for Production and Characterization of Proteins and Molecular Tags, p. 111).
- Some GTL projects are developing methods for creating new classes of affinity reagents for high-throughput global assays (e.g., on chips) of protein-expression and interaction partners to identify, track, remove, and disable corresponding proteins; locate protein complexes in living systems; and for other purposes (see Molecular Tags: Fusion Tags and Affinity Reagents, p. 126).
- Technologies are being refined, validated, and deployed in increasingly automated pilot pipelines to cultivate, isolate, stabilize, and characterize molecular complexes by making use of miniaturization and other developments, focusing on a number of microbes including *Rhodopseudomonas* and *Shewanella* (e.g., see 5.1. Facility for Production and Characterization of Proteins and Molecular Tags, p. 111, and sidebar, Capturing and Characterizing Protein Complexes, the Workhorses of the Cell, p. 68).
- Live-cell imaging, including colocalization and FRET-based techniques, are being used to observe complexes.
- Capabilities for modeling molecular-machine shapes and reaction surfaces are being developed and tested.

3.3.2. Research Highlights for Milestone 2: Cell and Community Function, Regulation, and Dynamics

- Platforms deploying advanced high-throughput separations and spectrometric instrumentation coupled with appropriate computational infrastructure are being developed by multiple projects for global proteomic analyses to identify and quantify large sets of proteins more comprehensively, including quantitative modalities

nature
biotechnology
VOLUME 22 NUMBER 1 JANUARY 2004
www.nature.com/naturebiotechnology
Rhodopseudomonas palustris genome sequence
Chemical genomics reveals drug pathways
Quantum dots in cancer biopsy

Metabolically Versatile Microorganism. Characteristic reddish colonies of the purple photosynthetic bacterium *Rhodopseudomonas palustris* are superimposed on images of this organism's rod-shaped cells visualized under the light microscope. The complete sequence of *R. palustris* was determined by the DOE Joint Genome Institute and reported in Larimer et al., *Nat. Biotechnol.* 22(1), 55–61 (2004).

Cover and caption used by permission from *Nature Biotechnology*, www.nature.com/nbt/

for analyzing small samples (e.g., see sidebar, Measuring Differential Expression of Cytochromes in the Metal-Reducing Bacterium *Geobacter,* p. 65).

- Some projects are focusing on uncultured microbes and microbial communities and the use of microfluidics and miniaturization with the goal of eventually reducing sample sizes to single or a few microbes.
- Functional imaging technologies are being improved to study the biochemistry of key microbial functions at the cellular and subcellular levels.
- Novel analysis approaches are being undertaken by several projects that are assessing the metabolome as a means of analyzing gene function.
- Projects investigating biological mechanisms having potential for alternative fuel production include investigations of the role of cellulose-binding modules in cellulolytic activity and large-scale analysis of the genes and metabolic pathways involved in photolytic hydrogen production.
- *R. palustris*, a common soil and water bacterium, is one of the most metabolically versatile organisms because it can make its living by converting sunlight into cellular energy, producing hydrogen as it degrades and recycles cellulose and lignins, and living off other substrates. Research goals are to use metabolic modeling to help optimize carbon sequestration and hydrogen evolution. One team of scientists has taken global approaches to ascertain mechanisms of metabolic regulation of carbon dioxide, hydrogen, nitrogen, aromatic acid, sulfur pathways, and other processes. A coordinated application of gene-expression profiling, proteomics, carbon-flux analysis, and computing approaches has been combined with more traditional studies of mutation analysis and cellular characterizations.
- Research on gene regulation in *Caulobacter crescentus* is focusing on the importance of master regulators (see sidebar, Genetic Regulation in Bacteria, p. 67).
- A goal for studies of environmental microbial systems biology is facile viewing of life processes—in real time. The molecules of life's complex choreography must be observed as its components carry out their specified activities inside and among cells interacting in dynamic microbial communities. A number of more in-depth systems biology projects are being undertaken on four organisms. These projects focus on integrating the results of the analyses of cellular proteomes, biochemistry, and imaging; they also model pathways.
  - The first two projects study cyanobacteria at the foundation of ocean food chains responsible for about half the photosynthesis (CO_2 fixation) on earth.
    - » Accomplishments of *Synechococcus* research include the *Synechococcus* Encyclopedia, which provides integrated access to genomics and proteomics databases to aid studies into the behavior of these abundant marine organisms important to global carbon fixation (see sidebar, *Synechococcus* Encyclopedia, p. 69). Other research efforts have given insight into the specificity of RuBisCO, an enzyme central to photosynthetic carbon fixation (see sidebar, New Imaging and Computational Tools Enable Investigations of Carbon Cycling in Marine Cyanobacteria, p. 66).
    - » Accomplishments regarding *Prochlorococcus* include explorations into gene expression in day-night cycles of this photosynthetic organism and gene transfer between it and phages (see sidebars, Modeling the Light-Regulated Metabolic Network of *Prochlorococcus marinus*, p. 64; and Transfer of Photosynthetic Genes Between Bacteria and Phages, p. 64).
  - The second two projects study organisms with capabilities to remove or detoxify metals from contaminated environments. Remediation projects are making an array of microbial-system measurements, including gene expression and qualitative and quantitative proteomics with modeling and simulation experiments on metal-reducing bacteria. The aim is to understand gene and operon regulation under natural environmental conditions that may affect the outcomes of metal-reduction and immobilization processes mediated by bacteria including *Geobacter sulfurreducens* and *S. oneidensis*:
    - » Some accomplishments of the *Shewanella* Federation include identifying global stress-response patterns to radiation, nitrate, and oxygen; identifying gene-expression patterns that are electron-

acceptor specific; determining the role of selected global regulators in anaerobic respiration; demonstrating expression of hypothetical genes and enhanced genome annotation; and elucidating mechanisms of electron transfer to metals and metal oxides (see sidebar, The *Shewanella* Federation, starting on p. 70).

- » Some accomplishments of *Geobacter* research include creating in silico models of *Geobacter* to predict responses to environmental conditions and aid in optimizing bioremediation and energy harvesting; demonstrating that *Geobacter* can generate electricity from a wide variety of organic wastes and renewable biomass; and determining that growth and activity of metal-reducing organisms in natural environments are enhanced by feeding microbes carbon sources such as acetate (see sidebars, *Geobacter*, p. 74; Harvesting Electricity from Aquatic Sediments with Microbial Fuel Cells, p. 76; and BER Research Advancing the Science of Bioremediation, p. 219).

3.3.3. Research Highlights for Milestone 3: Computing

- Progress is being made in data reduction and analysis for MS experiments, integration of databases containing heterogeneous data sets, and use of myriad approaches to metabolic and regulatory network modeling.
- The computational framework for comparative analysis of functional genomic data and computational models is being developed for data on the behavior of microbial gene regulatory networks in response to environmental conditions.
- Whole-cell flux-balance models are being used to understand aspects of natural behavior and for comparative analysis of different microbial strains.
- Computational methods are being developed to predict the wiring diagrams of various response networks, which consist of signaling, regulatory, and metabolic components. These include carbon fixation, phosphorus assimilation, and nitrogen-assimilation networks encoded in cyanobacterial genomes. Research is ongoing to apply the framework to *Shewanella*. These methods use predictions of operons and regulons and interaction relationships among candidate genes whose proteins appear to be expressed together or coordinately (see sidebar, *Synechococcus* Encyclopedia, p. 69).
- Computational models are being built to predict the activity of natural microbial communities for application of robust bioremediation technologies. Teams also are learning how to simulate growth and activity of metal-reducing organisms in their natural environments.
- Three institutes have been created to support the advancement of computational-biology research as an intellectual pursuit and provide innovative approaches to educating biologists as computational scientists. Using interdisciplinary teams of researchers drawn from the physical and life sciences, computational mathematics, and computer science, the institutes sponsor multidisciplinary scientific projects in which biological understanding is guided by computational modeling. They are training students to uncover biological mechanisms and pathways within microbial organisms through the use of computational biology and synergistic collaborations with experimental groups and will engage students in project-oriented research (www.doegenomestolife.org/compbioinstitutes/).

Chapter text continues on p. 77.

3.3.4. Sidebars Illustrating Details of Specific Research

The following section highlights progress in some GTL-supported projects.

GTL Progress: Rapid Deduction of Stress-Response Pathways in Metal- and Radionuclide-Reducing Bacteria

GTL researchers at the Virtual Institute of Microbial Stress and Survival (VIMMS, vimss.lbl.gov) at Lawrence Berkeley National Laboratory (LBNL) are studying how environmentally important microbes adapt and evolve at DOE-managed contamination sites and how their biogeochemical processes may be exploited to remediate these sites. Apart from the basic knowledge to be gained about microbial adaptability and evolvability, the ultimate goal is to integrate the findings into computational models of organism response to environments and each other. In enough detail, such models can help develop potential uses for these processes in bioremediation.

A variety of microbes coexist at contaminated field sites. These include *Desulfovibrio vulgaris*, which belongs to a class of sulfate-reducing bacteria found ubiquitously in nature, sulfur reducers like *Geobacter*, and other microbes. Sulfate-reducing bacteria represent a unique class of organisms that nonphotosynthetically and anaerobically generate energy through electron transfer–coupled phosphorylation using sulfite as a terminator; they thus play a critical role in sulfate cycling. Sulfate reducers also play an important role in global recycling of numerous other elements, especially in anaerobic environments. Aside from their role in biocorrosion and oil-well souring problems in the petroleum industry, their potential to bioremediate many toxic metals is of interest.[1] These bacteria can reduce such metal contaminants as uranium and chromium from a soluble, high-oxidation state in which they are mobile and toxic to less soluble forms, thus preventing their entry into the groundwater and reducing the compounds' toxicity. By developing an understanding of the molecular processes that enable these cells to reduce metals, investigators hope to derive optimal protocols for naturally stimulating the bacteria to higher reduction efficiencies.

To better understand the molecular processes occurring in these microbes at contaminated sites, VIMSS researchers have developed a pipeline to simulate field conditions in the laboratory and to produce quality-controlled, reproducible biomass under different stress and metabolic conditions. Analyses range from synchrotron infrared microscopy for evaluating gross physiological cell functions and cellular imaging to functional genomic measurements of gene expression, protein expression, and metabolite production.

To organize and interrelate these data in a genomic context, VIMSS researchers developed MicrobesOnline[1] (www.microbesonline.org), a publicly available comparative genomics resource that facilitates cross-comparison of the genome architecture and functional genomics of these and other microorganisms. Underlying its functionality is a pipeline of genome-annotation tools, novel operon- and regulon-prediction algorithms,[2] multispecies genome and Gene Ontology browsers, a comparative KEGG metabolic pathway viewer, the Bioinformatics Workbench for in-depth sequence analysis, and Gene Carts that allow users to save genes of interest for further study. In addition, VIMSS provides an interface for community-driven genome annotation. All data developed by VIMSS and imported from other projects are referenced to their respective organisms and, if appropriate, to genomic regions (gene identifications or functional sites). This allows researchers to cross-compare data on stress responses in multiple organisms. One result of creating this site and its algorithms and centralizing microbial functional genomic data is that VIMSS has lent strong support for one theory and against another regarding the formation of operons[3] and the origins of strand bias.[4]

Sequencing of four sulfate-reducing bacteria and development of comparative genomics tools by VIMSS GTL researchers and others have enabled the determination of sulfate-reducing–bacteria genomic "signatures." These signatures comprise some 50 genes unique to sulfate reducers. VIMSS has supported these predictions, using gene-expression data developed at VIMSS, by demonstrating that not only are the known genes coregulated, but so are a significant fraction of the unknown genes. Gene-expression data demonstrated a clustering of signature-gene responses over a number of conditions.

