DOE/SC-0090

DOE GENOMICS:GTL ROADMAP

SYSTEMS BIOLOGY FOR ENERGY AND ENVIRONMENT

AUGUST 2005

Office of Biological and Environmental Research
and
Office of Advanced Scientific Computing Research

Office of Science
U.S. DEPARTMENT OF ENERGY

DOEGenomesToLife.org

Genomics:GTL Programmatic Background

The Department of Energy's (DOE) Office of Science (SC) plays four key roles in U.S. research:

- Contributes essential scientific foundations to DOE's national energy and economic security missions;
- Invests in research at more than 280 universities, 15 national laboratories, and many international institutions;
- Builds and operates major research facilities for open access by the science community; and
- Supports core capabilities, theories, experiments, and simulations at the extreme limits of science.

An SC goal for its Office of Biological and Environmental Research (BER) is to "harness the power of our living world and provide the biological and environmental discoveries necessary to clean and protect our environment and offer new energy alternatives" [*Office of Science Strategic Plan* (2004)].

To address this priority, BER and SC's Office of Advanced Scientific Computing Research (OASCR) are sponsoring the Genomics:GTL program. Established in 2002, GTL uses microbial genome data to launch investigations of microbes with capabilities relevant to DOE energy and environmental missions. The GTL scientific program was developed with input from hundreds of scientists from universities, private industry, other federal agencies, and DOE national laboratories. Many genome sequences used in GTL and determined by BER programs have made important contributions to the understanding of biology, genetics, and evolution.

Scientific and technological progress during the Human Genome Project, initiated in 1986 by DOE, provides the foundation for GTL research.

SC's goal for OASCR is to deliver computing for the frontiers of science. OASCR's primary missions are to discover, develop, and deploy integrated computational and networking tools that enable researchers in scientific disciplines to analyze, model, simulate, and predict complex phenomena important to DOE. To this end, OASCR fosters and supports fundamental research in advanced scientific computing—applied mathematics, computer science, and networking—and operates supercomputer, networking, and related facilities. OASCR's leadership will be critical to GTL's success.

To aid program progress, GTL is in the process of populating its web site with such communications resources as research in progress, image galleries, presentations, fact sheets, topical web pages, and meeting calendars. Feedback and other interactions are welcome to assist in the development of dynamic, next-generation web tools for educational purposes and to facilitate the research of diverse scientific contributors and users of GTL resources and data.

Visit the roadmap web site for electronic download of this publication and its graphical content. To order hardcopies or submit comments, call or use the web:

- doegenomestolife.org/roadmap/
- 865.576.6669

David Thomassen
U.S. Department of Energy (SC-23)
Office of Biological and Environmental Research
301.903.9817, david.thomassen@science.doe.gov

Gary Johnson
U.S. Department of Energy (SC-21.1)
Office of Advanced Scientific Computing Research
301.903.5800, gary.johnson@science.doe.gov

DOE Office of Science: www.science.doe.gov

Suggested citation for this document: *Genomics:GTL Roadmap: Systems Biology for Energy and Environment*, U.S. Department of Energy Office of Science, August 2005.

DOE/SC-0090

DOE GENOMICS:GTL ROADMAP

SYSTEMS BIOLOGY FOR ENERGY AND ENVIRONMENT

August 2005

Prepared for the
U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research
Office of Advanced Scientific Computing Research
Germantown, MD 20874-1290

Prepared by
Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830
Managed by UT-Battelle, LLC
For the U.S. Department of Energy
Under contract DE-AC05-00OR22725
Comments: http://public.ornl.gov/hgmis/gtlroadmap.cfm

Preface

Welcome to the roadmap for the Department of Energy's (DOE) Genomics:GTL program (GTL). Prepared with the involvement of hundreds of scientists and technologists over the past 3 years, this document traces the path from DOE mission challenges to the science and technology base that will enable their biotechnological solutions. GTL's program goal is to use systems biology approaches to understand microbes so well that their diverse capabilities can be harnessed for many DOE and other national needs.