Figure 1 shows a schematic of *D. vulgaris'* pathway response to nitrite stress. We hope to understand the adaptation of each pathway in various environments and to learn how differences in these responses might support community interactions. A strong link seems to exist among this organism's metabolic responses to iron, sulfate, and nitrite. This might be expected simply because iron-dependent proteins are present in the nitrogen-response pathways; there also is evidence, however, of cross talk through coupled oxidation of ferrous iron. In addition, through comparative genomics VIMMS has predicted a link through the HcpR regulon that may regulate genes dealing with iron, sulfate, and nitrogen oxides.[5] VIMSS is in the process of comparing and contrasting the responses of *D. vulgaris, Shewanella oneidensis,* and *Geobacter metallireducens* to cover these conditions.

The physiological pipeline at VIMSS has been used to characterize a number of different responses to culture conditions of *D. vulgaris* and *S. oneidensis* at various levels of detail. For example, transcriptomic studies of the latter have uncovered the heat shock pathway homologous to that well characterized in *E. coli* and have identified key differences in metabolism regulation and membrane composition.[6] Analysis of sodium chloride shock in both *S. oneidensis*[7] and *D. vulgaris* has uncovered strong commonalities and distinctions between responses of the two organisms (different arrays of transporters, antiporters, osmoprotectants, and metabolic adjustments).

The pattern of expression both within and across organisms is mediated by the wiring of respective signal-transduction pathways. These pathways determine when homologous pathways are turned off and on in different organisms. Thus, VIMSS has focused on creating mutants in these components and on understanding the system's evolution in our target organisms. Two-component systems are the major signal-transduction pathway of bacteria. These systems are composed of a histidine kinase protein activated by an environmental signal and a response regulator that is affected by the histidine kinase and actuates a response. The response regulators may change the expression of genes or activate motility or perform other functions. A single histidine kinase might regulate a number of response regulators.

One interesting finding is that histidine kinases seem to undergo rapid lineage-specific family expansions. These expansions are particularly large and rapid in the target environmental microbes. Figure 2 shows a phylogenetic

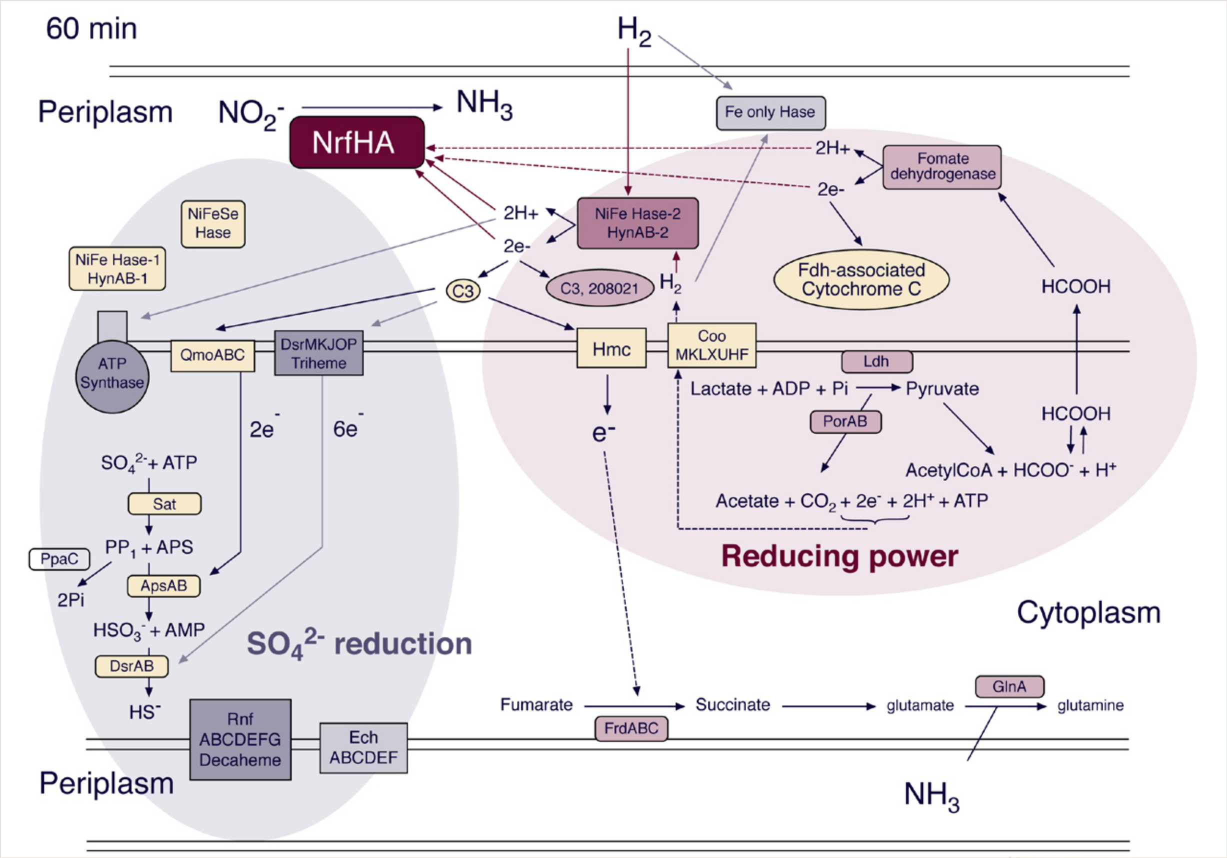


Fig. 1. Schematic Depiction of the Nitrate Response of *D. vulgaris*. Red systems are upregulated and blue systems downregulated.

tree of one cluster of proteins annotated as histidine kinases across more than 200 prokaryotic genomes. Examples of lineage-specific expansions are shown with colored ovals. Such expansions imply that these organisms evolve very rapidly and tune their signal-transduction systems to the precise environments in which they live and to the range of perturbation they encounter therein. [Adam Arkin, LBNL]

References

1. E. J. Alm et al., "The MicrobesOnline Website for Comparative Genomics," *Genome Res.* **15**, 1015–22 (2005).

2. M. N. Price et al., "A Novel Method for Accurate Operon Predictions in All Sequenced Prokaryotes," *Nucleic Acids Res.* **33**, 880–92 (2005).

3. M. N. Price et al., "Operons Formation is Driven by Coregulation, Not by Horizontal Gene Transfer," *Genome Res.* **15**, 809–819 (2005).

4. M. N. Price, E. Alm, and A. Arkin, "Interruptions in Gene Expression Drive Highly Expressed Operons to the Leading Strand of DNA Replication," *Nucleic Acids Res.* **33**(10), 3224–34 (2005).

5. D. A. Rodionov et al., "Reconstruction of Regulatory and Metabolic Pathways in Metal-Reducing Delta-Proteobacteria," *Genome Biol.* **5**, R90 (2004).

6. H. Gao et al., "Global Transcriptome Analysis of the Heat Shock Response of *Shewanella oneidensis*," *J. Bacteriol.* **186**, 7796–7803 (2004).

7. Y. Liu et al., "Transcriptome Analysis of *Shewanella oneidensis* MR-1 in Response to Elevated Salt Conditions," *J. Bacteriol.* **187**, 2501–7 (2005).

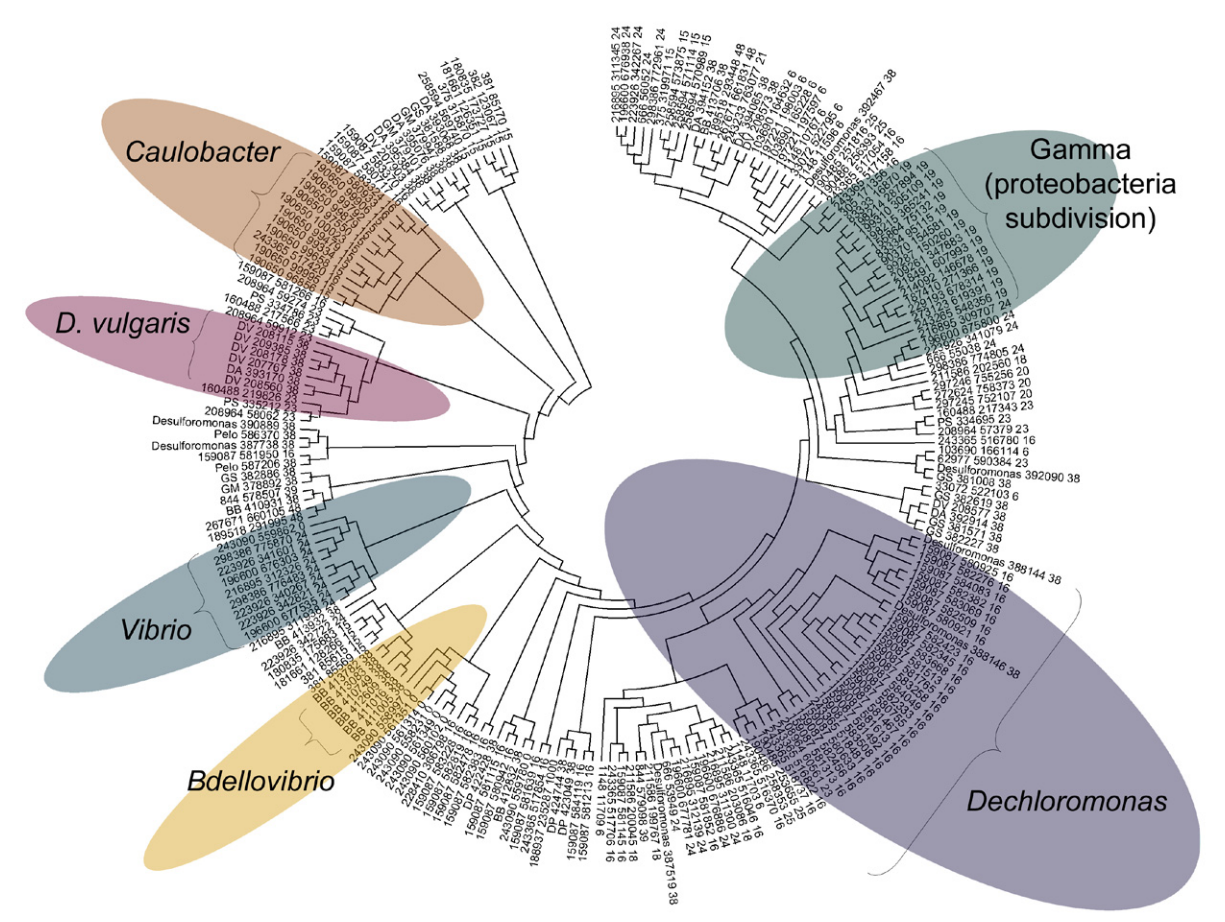


Fig. 2. Part of the Phylogenetic Tree of Bacteria Histidine Protein Kinases.

Metagenomics: Opening a New Window onto Natural Microbial Communities

Our understanding of microbial diversity and function has been limited severely by the inability to grow the vast majority of microbes in the laboratory. High-throughput DNA sequencing and other biotechnology tools now offer a new avenue for obtaining this knowledge. Genome fragments can be isolated and analyzed after being collected directly from environmental samples, whether liters of water from the open sea or a scraping from a slick film at the bottom of a highly acidic mine. Such environmental genomic (metagenomic) approaches have been applied to studying entire microbial communities in a specific locale as well as single genes, pathways, and whole organisms. Analyses of these data have revealed a broad spectrum of genomes, genes, and previously undiscovered functions.[1]

These studies will result in a multitude of new insights into the dynamics between microbes and their environments and will have the potential to catalyze development of numerous practical applications. Effective mining of the environment for fundamental knowledge and products, however, will require substantial investments in new high-throughput technologies. These biophysical and physiological techniques can help reveal the functions of new microbial proteins and compare the properties of large collections of genes of a particular type or function.[1, 2]

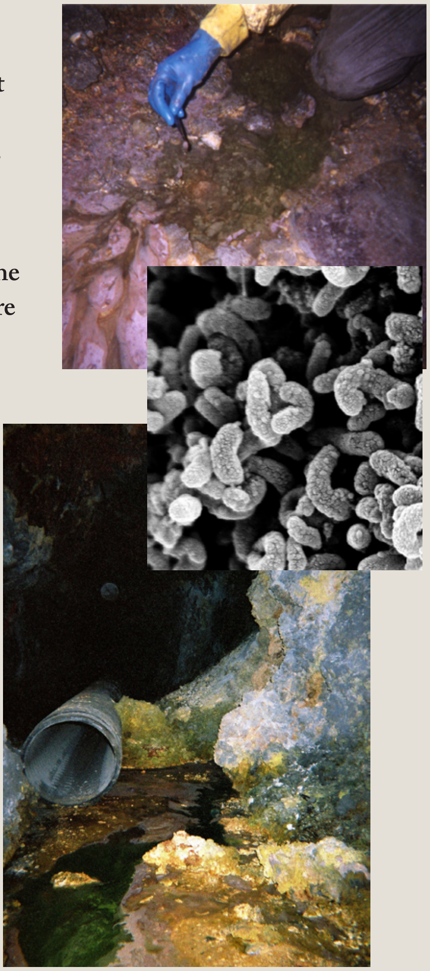
GTL supported the first sequencing of a microbial community directly from the environment at Iron Mountain, California,[3] and the first comprehensive study of gene expression in such a community,[4] as well as the vast environmental sample data set of large DNA fragments collected from the Sargasso Sea.[5] Some highlights of these studies are described below, and all data are available to the research community.[6] GTL also supports 11 other studies of natural microbial communities from sites as diverse as a boiling thermal pool in Yellowstone National Park, former uranium mining sites, and complex soil environments.

Microbial Community Thriving in Acid Mine Drainage

Direct environmental sampling led to the characterization of members of a microbial community in highly acidic water from an abandoned gold mine at Iron Mountain, one of the nation's worst Superfund sites (see bottom right). Acid mine drainage is caused by the complex interaction of various microbes with exposed iron ore and water, resulting in a mix so toxic (pH 0.83) that it can completely corrode shovels accidentally left overnight.

Samples were taken from a pink microbial biofilm (upper right) growing on the surface of acid mine drainage hundreds of feet underground within a pyrite ore body. A scanning electron microscope image of a piece of the biofim (middle) revealed a tight association of microbial cells. After extracting and cloning DNA from the biofilm, investigators were able to reconstruct the genomes of two hardy microbes and parts of three others capable of withstanding the harsh conditions. Four of the microbes had never been cultivated. Using genomic and mass spectrometry-based proteomic methods, the team later identified over 2000 proteins from the 5 most abundant species, including 48% of the predicted proteins from the dominant biofilm organism. One of the proteins (a cytochrome) from a minor organism is key in the production of acid mine drainage. More than 500 of the proteins seem to be unique to the biofilm bacteria.

Further analyses of these data and future studies on each of the species will provide insights into their metabolic pathways, the ecological roles they play, and how they survive in such an extreme environment. Obtaining this knowledge can help in developing future cleanup strategies. [Jillian Banfield, University of California, Berkeley]



J. Banfield, University of California, Berkeley (three images)

Snapshot of the Complex Microbial Communities in the Sargasso Sea

J. Craig Venter Institute

Environmental investigations in the nutrient-poor waters near Bermuda in the Sargasso Sea (see photo at right) led to the discovery of 1800 new species of bacteria and more than 1.2 million new genes. Scientists used a whole-genome shotgun sequencing technique to clone random DNA fragments from the many microbes present in the sample. The resulting data represent the largest genomic data set for any community on earth and offer a first glimpse into the broad ensemble of adaptations underlying diversity in the oceans. Because microbes generally are not preserved in the fossil record, genomic studies provide the key to understanding how their biochemical pathways evolved.

Hundreds of the new genes have similarities to the known genes called rhodopsins that capture light energy from the sun. Bacterial rhodopsins couple light-energy harvesting with carbon cycling in the ocean through nonchlorophyll-based pathways.[7] Future studies will allow more insights into how these molecules function as well as opportunities for mining and screening the data for specific applications. The vast data set provides a foundation for many new studies by other researchers. Analyses using iron-sulfur proteins as benchmarks led one researcher, for example, to conclude that these data reflect diversity equal to that in all the currently available databases, suggesting that microbial diversity thus far has been vastly underestimated.[8] [J. Craig Venter Institute]

References

1. C. S. Riesenfeld, P. D. Schloss, and J. Handelsman, "Metagenomics: Genomic Analysis of Microbial Communities," *Annu. Rev. Genet.* 38, 525–52 (2004).

2. P. G. Falkowski and C. deVargas, "Shotgun Sequencing in the Sea: A Blast from the Past?" *Science* **304**, 58–60 (2004).

3. G. W. Tyson et al., "Community Structure and Metabolism Through Reconstruction of Microbial Genomes from the Environment," *Nature* 428, 37–43 (2004).

4. R. J. Ram et al., "Community Proteomics of a Natural Microbial Biofilm," *Science* 308, 1915–20 (2005).

5. J. C. Venter et al., "Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science* **304**, 58–60 (2004).

6. Whole-genome shotgun sequencing project data from Iron Mountain and the Sargasso Sea available on the web (www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/sargasso.html).

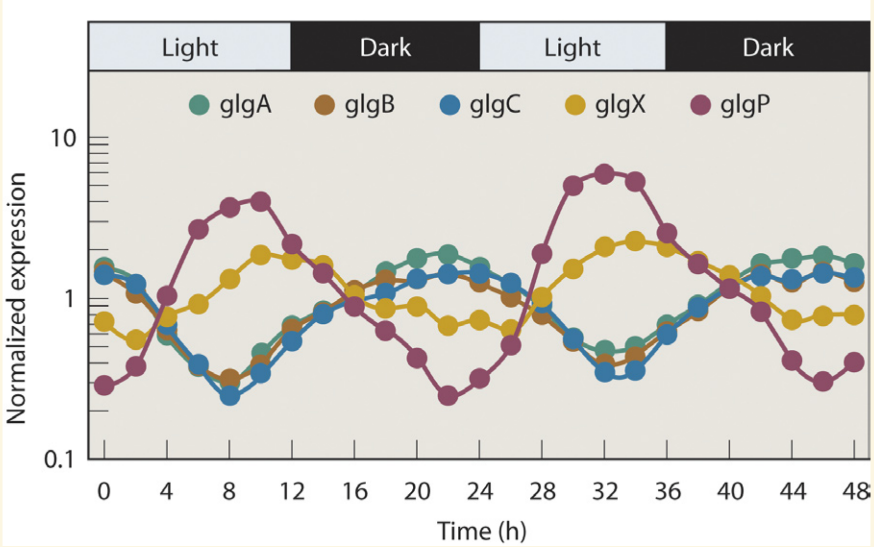
7. J. Meyer, "Miraculous Catch of Iron-Sulfur Protein Sequences in the Sargasso Sea," *FEBS Lett.* **570**, 1–6 (2004).

8. O. Beja et al., "Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea," *Science* 289, 1902–6 (2000).

Modeling the Light-Regulated Metabolic Network of *Prochlorococcus marinus*

The marine cyanobacterium *Prochlorococcus marinus* dominates the phytoplankton in the tropical and subtropical oceans and contributes to a significant fraction of global photosynthesis. To begin understanding metabolism at a systems level, GTL researchers are exploring day-night cycles known to play a central role in this bacterium's metabolism.

The graph summarizes the activities of five *Prochlorococcus* genes grown on a light-dark cycle and involved in a single metabolic pathway for central carbon metabolism. Data currently are being analyzed. Note that genes achieve maximal expression at different times, and some cycle less than others. The whole-flux balance model being developed for *Prochlorococcus* will be useful for generating hypotheses about the natural behavior and different strains of this important ocean organism. [George Church, Harvard University, and Penny Chisholm, Massachusetts Institute of Technology]



E. R. Zinser, Massachusetts Institute of Technology

Transfer of Photosynthetic Genes Between Bacteria and Phages

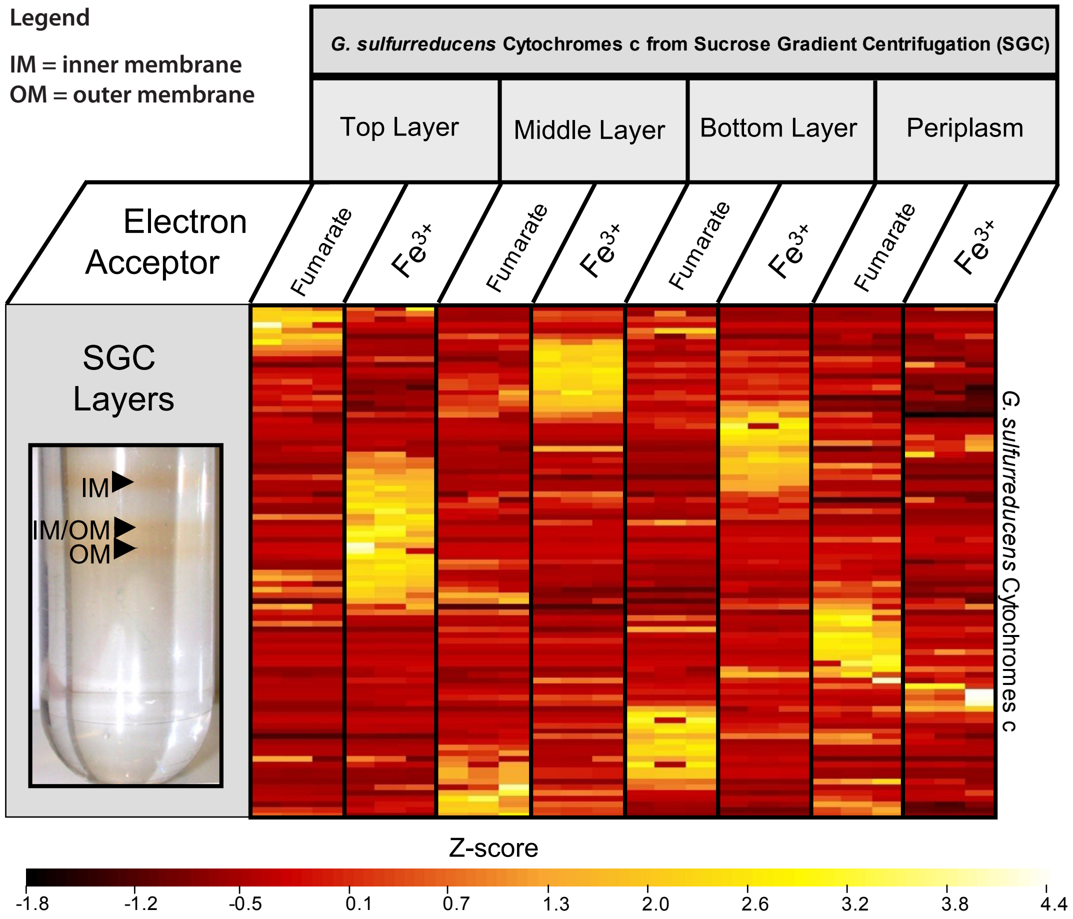
Viruses (phages) infecting the oceanic cyanobacteria *Prochlorococcus* are thought to mediate population sizes and affect the evolutionary paths of their hosts. GTL researchers analyzed genomes from three *Prochlorococcus* phages: a podovirus and two myoviruses. They appear to be variations of two well-known phages (T4 and T7) but also contain genes common to cyanobacteria that may help maintain host photosynthetic activity during infection by phages. Transferring these genes back to their hosts after a period of evolution in the phage could impact the evolution of both phages and hosts in the surface oceans. Phages in other environments also have been found to carry genes required by their hosts. Researchers hypothesize that these processes may represent a general phenomenon of metabolic facilitation of key host processes that could lead to specialization and possibly speciation. [Penny Chisholm, Massachusetts Institute of Technology]

Reference

D. Lindell et al., "Transfer of Photosynthesis Genes to and from *Prochlorococcus* Viruses," *Proc. Natl. Acad. Sci. USA* **101**, 11013–18 (2004).