A key element of the GTL program is an integrated computing and technology infrastructure, which is essential for timely and affordable progress in research and in the development of biotechnological solutions. In fact, the new era of biology is as much about computing as it is about biology. Because of this synergism, GTL is a partnership between our two offices within DOE's Office of Science—the Offices of Biological and Environmental Research and Advanced Scientific Computing Research.

Only with sophisticated computational power and information management can we apply new technologies and the wealth of emerging data to a comprehensive analysis of the intricacies and interactions that underlie biology. Genome sequences furnish the blueprints, technologies can produce the data, and computing can relate enormous data sets to models linking genome sequence to biological processes and function.

DOE is committed to establishing the necessary science base, which will be translated into important applications by programs across DOE. Because grand challenges will not submit to incremental approaches, the GTL program will build four advanced biology user facilities. Their research environment will comprise suites of technologies, methods, and computing, along with training to use facility resources. The new infrastructure will be a resource not only for the scientific community but for industry, allowing rapid translation of science into new technologies.

We believe the roadmap will serve as the foundation for involving scientists, engineers, and technologists from academia, industry, and the national laboratories in GTL research and in the design and development of GTL user facilities, in the conduct of necessary research and technology development, and in preparing the scientific community to use the new resources. We hope this document and related information available on the supporting web site (www.doegenomestolife.org) will inspire and encourage participation in this important challenge.

Pursuing mission science goals often has required grappling with seemingly intractable challenges, but they have taken us to fascinating places where we have made surprising discoveries. We expect our new quest on the frontier of biology to prove equally exciting.

Aristides A. N. Patrinos, Director
Office of Biological and Environmental Research
U.S. Department of Energy
ari.patrinos@science.doe.gov

Robin Staffin, Acting Director
Office of Advanced Scientific Computing Research
U.S. Department of Energy
robin.staffin@science.doe.gov

Executive Summary

Providing solutions to major national problems, biology and industrial biotechnology will serve as an engine for economic competitiveness in the 21st Century. Department of Energy (DOE) missions in energy security, environment, and climate are grand challenges for a new generation of biological research. As a mission agency, DOE can bring together biological, computing, and physical sciences for the focused and large-scale research effort needed—from scientific investigations to commercialization in the marketplace.

Our investment in genomics over the past 20 years now allows us to rapidly determine and interpret the complete DNA sequence of any organism. Because it reveals the blueprint for life, genomics is the launching point for an integrated and mechanistic systems understanding of biological function. It is a new link between biological research and the development of biotechnologies. With genomics data as a starting point, the Genomics:GTL program (GTL) will use a systems biology approach to fundamentally transform the way scientists conduct biological investigations and describe living systems.

GTL's goal is simple in concept but complicated in practice—to reveal how the static information in genome sequences drives the intricate and dynamic processes of life. Through predictive models of these life processes and supporting research infrastructure, we seek to harness the capabilities of microbes and complex microbial communities, which are the foundation of the biosphere and sustain all life on earth. Gaining reliable use of microbial processes requires understanding the whole living system, not just genomic DNA sequences or collections of proteins or cell by-products. GTL will study critical microbial properties and processes on three systems levels—molecular, cellular, and community—each requiring advances in fundamental capabilities and concepts.

Already, discoveries in the microbial world are changing our view of the origins, limits, and capabilities of life. Unique microbial biochemistries amassed over eons in every niche on the planet now offer a deep and virtually limitless resource that can be applied to help enable biobased solutions for national needs. GTL research will reveal processes by which microbes produce energy, including ethanol and hydrogen, and other capabilities that may be used to clean up environmental contaminants and control the cycling of carbon.

Elucidating the design principles of microbial systems in their diverse environments entails analyses of unprecedented scale and complexity. DOE-relevant microbial systems can have millions of genes and thousands of genetic and regulatory processes and community interactions that underlie diversity and adaptability. Achieving GTL goals requires a major advance in our ability to measure the phenomenology of living systems and to incorporate their operating principles into computational models and simulations that accurately represent biological systems—the ultimate level of integrated understanding generated by GTL research.

To make GTL science and biological research more broadly tractable, timely, and affordable, GTL will develop four user facilities, delivering economies of scale and enhanced performance. These facilities will provide the advanced technologies and state-of-the-art computing needed to better understand microbial genomic potential, cellular responses, regulation, and community behaviors in any environment. Making such capabilities available to the broad research and technology-development communities will democratize access to forefront scientific resources and enlist an expanded community in exciting science for national needs.