Measuring Differential Expression of Cytochromes in the Metal-Reducing Bacterium *Geobacter*

Microbial proteomics involves the comprehensive measurement of cellular proteins to achieve a fundamental understanding of cell processes. New and innovative separation and mass spectrometry technologies enable cellular proteins to be identified and their cellular location and relative abundance to be determined. Cytochromes are an important class of proteins involved in dissimilatory metal reduction in the membranes of the bacterium *Geobacter sulfurreducens.* As part of Genomics:GTL, cytochrome distribution is providing important insights into how this organism responds and adjusts its electron-transport system to different environmental stimuli. Global measurements led to the determination that the relative abundance of certain c-type cytochromes varied markedly during growth on Fe(III), indicating that they would play an essential role in Fe(III) reduction. Such global measurements are important for simultaneously characterizing microbial proteomes and for achieving a systems-level understanding of how microorganisms can be manipulated to achieve desired outcomes for bioremediation, energy production, or carbon sequestration. [Mary Lipton, Pacific Northwest National Laboratory, and Derek Lovley, University of Massachusetts]



Relative Abundance Data Plot for 91 Cytochromes with One or More Unique Peptides Identified per Open Reading Frame from Cell-Fraction Preparations of *G. sulfurreducens*. The color represents the relationship of protein abundance to the average seen over all conditions. Darker colors represent lower abundance, and lighter colors represent increased abundance.

New Imaging and Computational Tools Enable Investigations of Carbon Cycling in Marine Cyanobacteria

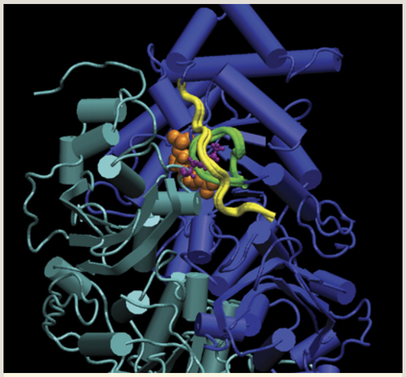
GTL research teams led by Sandia National Laboratories and Oak Ridge National Laboratory are developing new experimental and computational tools to investigate carbon-sequestration behavior in marine cyanobacteria, in particular, *Synechococcus* and *Synechocystis*. These abundant marine microbes are known to play an important role in the global carbon cycle.

Whole-cell imaging using a newly developed 3D hyperspectral microscope enabled researchers to detect the distribution of photosynthetic pigments in individual *Synechocystis* cells. The GTL team also improved the quality and information content in DNA microarray technology by combining hyperspectral imaging technology and patented multivariate statistical analysis.

The new system collects a full fluorescence emission spectrum at each pixel, as compared to the single bands of a spectrum collected by current scanners. All relevant wavelengths of light thus are measured at each point across a surface rather than simply at predefined bands of wavelengths. This approach enables the identification, modeling, and correction of gene expressions for unknown and unanticipated emissions; increases throughput by accommodating many spectrally overlapped labels in a single scan; and improves sensitivity, accuracy, dynamic range, and reliability. The scanner is being modified to allow 3D imaging of many fluorescently tagged molecules in cells and tissues.

New massively parallel modeling and simulation tools also developed by the team have yielded structural insight into the specificity of RuBisCO, an enzyme central to photosynthetic carbon fixation. The team also developed the computational capability to track spatial and temporal variations in protein species concentrations in realistic cellular geometries for important cyanobacterial subcellular processes. These tools include the Large-Scale Atomic/Molecular Massively Parallel Simulator (LAMMPS, www.cs.sandia.gov/~sjplimp/lammps.html), a molecular simulation tool; and ChemCell, a whole-cell modeling tool that captures those and other results into a spatially realistic metabolic pathway simulation.

LAAMPS enables investigations of the protein-sequence effect on different RuBisCO specificities and reaction rates in various species. Using this tool, researchers discovered that mutations in RuBisCO's amino acid sequence substantially altered the free-energy barrier for gating the binding pocket. This result provided a molecular-level explanation for the experimentally observed species variations in RuBisCO performance (see illustration). LAMMPS was released as open-source software in September 2004 and has been downloaded over 4000 times to June 2005. Via 3D simulations of diffusion and reaction in realistic geometries, ChemCell captured the carbon-fixation process carried out by RuBisCO in the carboxysome, a subcellular organelle. [Grant Heffelfinger, Sandia National Laboratories]



RuBisCO Carbon-Fixation Enzyme. RuBisCO has an active site (binding pocket) that binds ribulose-1,5-bisphosphate (RuBP) and catalyzes the reaction between RuBP and CO_2 or O_2. In the figure, the two large RuBisCO subunits (blue and cyan) sandwich an RuBP molecule (orange) in the active site. The site is gated by the C-terminus (yellow), lysine 128 (purple), and loop 6 (green), which undergo periodic conformational changes that open or close the site. Reactants enter and products escape while it is in an open state, and carbon-fixation reactions occur during the closed state. Simulations of this gating mechanism allow predictions of the gating rate, which can be linked to RuBisCO performance characteristics.

Reference

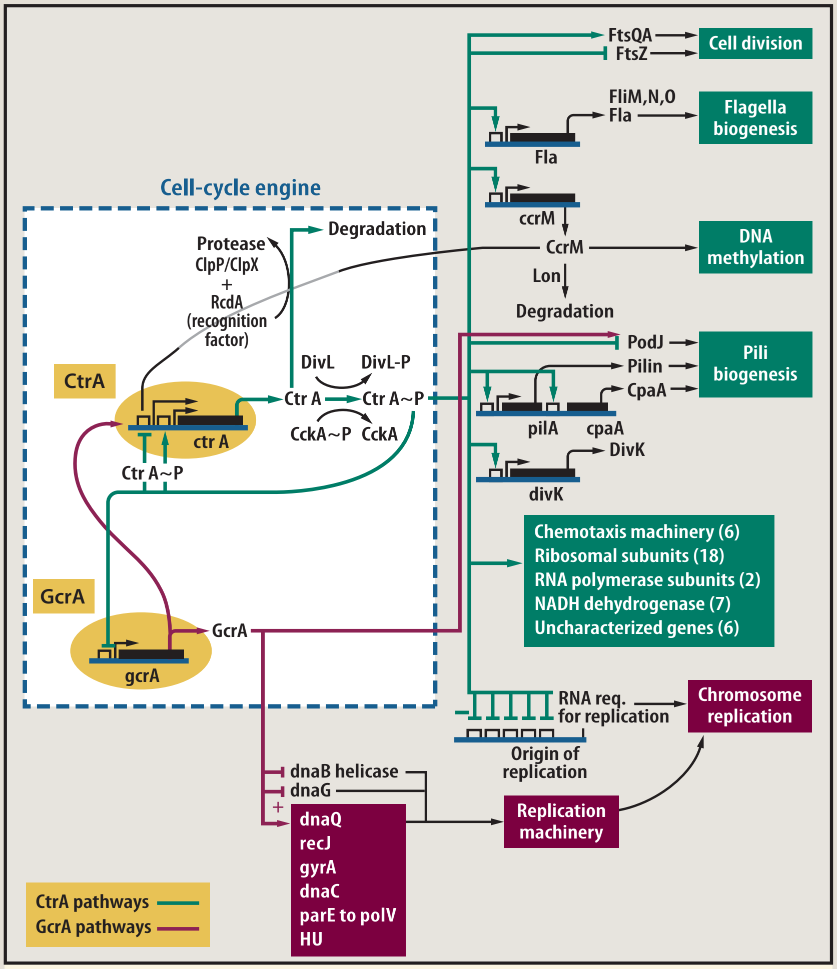
M. B. Sinclair et al., "Design, Construction, Characterization, and Application of a Hyperspectral Microarray Scanner," *Appl. Optics* 43, 2079–88 (2004).

Genetic Regulation in Bacteria

Progression through the cell cycle requires precise coordination of DNA replication, chromosome segregation, cell division, and cell growth. Study of the aquatic bacterium *Caulobacter crescentus* has shown that a small number of "master regulator" genes and their proteins provide this control. These proteins (CtrA and GcrA on the left side of the figure) interact with each other to form top-level regulatory circuitry that produces both temporal and spatial oscillations in their intracellular concentrations.[1] Changing concentrations of these regulatory proteins activate or repress key genes to initiate modular functions that implement the cell cycle through such activities as chromosome replication, cytokinesis, and the timing of construction and destruction of polar organelles.

The general features of the top-level genetic circuits comprising the cell's control system are emerging. The control system is hierarchical, modular, and asynchronous. Genes are expressed "just in time"—that is, only when their protein products are needed to perform their function—and then quickly degraded. The number of master regulator proteins is relatively small, and their expression and proteolysis are very tightly controlled. Environmental and cell status signals also tend to flow through master regulators.

The set of master regulators tends to be conserved as a system in related bacterial species, but the set of controlled genes is less conserved.[2] Bacterial species' fitness strategies are embodied in their master regulator genetic circuitry. The function of bacterial cells, and indeed all cells, is very machine-like, with every cell's processes for growth, division, and responses to internal and external signals tightly and predictably controlled by the embedded biochemical and genetic logic circuits. [Harley McAdams, Stanford University]



Combined CtrA and GcrA Transcriptional Network Creates Engine that Drives the Cell Cycle Forward. A complex oscillatory genetic circuit controls *Caulobacter crescentus* cell-cycle progression and asymmetric polar morphogenesis. Two tightly regulated master regulatory proteins, CtrA and GcrA, recently were shown to form the core oscillator.[1] Their intracellular concentrations activate or repress numerous cell cycle–regulated genes. Many of these genes are themselves top-level regulators of modular functions that execute the functions involved in cell-cycle progression (e.g., chromosome replication). Recent results elaborating this circuit include characterization of the regulons of two additional key *Caulobacter* cell-cycle regulatory proteins (publications in preparation).

References

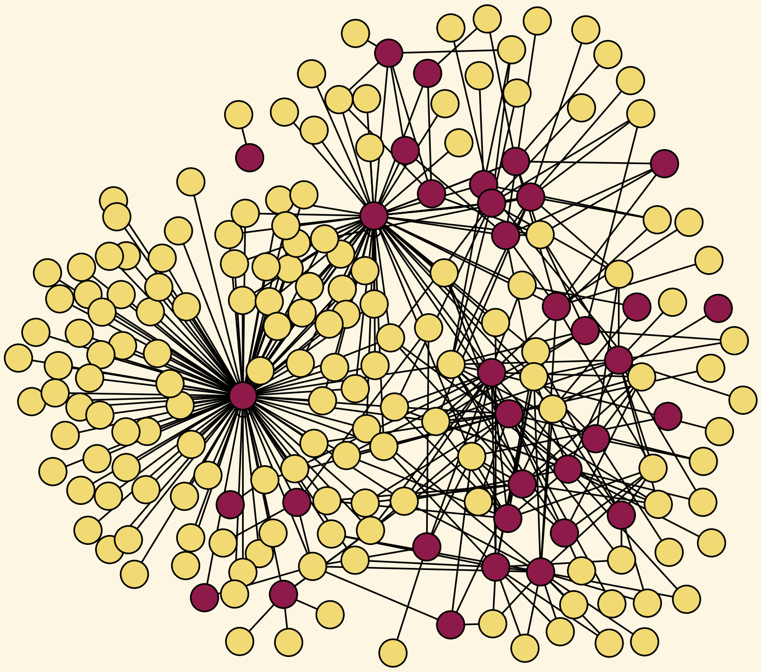
1. J. Holtzendorff et al., "Oscillating Global Regulators Control the Genetic Circuit Driving a Bacterial Cell Cycle," *Science* 304, 983–7 (2004).