Central to the success of the GTL program are computing and information technologies, which will allow us to surmount the barrier of complexity now preventing deduction of biological function directly from genome sequence. GTL will create an integrated computational environment linking experimental data of unprecedented quantity and dimensionality with theory, modeling, and simulation to uncover fundamental biological principles and to develop and test systems theory for biology.

This roadmap was developed from a process of broad community participation. It traces the path from DOE mission science through systems microbiology to the promise of emerging technologies, integrated computing, and a new research infrastructure. It describes opportunities, research strategies, and solutions at the nexus of the challenges of this new science as applied to microbes and the complexities of mission problems.

This Document: A Roadmap for the Future of GTL and Systems Microbiology

Roadmaps are pathways to the future. By their nature, they are pulled by needs rather than pushed by technology. They should clearly establish and communicate those needs and expectations, and they can serve as a handshake among all parties—end users, policymakers, science and technology leaders, and scientists and technologists. Roadmaps provide a basis for planning and coordination, allocation of resources, organization, and setting of strategy and priorities. They are the foundation of a creative and energetic venue for scientists and technologists to pursue the frontiers aggressively while collaborating on achieving higher goals. This roadmap describes details of a three-phase implementation of the Genomics:GTL program (see Table 1. GTL Science, Technologies, and Applications Roadmap, p. 11).

This 2005 roadmap builds on and expands the first GTL roadmap published in 2001 (www.doegenomestolife.org/roadmap/GTLcomplete_web.pdf). It traces connections among technical DOE mission objectives and science needs and the GTL goals and milestones in biological research, technology, and infrastructure, including four world-class user facilities. This roadmap is the result of 3 years of collaboration among hundreds of scientists and technologists via a number of workshops and other activities covering all relevant aspects of science, Department of Energy (DOE) missions, technologies, and computing (see Appendix D. GTL Meetings, Workshops, and Participating Institutions, p. 239). Drawing heavily on the output and insights from these discussions, the roadmap presents a baseline for the science, technologies, computing, and research facilities. It is meant to begin the dialogue that will determine their ultimate functionality and form. A vigorous process to refine these ideas and incorporate progress and revolutions as they occur will be central to implementation of this plan.

The GTL roadmap is grounded in DOE missions. First, "GTL Roadmap Strategy" connects the tremendous promise of 21st Century biology to the needs of the nation. Genomics, systems biology, the amazing world of microbes, computing, and the creation of major facilities to provide a new biology venue are discussed. The three phases of the GTL program set a timeline and logical construct for all that follows. Second, "Missions Overview" explains the ultimate focus of GTL research, laying out the technological objectives of energy production, environmental remediation, and carbon cycling and sequestration. Outlining the ways bioscience can support application progress in these areas, it presents a high-level science roadmap for addressing mission challenges. The "GTL Research Program" chapter states the overarching science goal, mission science goals, and four science and technology milestones that, when achieved, will provide the intellectual and technical basis for microbial systems biology and a tractable strategy for solving exceedingly complex mission problems. Highlights of ongoing research related to individual milestones are presented. These concepts and technologies will be integrated and scaled up in four research facilities that will serve as an engine of discovery and innovation for the GTL research program. This chapter also provides a discussion of governance, training, and ethical, legal, and social implications and issues.

"Creating an Integrated Computational Environment for Biology" discusses the central role of computing in this endeavor. It describes how modeling and simulation, data and data analyses, theory, community access, and a computational infrastructure (a roadmap for each is described) can come together to serve as the "central nervous system" of GTL research projects and facilities.