2. H. H. McAdams, B. Srinivasan, and A. P. Arkin, "The Evolution of Genetic Regulatory Systems in Bacteria," *Nat. Rev. Genet.* 58, 169–78 (2004).

Capturing and Characterizing Protein Complexes, the Workhorses of the Cell

Comprehensively analyzing the molecular complexes that perform life's most essential functions presents many challenges due to their large number, biochemical variations, and dynamic nature. Some, such as ribosomes and other components of the cell's basic biosynthetic machinery, are present under most, if not all, growth conditions and are relatively stable. Other proteins and their complexes are expressed only under particular conditions and on an as-needed basis. Isolating and characterizing the range of molecular complexes present in microbial organisms require the development and validation of robust and complementary techniques.

GTL researchers at the Center for Molecular and Cellular Systems [a joint project of Oak Ridge National Laboratory (ORNL) and Pacific Northwest National Laboratory (PNNL)] have developed an integrated analysis pipeline that combines two complementary isolation approaches with mass spectrometry (MS) and computational tools for identifying protein complexes. This analysis pipeline uses molecular biology tools for expression of affinity-labeled proteins, highly controlled cell growth, affinity-based isolation of the complexes, and analysis of constituent proteins by MS. In addition, a bioinformatics infrastructure supports the entire pipeline, following samples "from cradle to grave" using a laboratory information management system integrated with data analysis and storage. This pipeline has been in continuous operation for over a year, focusing on two microbes relevant to DOE energy and environmental missions, *Rhodopseudomonas palustris* and *Shewanella oneidensis*. Extensive data are available for these organisms, including completed genome sequences.



Visualizing Interaction Networks. Graphical maps display protein interaction data in an accessible form. These visualizations summarize data from multiple experiments and also allow quick determinations of proteins that might be core constituents of a particular protein complex and those that might play roles in bridging interactions among different complexes. The figure above, generated using Cytoscape (www.cytoscape.org), summarizes affinity purification data from *Shewanella oneidensis*. Nodes (yellow or red circles) represent proteins identified from the integrated pipeline at the Center for Molecular and Cellular Systems, using both endogenous and exogenous protocols (see sidebar text). Probe proteins for affinity purifications are shown as red circles. Edges (black lines connecting nodes) are drawn between probe proteins and any other proteins confidently identified from a particular affinity-isolation experiment.

To isolate the complexes, the center employs two complementary affinity-based approaches in which tagged proteins are expressed either endogenously (in *Rhodopseudomonas* or *Shewanella* cells) or exogenously (in *Escherichia coli* or another surrogate cell) under specific experimental conditions. Combined liquid chromatography tandem mass spectrometry (LC MS/MS) is used to identify the isolated complexes.

Once a protein complex is identified, additional analytical tools are used to validate the complex. For example, imaging tools are employed to confirm the interactions of protein pairs in live cells using proteins expressed with fluorescent tags. At ORNL, high-performance Fourier transform ion cyclotron (FTICR) MS has been added to the analysis pipeline. This "top-down" approach analyzes the intact protein, relying on the high mass resolving power of FTICR MS to identify the full range of truncations and modifications present on the protein. The "bottom-up" conventional LC MS/MS method analyzes protein fragments and relies on databases to identify the original protein but cannot identify the full range of protein modifications. Thus, integrating the two types of MS provides detailed insights into the full identity of protein complex constituents.

Using these integrated methods to study 70S ribosomes from *R. palustris*, investigators obtained 42 intact protein identifications by the top-down approach, and 53 of 54 orthologs to *E. coli* ribosomal proteins were identified via bottom-up analysis. Scientists were able to assign post-translational modifications to specific amino acid positions and distinguish between isoforms. The combined MS data also allowed validation of gene annotations for three unusual ribosomal proteins (S2, L9, and L25) that were predicted to possess extended C-termini.[1] The low-complexity, highly repetitive sequences common to eukaryotes had not previously been identified experimentally at the protein level in prokaryotes.[2]

These early results underscore the need for multiple technologies to identify and characterize the thousands of protein complexes GTL studies will require each year and to eliminate the many bottlenecks that remain. [Michelle Buchanan, ORNL, and Steven Wiley, PNNL]

References

1. F. W. Larimer et al., "Complete Genome Sequence of the Metabolically Versatile Photosynthetic Bacterium *Rhodopseudomonas palustris*," *Nat. Biotechnol.* 22, 55–61 (2004).

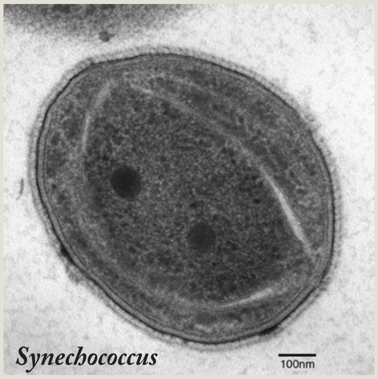
2. M. B. Strader et al., "Characterization of the 70S Ribosome from *Rhodopseudomonas palustris* Using an Integrated 'Top-Down' and 'Bottom-Up' Mass Spectrometric Approach," *J. Proteome Res.* 3, 965–78 (2004).

Synechococcus Encyclopedia: Integrating Heterogeneous Databases and Tools for High-Throughput Microbial Analysis

http://modpod.csm.ornl.gov/gtl/

High-throughput experimental data are extremely diverse in format and source and distributed across many internet sites, making integrated access to the information difficult. To address this challenge, GTL researchers at Sandia National Laboratories and Oak Ridge National Laboratory developed the *Synechococcus* Encyclopedia. This new computational infrastructural capability provides integrated access to 23 genomic and proteomic databases via an advanced-query language for browsing across multiple data sources. Sources include databases for sequence annotations, protein structure, protein interactions, pathways, and raw mass spectrometry and microarray data. Integrative analysis will yield major insights into the behavior of these abundant marine cyanobacteria and their importance to global carbon fixation. Also available are web-based analysis tools for exploration and analysis of information on the *Synechococcus* species.

These resources are enabling biologists to combine knowledge and see relationships that previously were obscured by the distributed nature and diverse data types present in biological databases. GTL researchers are using the tools to create knowledgebases for other organisms as well (e.g., *R. palustris* and *Shewanella*). [Grant Heffelfinger, Sandia National Laboratories]



Synechococcus

J. McCarren, Scripps Institution of Oceanography

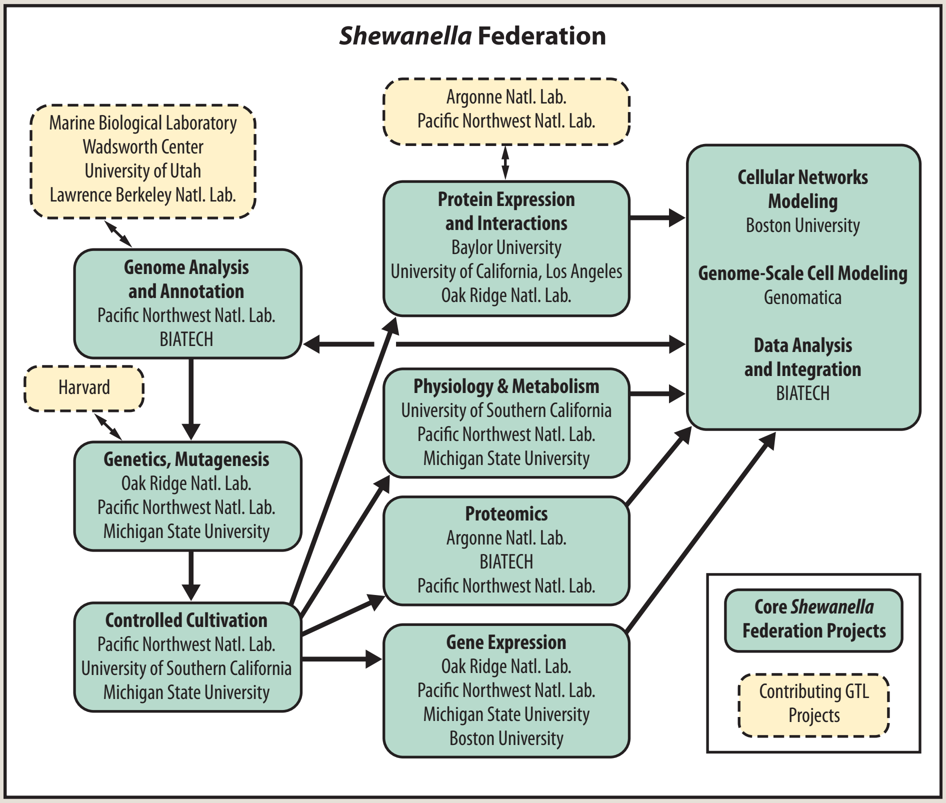
GTL RESEARCH PROGRAM

The Shewanella *Federation*

The *Shewanella* Federation, a multi-institutional consortium assembled by DOE, is applying high-throughput approaches for measuring gene and proteome expression of *Shewanella oneidensis* MR-1. The federation seeks to achieve a systems-level understanding of how this respiration-versatile microorganism regulates energy and material flow and uses its electron-transport system to reduce metals and nitrate. Leveraging substantial DOE investments in capability development and scientific knowledge, the *Shewanella* Federation employs an approach to systems that capitalizes on the relative strengths, capabilities, and expertise of each federation group. The federation conducts integrated and coordinated investigations that incorporate many facets of biological research and technologies across a number of disciplines and, hence, serve as a model for systems biology studies within the Genomics:GTL program. Federation members share information and resources and collaborate on projects consisting of a few investigators focused on a defined topic and on larger experiments combining their capabilities to address complex scientific questions. Several recent accomplishments are provided as examples below.

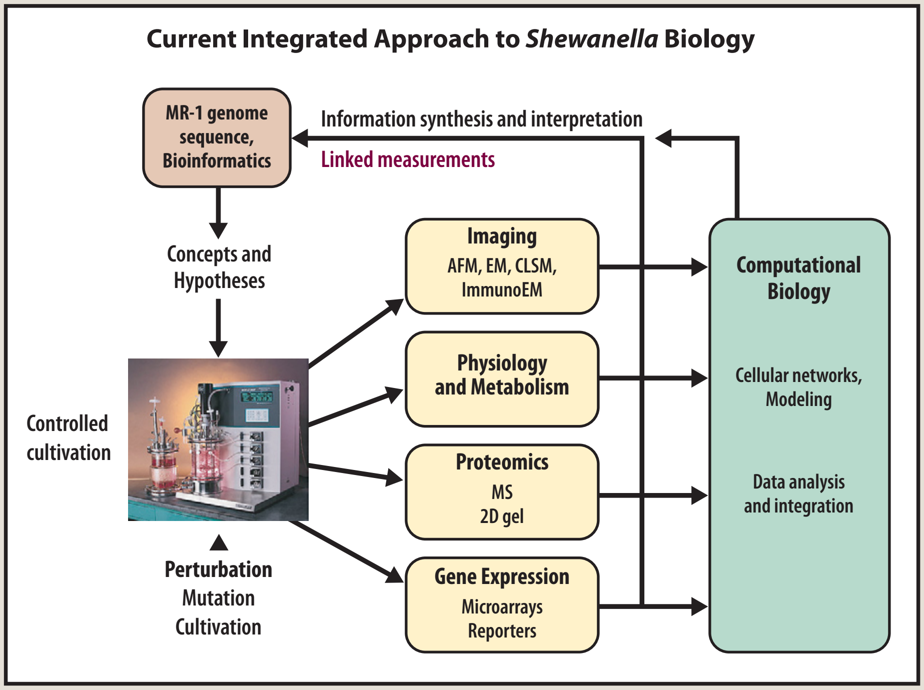
Combining Computational and Experimental Approaches to Enhance *Shewanella* Genome Annotation

Genomics, the study of all the genetic sequences in living organisms, has leaned heavily on the blueprint metaphor. A large part of the blueprint unfortunately has been unintelligible, requiring a way to link genomic features to what's happening in the cell. The *Shewanella* Federation has taken a significant step toward improving the interpretation of the blueprint for *S. oneidensis* MR-1. Federation members have applied a powerful new approach that integrates experimental and computational analyses to ascribe cellular function to genes that had been termed "hypothetical"—sequences that appear in the genome but whose biological expression and purpose previously were unknown. This approach currently offers the most-comprehensive "functional annotation," a way of assigning biological function to the mystery sequences and ranking them based on their similarity to genes known to encode proteins. Before this study, 1988 (nearly 40%) of the predicted 4931 genes in *S. oneidensis* were considered hypothetical.