"GTL Facilities" contains an overview and sections describing on several levels each of the four user facilities, expected to achieve unprecedented levels of performance, throughput, efficiency, quality, and cost-effectiveness. They are the Facility for Production and Characterization of Proteins and Molecular Tags; Facility for Characterization and Imaging of Molecular Machines; Facility for Whole Proteome Analysis; and Facility for Modeling and Analysis of Cellular Systems. The DOE Office of Science includes all four in its 2003 *Facilities for the Future of Science: A Twenty-Year Outlook* (www.science.doe.gov/Sub/Facilities_for_future/facilities_future.htm). Each section discusses the particular facility's science and technology drivers and rationale, components and functions, and relevant technologies. Roadmaps for each facility explain how development of an

appropriate mix of technologies and computing needs will tie its components together, integrating each facility into the GTL research enterprise. The "GTL Development Summary" chapter describes global, crosscutting, and long-lead management and technological issues that must be addressed in a comprehensive way to achieve the best overall science, technology, and impact.

Three appendices on DOE missions (Energy, Environmental Remediation, and Carbon Cycling and Sequestration) present detailed discussions of mission problems, the vision for the future with existing gaps and necessary science foundation, and research strategies to meet those challenges. Other appendices provide more details on program background, relationships, and research projects, as well as references and a glossary.

Contents

Preface iii
Executive Summary v
This Document: A Roadmap for the Future of GTL and Systems Microbiology vii

1.0. Genomics:GTL Roadmap Strategy 1
1.1. Industrial Biotechnology and DOE Missions 2
1.2. Genomics and Systems Biology 4
1.3. Genomes to Life: Achieving a Predictive Understanding of Microbial Function 5
1.3.1. GTL User Facilities: Performance, Throughput, and Cost 6
1.3.2. Computing and Information Science 8
1.3.3. Power of the GTL Knowledgebase: Economies Provided by Nature 8
1.3.4. Bridging the Gap Between Big and Small Science—The Need for a Third Model 9
1.3.5. Department of Energy: Experienced at Large-Scale Projects 9
1.3.6. Implementation: What is the Time Frame, Who will be Involved, and How will Decisions be Made? 10

The Microbial World: A Challenging Frontier 13

2.0. Missions Overview: The Role of Microbial Systems in Energy Production, Environmental Remediation, and Carbon Cycling and Sequestration 21
2.1. Introduction to GTL Goals for DOE Missions 22
2.2. GTL Research Analyzing Mission-Relevant Systems 22
2.2.1. Engineered Systems 23
2.2.2. Natural Ecosystems 23
2.2.3. Shortening the Missions Technology Cycle 23
2.3. Basic Energy Research: Develop Biofuels as a Major Secure Energy Source 24
2.3.1. Ethanol Production from Cellulose 24
2.3.2. Biophotolytic Hydrogen Production 29
2.4. Environmental Remediation: Develop Biological Solutions for Intractable Environmental Problems 31
2.4.1. Science and Technology Objectives 32
2.5. Microbial Roles in Carbon Cycling and Sequestration: Understand Biosystems' Climate Impacts and Assess Sequestration Strategies 33
2.5.1. Science and Technology Objectives 35
2.5.2. Marine Microbial Communities 35
2.5.3. Terrestrial Microbial Communities 36
2.6. Summing Up the Challenges 37

3.0. GTL Research Program 41
3.1. Background and Approach 42
3.1.1. Phase I Implementation: Current GTL and Related Projects 43
3.2. Scientific Goals and Milestones 43
3.2.1. Missions Science Goals 44
3.2.2. Science and Technology Milestones 44

3.3. Highlights of Research in Progress to Accomplish Milestones 55
3.3.1. Research Highlights for Milestone 1: Sequences, Proteins, Molecular Complexes 56
3.3.2. Research Highlights for Milestone 2: Cell and Community Function, Regulation, and Dynamics 56
3.3.3. Research Highlights for Milestone 3: Computing 58
3.3.4. Sidebars Illustrating Details of Specific Research 58
3.4. GTL Program and Facility Governance 77
3.4.1. Facility User Access 77
3.4.2. Collaborative Environment 77
3.4.3. Facility Governance 78
3.5. Training 78
3.6. Ethical, Legal, and Social Issues (ELSI) 79
3.6.1. GTL Commitment to Explore ELSI Impacts 79
3.6.2. The Path Forward 80