Shewanella Federation
Argonne Natl. Lab.
Pacific Northwest Natl. Lab.
Marine Biological Laboratory
Wadsworth Center
University of Utah
Lawrence Berkeley Natl. Lab.
Protein Expression and Interactions
Baylor University
University of California, Los Angeles
Oak Ridge Natl. Lab.
Cellular Networks Modeling
Boston University
Genome-Scale Cell Modeling
Genomatica
Data Analysis and Integration
BIATECH
Genome Analysis and Annotation
Pacific Northwest Natl. Lab.
BIATECH
Harvard
Physiology & Metabolism
University of Southern California
Pacific Northwest Natl. Lab.
Michigan State University
Genetics, Mutagenesis
Oak Ridge Natl. Lab.
Pacific Northwest Natl. Lab.
Michigan State University
Proteomics
Argonne Natl. Lab.
BIATECH
Pacific Northwest Natl. Lab.
Controlled Cultivation
Pacific Northwest Natl. Lab.
University of Southern California
Michigan State University
Gene Expression
Oak Ridge Natl. Lab.
Pacific Northwest Natl. Lab.
Michigan State University
Boston University
Core *Shewanella* Federation Projects
Contributing GTL Projects

To gain insight into whether the sequences in fact produced proteins and the importance and function of any expressed hypothetical genes, a rigorous experimental approach was used. This approach involved growing the cells under a range of conditions to elicit expression of as many genes as possible, followed by comprehensive comparative analyses using a wide assortment of databases. High-throughput proteome and transcriptome analyses of MR-1 cells grown under a variety of conditions revealed that 538 of the hypothetical genes were expressed (proteins and mRNA) under at least one condition. The analyses confirmed that these are true genes used for one or more cellular processes.

Searches were undertaken to determine if existing databases could provide high-confidence insights into putative functions for these expressed genes (initially hypothetical). Of the 538 genes, 97% were identified as having homologs in other genomes, and general functional assignments were possible for 256 of them. Given the current amount and quality of experimental data in public genome databases, however, assigning exact biochemical function was possible for only 16 genes. These results and other arguments (Roberts 2004; Roberts et al. 2004) point to the need for new methods for understanding gene, protein, and, ultimately, organism function.


Current Integrated Approach to *Shewanella* Biology

The ability to rank hypothetical sequences according to their likelihood to encode proteins will be vital for any further experimentation and, eventually, for predicting biological function. The method not only portends a way to fill in the blanks in any organism's genome but also to compare the genomes of different organisms and their evolutionary relationships. In many cases, it is not known if a computationally annotated gene expresses a protein. With growing confidence that many hypothetical genes are expressing proteins, follow-on analyses now can be used to establish the role these proteins play.

Reference

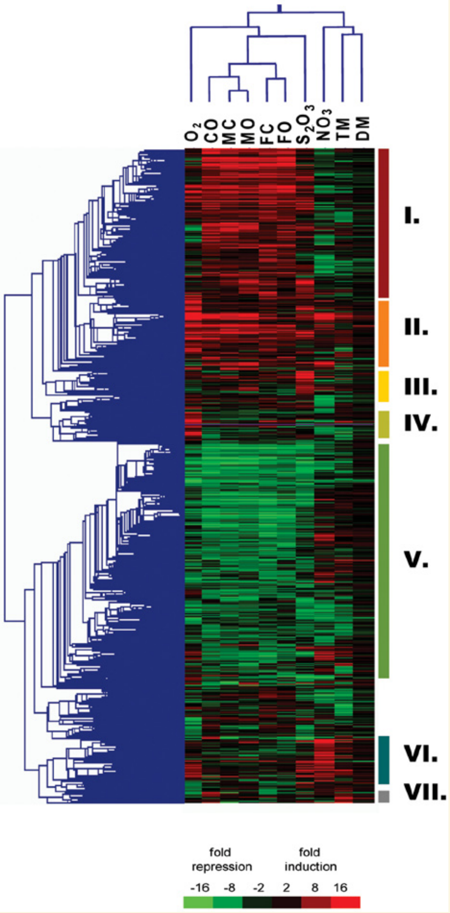
E. Kolker et al., "Global Profiling of *Shewanella oneidensis* MR-1: Expression of Hypothetical Genes and Improved Functional Annotations," *Proc. Natl. Acad. Sci. USA* 102, 2099-2104 (2005).

Physiologic, Genetic, and Proteome Response of *Shewanella oneidensis* to Electron Acceptors

As a facultative anaerobe and dissimilatory metal-reducing bacterium, *S. oneidensis* MR-1 can shift its metabolism and flexible electron-transport system to allow it to thrive in environments with steep redox gradients. It can accommodate O_2 as a terminal electron acceptor, or it can generate energy from anaerobic respiration using a variety of soluble (e.g., nitrate, thiosulfate) and insoluble electron acceptors such as Fe(III) and Mn(IV). This major shift in lifestyle probably requires rewiring of electron transport and metabolism by sensing changes in the environment and making the necessary changes in cellular proteins or the proteome. To begin to understand how MR-1 cells respond at the whole-cell or "system" level to this transition to anaerobicity, the federation initiated a series of experiments in which MR-1 was grown under changing conditions in continuous culture. These experiments revealed that MR-1 cells growing at high oxygen concentrations formed cell aggregates—the precursor to biofilms. They also exhibited elevated expression levels of genes involved in attachment and autoaggregation including fimbrae (curli, pili, flagella), extracellular polysaccharides, lectins, and surface antigens. These studies indicated that aggregation in

S. oneidensis MR-1 may serve as a mechanism to facilitate reduced O_2 tensions to cells within the aggregate interior, avoiding the oxidative stress associated with production of reactive oxygen species during metabolism.

To gain insight into the complex structure of the energy-generating networks in MR-1, global mRNA patterns were examined in cells exposed to a wide range of metal and nonmetal electron acceptors. Gene-expression patterns were similar regardless of which metal ion was used as electron acceptor, with 60% of the differentially expressed genes showing similar induction or repression relative to fumarate-respiring conditions (Fig. A). Several groups of genes exhibited elevated expression levels in the presence of metals, including those encoding putative multidrug efflux transporters, detoxification proteins, extracytoplasmic sigma factors, and PAS-domain regulators. Only one of the 42 predicted *c*-type cytochromes in MR-1, SO3300, displayed significantly elevated transcript levels across all metal-reducing conditions. Genes encoding decaheme cytochromes MtrC and MtrA, which were linked previously to reduction of different forms of Fe(III) and Mn(IV), exhibited only slight decreases in relative mRNA abundances under metal-reducing conditions. In contrast, specific transcriptome responses were displayed to individual nonmetal electron acceptors, resulting in identification of unique groups of nitrate-, thiosulfate- and TMAO-induced genes including previously uncharacterized multicytochrome gene clusters. Collectively, gene-expression results reflect the fundamental differences between metal and nonmetal respiratory pathways of *S. oneidensis* MR-1, in which the coordinate induction of detoxification and stress-response genes play a key role in adaptation of this organism under metal-reducing conditions. [*Shewanella* Federation]



Electron-Acceptor Abbreviations

O_2	oxygen
CO	Co(III) EDTA
FC	ferric citrate
FO	hydrous ferric oxide
MC	colloidal manganese(IV)
MO	manganese dioxide
S_2O_3	sodium thiosulfate
NO_3	sodium nitrate
TM	trimethylamine N-oxide
DMSO	dimethyl sulfoxide

Fig. A. Hierarchical Clustering of 2827 *S. oneidensis* Genes Exhibiting Differential Expression in Cultures Exposed to Metal and Nonmetal Electron Acceptors. Each column indicates an individual electron acceptor. Relative transcript abundances represent a ratio of mRNA levels in test cultures to these displayed by the fumarate-grown reference.

Reference

A. S. Beliaev et al., "Global Transcriptome Analysis of *Shewanella oneidensis* MR-1 Exposed to Different Terminal Electron Acceptors," *J. Bacteriol*., accepted for publication.

Bacteria Use "Nanowires" to Facilitate Extracellular Electron Transfer

GTL science and capabilities are being leveraged to identify and characterize the composition, function, and expression of extracellular appendages grown by some bacteria to facilitate electron transfer in challenging environments important to DOE missions. These appendages are electrically conductive and are hypothesized to function as biological "nanowires." Below are some highlights of research on nanowires in *Shewanella* and *Geobacter* species. In addition to providing insights into microbes with potential uses in bioremediation strategies, these remarkable structures may one day have commercial applicability.

Shewanella

Nanowires were revealed in *S. oneidensis* MR-1 cells experiencing electron-acceptor limitation (EAL) using scanning tunneling microscopy (STM) and tunneling spectroscopy. *Shewanella* is a metabolically versatile bacterium that uses a variety of electron acceptors, including nitrate, metals such as solid-phase iron and manganese oxides, and radionuclides such as uranium and technetium. A GTL *Shewanella* collaborative team uses an integrated approach to study this organism's electron-transport and energy-transduction systems.

Nanowires Facilitate Extracellular Electron Transfer via *c*-Type Cytochromes. *Shewanella* nanowires were observed using scanning electron microscopy (SEM, Fig. A) and STM (Fig. B) of MR-1 cells grown in chemostats under EAL. The ability to be imaged by STM indicated that the material is conductive, allowing electrons to tunnel from the probe tip to the underlying graphite surface. Peptide-specific antibodies against outer membrane cytochromes MtrC and OmcA were used in immunoEM experiments to investigate their cellular location. ImmunocytoTEM (transmission electron microscopy) analysis of MR-1 cells grown under EAL revealed that MtrC (Fig. C) and OmcA (not shown) are associated with extracellular structures morphologically identical to the MR-1 nanowires observed by SEM and STM.

Nanocrystalline Magnetite Particles are Associated with Nanowires. Since nanowires can conduct electrons in vitro, investigations have been made of the association between nanowires and the Fe(III) mineral

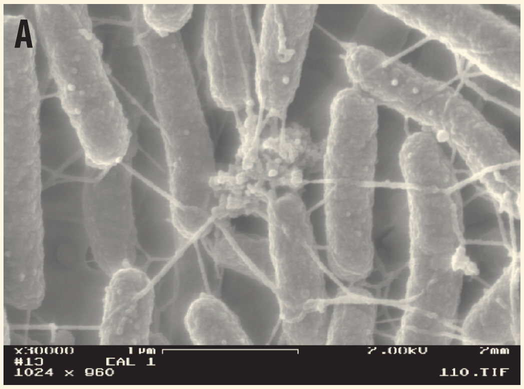


Fig. A. SEM MR1 Grown in a Bioreactor under EAL Conditions. Sample was prepared by critical-point drying.

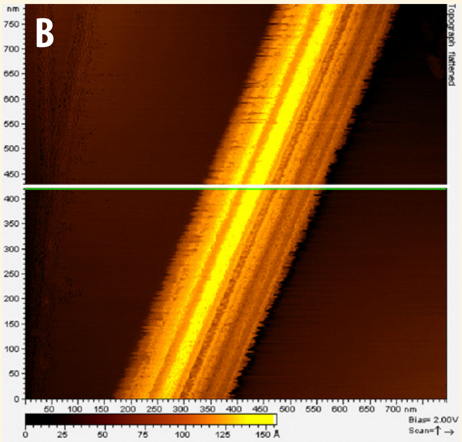


Fig. B. STM Image of Isolated Nanowire from Wild-Type MR-1. Nanowire has lateral diameter of 100 nm and topographic height of 5 to 10 nm. High magnification shows ridges and troughs running along the structure's long axis.