4.0. Creating an Integrated Computational Environment for Biology 81
4.1. An Essential Foundation 82
4.2. Capabilities for an Integrated Computational Environment 85
4.2.1. Theory, Modeling, and Simulation Coupled to Experimentation of Complex Biological Systems 85
4.2.2. Sample and Experimental Tracking and Documentation: Laboratory Information Management System (LIMS) and Workflow Management 91
4.2.3. Data Capture and Archiving 92
4.2.4. Data Analysis and Reduction 93
4.2.5. Computing and Information Infrastructure 96
4.2.6. Community Access to Data and Resources 97
4.2.7. Development Requirements 99

5. GTL Facilities 101

5.0. Facilities Overview 101
5.0.1. Science and Technology Rationale 102
5.0.2. A New Trajectory for Biology 104
5.0.3. Capsule Facility Descriptions 104
5.0.4. Relationships and Interdependencies of Facilities 106
5.0.5. Research Scenarios 107
5.0.6. Facility Development 107

5.1. Facility for Production and Characterization of Proteins and Molecular Tags 111
5.1.1. Scientific and Technological Rationale 112
5.1.2. Facility Description 116
5.1.3. Development of Methods for Protein Production 118
5.1.4. Development of Methods for Protein Characterization 123
5.1.5. Development of Approaches for Affinity-Reagent Production 126
5.1.6. Development of Data Management and Computation Capabilities 131
5.1.7. Facility Workflow Process 131

5.2. Facility for Characterization and Imaging of Molecular Machines 139
5.2.1. Scientific and Technological Rationale 140
5.2.2. Facility Description 141
5.2.3. Technology Development for Expression, Isolation, and Purification of Molecular Machines 143
5.2.4. Technology Development for Identification and Characterization of Molecular Machines 144
5.2.5. Technology Development for Biophysical Characterization 149
5.2.6. Development of Computational and Bioinformatics Tools 153

5.3. Facility for Whole Proteome Analysis 155
5.3.1. Scientific and Technological Rationale 156
5.3.2. Facility Description 158
5.3.3. Technology Development for Controlled Microbial Cultivation and Sample Processing 159
5.3.4. Large-Scale Analytical Molecular Profiling: Crosscutting Development Needs 162
5.3.5. Technology Development for Transcriptome Analysis 162
5.3.6. Technology Development for Proteomics 164
5.3.7. Technology Development for Metabolomics 167
5.3.8. Technology Development for Other Molecular Analyses 169
5.3.9. Development of Computational Resources and Capabilities 169

5.4. Facility for Analysis and Modeling of Cellular Systems 173
5.4.1. Scientific and Technological Rationale 174
5.4.2. Facility Description 178
5.4.3. Technology Development for Cultivation of Microbial Communities 179
5.4.4. Development of Genomic Capabilities 181
5.4.5. Technology Development for Imaging and Spectroscopy 181
5.4.6. Development of Computing Capabilities 187

6.0. GTL Development Summary: Global, Crosscutting, and Long-Lead Issues 191
6.1. Coordinated GTL Program and Facility Development 192
6.2. Biology Drivers and Issues 192
6.2.1. Recalcitrant Proteins and Complexes 192
6.2.2. Biosample Growth and Culturing 192
6.2.3. Affinity Reagent Libraries 193
6.2.4. Characterization of Proteins and Complexes 193
6.3. Technology Drivers and Issues 193
6.3.1. Technologies for Measurement of Proteins, Metabolites, and Molecular Machines 193
6.3.2. MEMS, Microfluidics, and Nanotechnology 193
6.3.3. Single-Cell Analysis 194
6.3.4. Imaging 194
6.3.5. Data-Quality Standards 194
6.4. Computing, Communications, and Information Drivers and Issues 194
6.4.1. Computational Methods for Experimental Data Analysis 194
6.4.2. Process Control, LIMS, Workflow Management 194
6.4.3. Data Architecture, Modeling, and Integration 195
6.4.4. Computing Hardware and Networking Infrastructure 195

6.4.5. Computational Models for Establishing Networks and Simulations 195
6.4.6. Genome Annotation 195
6.4.7. Computing, Communications, and Information 196
6.5. Other Issues 196
6.5.1. Ethical, Legal, and Social Issues (ELSI) 196
6.5.2. Technology Transfer 196
6.5.3. Industrial Involvement 196