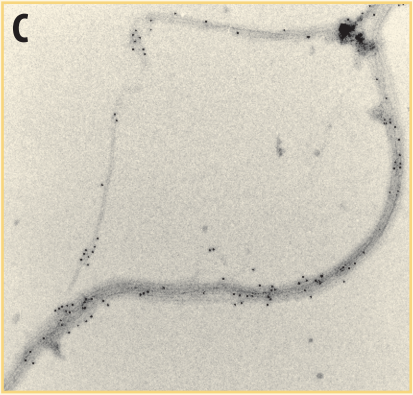


Fig. C. Immunogold Labeling of MtrC on Nanowires. TEM images of whole mounts of MR-1 nanowires from cells grown in continuous culture under EAL conditions reveal that MtrC is localized specifically to the nanowires.

ferrihydrite in vivo. TEM analyses of MR-1 grown anaerobically in the presence of ferrihydrite revealed nanocrystalline magnetite arranged in linear arrays along features consistent with nanowires (Fig. D).

In addition, a mutant deficient in the outer membrane decaheme cytochromes MtrC and OmcA was unable to reduce hydrous ferric oxide or transfer electrons directly to electrodes in a mediator-less fuel cell, directly linking these cytochromes to extracellular electron transfer in MR-1. Also observed was the production of nanowires in several other microbes in direct response to electron-acceptor limitation, including *Geobacter sulfurreducens* and *Desulfovibrio desulfuricans*, suggesting that nanowires may be common to other bacteria and microbial consortia dependent on electron transfer. Furthermore, nanowires could be responsible for cell-to-cell electron-transfer processes in biofilms and complex microbial mat communities. [Yuri Gorby and Jim Fredrickson, Pacific Northwest National Laboratory]

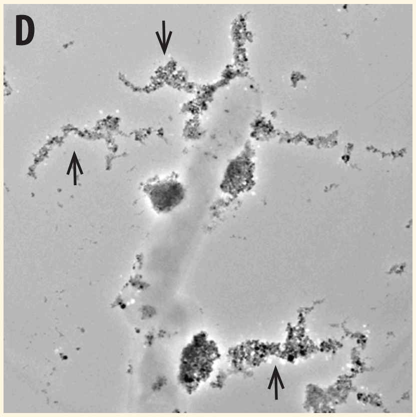


Fig. D. Magnetite Associated with Nanowires. TEM images of whole mounts of MR-1 cells incubated with the Fe(III) mineral ferrihydrite revealed the formation of nanocrystalline magnetite along the nanowires (indicated by arrows).

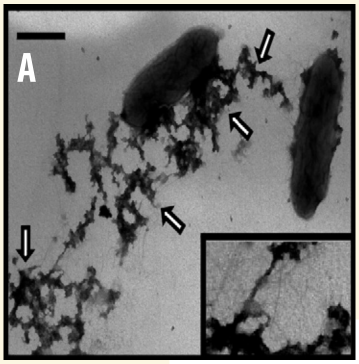
Reference

A. S. Beliaev et al., "MtrC, an Outer Membrane Decahaem *c* Cytochrome Required for Metal Reduction in *Shewanella putrefaciens* MR-1," *Mol. Microbiol.* 39, 722–30 (2001).

Geobacter

Field experiments have demonstrated that stimulating the growth of *Geobacter* species in uranium-contaminated subsurface environments precipitates the uranium from the groundwater and prevents its spread. To support their growth, *Geobacter* species require Fe(III) oxide minerals, naturally present in the subsurface, as an electron acceptor. Transferring electrons outside the cell onto an insoluble mineral represents a physiological challenge not faced by microorganisms that use such commonly considered, soluble electron acceptors as oxygen, nitrate, and sulfate. Understanding electron transfer to Fe(III) oxide is essential to optimize strategies for the in situ bioremediation of uranium-contaminated groundwater.

Fig. A. *Geobacter* Pilin Nanowires with Fe(III) Oxide Attached.



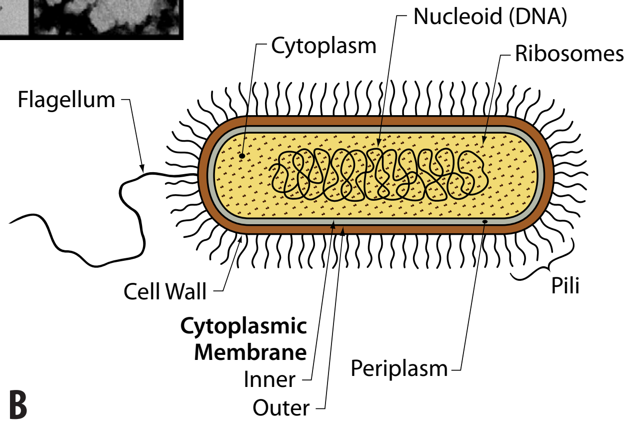


Fig. B. View of Simplified Microbial Anatomy.

Pili Extend as Nanowires to Transfer Electrons. Investigators noted that *Geobacter* species specifically produced fine, hair-like structures known as pili on one side of the cell during growth on Fe(III) oxide.[1] Knocking out a key gene for pili production prevented *G. sulfurreducens* from growing on insoluble Fe(III) oxides but had no effect on growth with soluble electron acceptors. Although pili in other organisms often function in attachment to surfaces, the mutant strain could attach to Fe(III) oxides as well as the wild-type strain. Further investigation with an atomic force microscope fitted with a tip capable of conducting electrical current demonstrated that the pili of *G. sulfurreducens* are highly conductive. These results suggest that *Geobacter* species are able to transfer electrons onto Fe(III) oxide with conductive pili that extend as nanowires from the cell.[2] Mechanisms for pili conductivity and electron transfer have yet to be eludicated.

Potential Applications of Nanowires. Conductive pili produced by *G. sulfurreducens* are only 3 to 5 nm wide. A wire this thin that can be mass-produced biologically may have a variety of nanoelectronic applications. Furthermore, genetically modifying *G. sulfurreducens* pili structure or composition to generate nanowires with different functionalities may have significant commercial value. [Derek Lovley, University of Massachusetts]

References

1. S. E. Childers, S. Ciufo, and D. R. Lovley, "*Geobacter metallireducens* Accesses Fe (III) Oxide by Chemotaxis," *Nature* **416**, 767–69 (2002).

2. G. Reguera et al., "Extracellular Electron Transfer Via Microbial Nanowires," *Nature* **435**, 1098–1101 (2005).

Synthetic Genome Research

Accurate, Low-Cost Gene Synthesis from Programmable DNA Microchips

Technologies are needed for accurate and cost-effective gene and genome synthesis to support protein production and test the many hypotheses from genomics and systems biology experiments. GTL researchers have developed a microchip-based technology enabling multiplex gene synthesis suitable for large-scale synthetic biology projects. In this approach, pools of thousands of "construction" oligonucleotides (oligos) and tagged complementary "selection" oligos are synthesized on photo-programmable microfluidic chips, released, amplified, and selected by hybridization to reduce synthesis errors ninefold. The oligos then are assembled into multiple genes using a one-step polymerase assembly multiplexing reaction.

These microchips were used to synthesize all the 21 protein-encoding genes making up the *Escherichia coli* small ribosomal subunit, with translation efficiencies optimized via alteration of codon usage. Researchers estimate that the chip's synthetic capacity may potentially increase cost-efficiency in oligo yields from 9 bp to 20,000 bp per dollar, depending on the microchip and number of oligos. This technology represents a powerful tool for synthetic biology and complex nanostructures in general. [George Church, Harvard University]

Reference

J. Tian et al., "Accurate Multiplex Gene Synthesis from Programmable DNA Microchips," *Nature* **432**, 1050–54 (2004).

Generating a Synthetic Genome

Researchers at the Institute for Biological Energy Alternatives (IBEA, now called the J. Craig Venter Institute) have advanced methods to improve the speed and accuracy of genomic synthesis. The team assembled the 5386-bp bacteriophage φX174 (phi X), using short, single strands of synthetically produced, commercially available DNA (oligonucleotides). Researchers employed an adaptation of the polymerase chain reaction (PCR) known as polymerase cycle assembly (PCA) to build the phi X genome. Like PCR, PCA is a technique that produces double-stranded copies of individual gene sequences based on single-stranded templates. IBEA assembled the synthetic phi X in just 14 days.

Reference

H. O. Smith et al., "Generating a Synthetic Genome by Whole Genome Assembly: φX174 Bacteriophage from Synthetic Oligonucleotides," *Proc. Natl. Acad. Sci.* **100**(26), 15440–445 (2003).

Harvesting Electricity from Aquatic Sediments with Microbial Fuel Cells

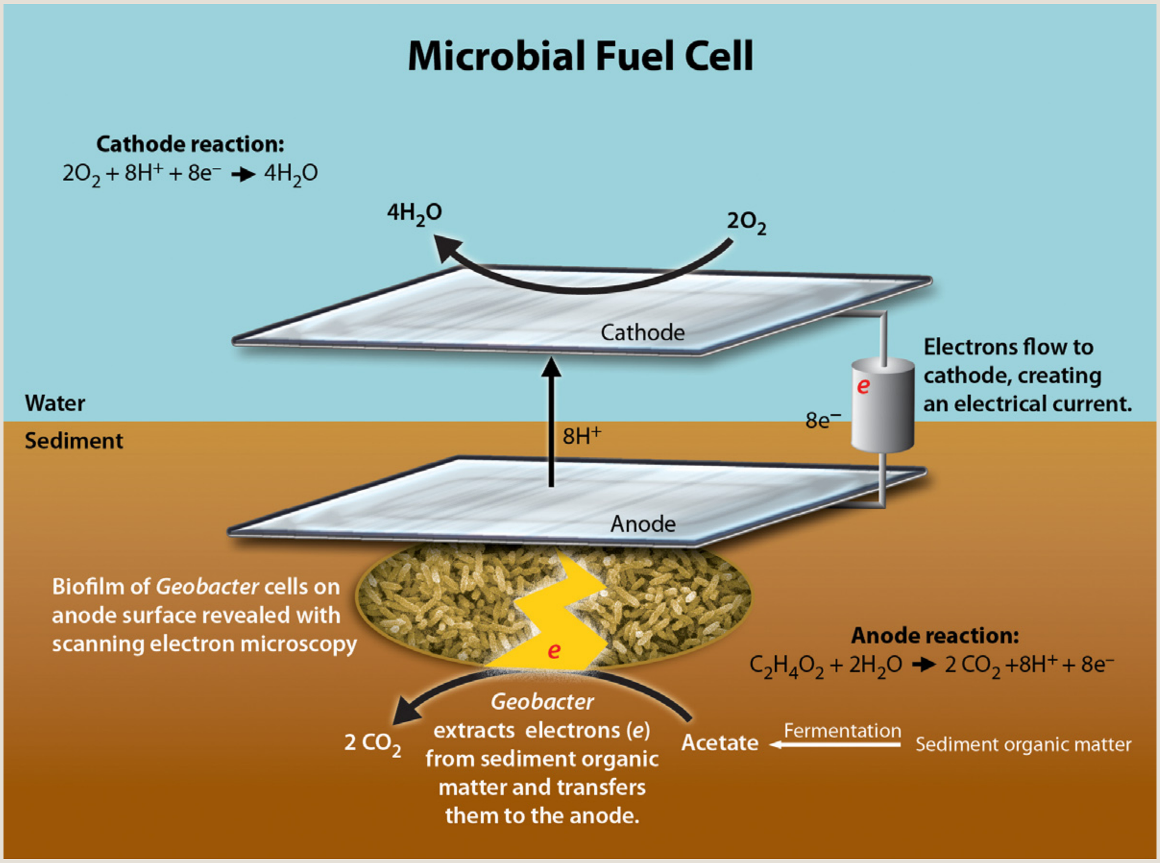
Microorganisms known as "electricigens" can efficiently convert organic wastes, renewable biomass, and even mud into electricity and harmless by-products. This capability offers the potential for using microbes (or their components) to generate electricity at low cost while transforming industrial, domestic, and farm wastes. GTL studies are exploring how some microbes accomplish these processes naturally.

The family *Geobacteraceae* can metabolize organic compounds directly at electrode surfaces, transferring electrons and producing an electrical current. Genome-scale analysis revealed that when *G. sulfurreducens* grows on electrodes, it produces high levels of a cytochrome (OmcS), displaying it on the outside of the cell. These studies also demonstrated that OmcS is required for power production, which stops when OmcS is removed and resumes when the gene is restored.

GTL investigators are collaborating with the automotive industry to use this information for designing improved microbial fuel cells—microbe-powered batteries that can convert organic matter to electricity. In contrast to commonly considered hydrogen fuel cells requiring highly refined clean fuels, microbial fuel cells can harvest electricity from relatively low-quality, dirty fuels or from biomass without extensive preprocessing. By engineering electrodes that interact better with OmcS or microbes that make more OmcS, increasing the power output of microbial fuel cells and expanding their practical applications is possible. Potential uses range from powering small electronic devices and robots that can "live off the land" to serving as localized domestic power sources for household uses. [Derek Lovley, University of Massachusetts]

Reference

D. R. Bond et al., "Electrode-Reducing Microorganisms that Harvest Energy from Marine Sediments," *Science* 295, 483–85 (2002).


Microbial Fuel Cell
Cathode reaction:
$2O_2 + 8H^+ + 8e^- \rightarrow 4H_2O$
$4H_2O$
$2O_2$
Cathode
Electrons flow to cathode, creating an electrical current.
e
8e⁻
Water
Sediment
$8H^+$
Anode
Biofilm of *Geobacter* cells on anode surface revealed with scanning electron microscopy
e
Anode reaction:
$C_2H_4O_2 + 2H_2O \rightarrow 2\,CO_2 + 8H^+ + 8e^-$
$2\,CO_2$
Geobacter extracts electrons (*e*) from sediment organic matter and transfers them to the anode.
Acetate
Fermentation
Sediment organic matter

3.4. GTL Program and Facility Governance

The GTL program will establish a governance process to ensure advancement of DOE, GTL, and research-community objectives. The program will continue to fund a balanced portfolio of merit-reviewed research projects at universities and national laboratories and in the private sector. A mix of large multi-institutional, interdisciplinary research projects and single-investigator studies will focus on fundamental GTL science, GTL facility pilots, and technology development.

Excellence in every facet of research and operations will be the hallmark of the GTL facilities, including optimized operations and continuous facility and equipment enhancement. Relevance to missions and the scientific community's foci will be supported by continuous peer review and oversight. Key operational and management governance processes and their objectives for the GTL program and facilities include:

- Process of review for excellence and relevance to guarantee the best science and efficient capacity allocation and to achieve optimum performance and output.
- Facility and program-development mechanisms to keep research objectives on track with new discoveries.
- Workflow management for efficient and effective facility and equipment operations.
- Appropriate user-community access to provide open but secure and prioritized use of facilities and access to products and data.
- Performance-measurement metrics to assess quality of outputs.

3.4.1. Facility User Access

In the tradition of a long history of DOE user facilities, access will be open, based on a peer-review process that will judge science quality and relevance and the need to use these valuable national assets. Factors in judging proposals will include inventiveness of the science, relevance to solution of mission problems, quality and breadth of interdisciplinary teams, institutional capabilities to execute the science, performance records of investigators, and quality of the plan to use facility outputs (e.g., computational analyses, high-throughput research technologies, systems biology concepts, and research resources).

This formula allows for the study not only of systems with direct relevance to DOE missions but also of model systems that could shed light on DOE microbes that by their nature are less studied, unstudied, or vastly more complex. Model systems thus would be used to test facility technologies, resources, methods, and concepts and disseminate concepts to less-defined systems.

GTL's dedicated user facilities will provide the broader scientific community with technologies, research resources, and computing and information infrastructure. Based on a policy of open access, the most sophisticated and comprehensive capabilities, reagents, and data will be available to investigators lacking such integrated technology suites in their own laboratories or institutions. The facilities also will provide a venue for user groups to develop scientific approaches and technologies to make optimum use of facility outputs and advance the practice of systems microbiology.

Pilot studies will enable development of large-scale systems biology experimental protocols, including those for remote facility access. Other factors contributing to the operational and scientific success of GTL facilities will be advisory, review, and community meetings to facilitate feedback and sharing of information and lessons learned. A peer-review process for selecting user projects and principles for experiment prioritization will be established, clear lines of communication between each facility and research constituencies will be created, and data-access sites will be community oriented.

3.4.2. Collaborative Environment

The GTL program and facilities will foster communication and collaboration to share concepts used by the multi-institutional, multidisciplinary teams working on complex systems-microbiology problems. Specialized web interfaces and state-of-the-art electronic conferencing mechanisms will be used routinely, and, as requested, jamborees for data analysis and setting of research strategies could be held by facility staff in collaboration with research-team leaders.

3.4.3. Facility Governance

Management and financing of programs have evolved over the years, particularly for facilities, and most user facilities are now managed with what is termed the "Steward-Partner model."* This model was developed to ensure that user facilities provide the maximum scientific benefit to the broadest possible research community in the most cost-effective manner (see footnote for details).

In this model, DOE (the steward) would manage and fund the core facilities. Research would be conducted by scientists (the partners) supported by the steward and by other federal agencies, industry, or private institutions. Principles governing the Steward-Partner model would be used to provide GTL facility resources to the scientific community. Good stewards of the investments and trust would determine use and access by objective merit-based peer review of both scientific quality and programmatic relevance.

In this spirit, management and operations of GTL facilities will be the host institution's responsibility with appropriate provisions, measurements, and metrics in accordance with its management contract. Facility management will have input from advisory panels focused on six topics: Science, Technology, User Access, Programmatic Impact, Prioritization, and Interfacility Coordination, each appointed by and under the direction of host-institution management with DOE involvement and approval. The panels, with guidance and feedback from the user community, will help establish user-access procedures.

Facility management and DOE will work with user and advisory groups to consider management structure, operations team, research and development priorities, process and capacity-allocation metrics, reporting mechanisms, advisory panels, remote vs local facility use, QA/QC protocols, and experimentation and scope requirements. Ongoing objectives will be to establish R&D teaming, QA/QC milestones, and production goals and metrics. Guidelines to be set include operational rules, facility community-access rules, and user and broad community access to data and computing tools, protocols, and experimental details.

3.5. Training

The GTL program also is committed to training as a means of enlarging the workforce involved in large-scale quantitative biology to help solve DOE mission problems and to ensure an efficient and safe work environment. Training must fit users at any stage of their careers, whether undergraduate, postgraduate, or senior scientists. Educational and collaboration-fostering activities will focus on single and crosscutting technologies and computing that encompass capabilities provided by a facility or used by research programs. These activities include interfacing with analytical technologies in investigators' laboratories and integrating next-generation strategies and technologies into existing strategies. Different training modes such as web-based information and courses, onsite workshops, minicourses, and symposia at major scientific meetings must be established.

*The Steward-Partner Model was implemented in the report, *Synchrotron Radiation for Macromolecular Crystallography, Office of Science and Technology Policy* (January 1999). The model is described in some detail in the National Research Council report, *Cooperative Stewardship: Managing the Nation's Multidisciplinary User Facilities for Research with Synchrotron Radiation, Neutrons, and High Magnetic Fields* (National Academy Press, 1999). It also is followed in the more recent report, *Office of Science and Technology Policy Interagency Working Group on Neutron Science: Report on the Status and Needs of Major Neutron Scattering and Instruments in the United States* (June 2002). (Reports: http://clinton2.nara.gov/WH/EOP/OSTP/Science/html/cassman_rpt.html; www.nap.edu/books/0309068312/html/; and www.ostp.gov/html/NeutronIWGReport.pdf, respectively.)

Of particular interest to the broad research community will be data, models, and concepts in the GTL Knowledgebase for application to other areas of biology. Web-based documentation, publications, tutorials, workshops, and symposia at scientific and computing meetings will facilitate knowledgebase use.

The laboratory information management system (LIMS) is central to facility operations and data output and integration, so all users will be trained in relevant aspects of the facility's LIMS. Online documentation and tutorials will be central to this process and to learning about computational analysis and database use.

Ongoing oversight and peer review of training operations will ensure the curriculum's continued excellence and relevance.

3.6. Ethical, Legal, and Social Issues (ELSI)

The Human Genome Project (HGP) was a technology- and data-development project, whose infrastructure and tools ultimately would enable human genetics and medical questions to be answered. A program was established within HGP to identify and explore the ethical, legal, and social issues (ELSI) that were expected to arise. The HGP ELSI program became a significant contributor to genetics policy in the United States as well as a model for additional bioethics programs both here and in other countries.

Ultimately, the formulation of policy is a public responsibility to which scientists should contribute in two ways: (1) as citizens with an obligation to become informed and participate in the discussion and (2) as sources of reliable, accurate, objective, and relevant information.

3.6.1. GTL Commitment to Explore ELSI Impacts

Genomics:GTL is largely a microbiology program at this point, but it encompasses a number of scientific activities that could be expected to impact society and the environment. GTL's explicit intention and commitment are to explore these impacts where appropriate and nonoverlapping with others' efforts. Furthermore, GTL commits to stressing the close coordination of ELSI activities and studies with ongoing scientific research. ELSI will be one of the topics covered under the crosscutting management process as GTL moves forward (see 6.0. GTL Development Summary, p. 191).

3.6.1.1. Examples of Potential GTL ELSI Issues

ELSI concerns are being raised both by the missions being addressed and by specific topics of scientific research. Consideration of topics will be guided by the uses to which research might be put. These uses include developing new energy sources, cleaning up environmental contaminants, and exploring biological ways of managing excess carbon dioxide in the atmosphere. Items below are meant to serve only as examples and will evolve as the GTL program develops, as external guidance and counsel are obtained, and when outside societal impacts occur or are anticipated. ELSI issues expected to be important to GTL may include (but are not limited to)

- The impact of "synthetic" biology, which is the ability to "engineer" simple life forms with specific properties.
- Societal impacts of progress in elucidating microbial mechanisms of energy production or more effective toxic-metal and radionuclide cleanup.

3.6.1.2. Using Microbial Diversity for Practical Applications

GTL researchers use genome data and tools to isolate and manipulate genes and their products as they search for insights into the fundamental workings of life processes. These tools support a broad range of activities required for GTL, from high-throughput protein production (beginning with gene sequence) to manipulation of genes to aid in characterizing the physical and functional differences in protein products and their interactions.

Practical applications of GTL research will be based on microbial enzymes, which may be optimized and used as isolated components or within microbes residing in controlled environments, for example, in fermentors. Synthetic systems or genetically engineered releases, to date, have not proven to be cost-effective or necessary in practical applications. Indeed, through environmental genomics we are discovering such great diversity in nature's tool kit that it may suffice to simply pick and choose the correct microorganism and encourage natural systems to work for us.

3.6.2. The Path Forward

GTL managers and planners recognize that this list of possible ELSI issues is only a starting point. New societal issues are expected to arise, and GTL will explore those linked to the GTL program and DOE mission applications. The goal for the GTL ELSI program is to seek insights into science implications in a way and at a time when course corrections for the program, if needed, can be suggested with the least disruption. Larger social issues are the purview of other agencies. For example, the President's Council on Bioethics and the National Science Advisory Board for Biosecurity (www.biosecurityboard.gov) provide guidance to federal agencies on strategies for appropriate conduct in biotechnology research.