Appendix A. DOE Mission: Energy Security 197
A.1.1. The Energy Challenge 198
A.1.2. The Role of Biology and Biotechnology in America's Energy Future 199
A.1.3. GTL's Vision for Biological Energy Alternatives 201
A.1.4. Ethanol from Biomass 203
A.1.4.1. Cellulose Degradation and Conversion 203
A.1.4.2. Bioethanol Research Targets for GTL 205
A.1.5. Biohydrogen Production 208
A.1.5.1. Biophotolysis of Water 209
A.1.5.2. Biohydrogen Research Targets for GTL 209
A.1.6. Summary 214

Appendix B. DOE Mission: Environmental Remediation 215
B.1.1. Environmental Remediation Challenge 216
B.1.2. The Role of Microbial Systems in Remediation 217
B.1.2.1. Benefits and Impacts 218
B.1.2.2. Establishing the Link Between Biology and Geochemistry 218
B.1.3. Using Genome Sequences as a Launch Point to Understand Communities 218
B.1.3.1. Modeling Microbial Metabolic Activities 220
B.1.3.2. Merging Metabolic and Field-Scale Models 220
B.1.4. GTL's Vision for Environmental Remediation and Restoration 221
B.1.4.1. Gaps in Scientific Understanding 221
B.1.4.2. Scientific and Technological Capabilities Required to Achieve Milestones 222

Appendix C. DOE Mission: Carbon Cycling and Sequestration 227
C.1.1. The Climate Change Challenge 228
C.1.2. The Role of Microbes 229
C.1.3. Microbial Ocean Communities 231
C.1.3.1. Photosynthetic Capabilities 231
C.1.3.2. Strategies for Increasing Ocean CO_2 Pools 231
C.1.3.3. GTL's Vision for Ocean Systems 233
C.1.4. Terrestrial Microbial Communities 235
C.1.4.1. Influence on Plant Growth 235
C.1.4.2. Strategies for Increasing Stable Carbon Inventories 236
C.1.4.3. Terrestrial Systems Vision 236

CONTENTS

Appendix D. GTL Meetings, Workshops, and Participating Institutions 239

Appendix E. GTL-Funded Projects 245

Appendix F. Strategic Planning for CCSP and CCTP 249

Appendix G. Microbial Genomes Sequenced or in Process by DOE 253

Appendix H. Programs Complementary to GTL Research 265

References 275

Glossary 281

1.0. Genomics:GTL Roadmap Strategy

1.1. Industrial Biotechnology and DOE Missions 2

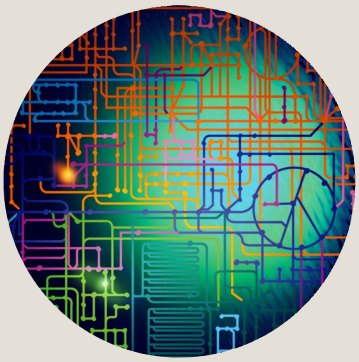
1.2. Genomics and Systems Biology 4

1.3. Genomes to Life: Achieving a Predictive Understanding of Microbial Function 5

1.3.1. GTL User Facilities: Performance, Throughput, and Cost 6
1.3.2. Computing and Information Science 8
1.3.3. Power of the GTL Knowledgebase: Economies Provided by Nature 8
1.3.4. Bridging the Gap Between Big and Small Science—The Need for a Third Model 9
1.3.5. Department of Energy: Experienced at Large-Scale Projects 9
1.3.6. Implementation: What is the Time Frame, Who will be Involved, and How will Decisions be Made? 10
1.3.6.1. Three-Phase Implementation of the GTL Program 10
1.3.6.2. Integrated Management and Development 11
1.3.6.2.1. GTL Program and Facility Governance 12
1.3.6.2.2. Facility Development and Acquisition Process 12
1.3.6.2.3. Community Involvement in GTL Technology Development 12
1.3.6.2.4. Communication in a Multidisciplinary Environment 12
1.3.6.2.5. Facility User Access 12

The Microbial World: A Challenging Frontier 13

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, the Bioenergy Research Center, will concentrate on overcoming biological barriers to the industrial production of biofuels from biomass and on other potential energy sources such as biophotolytic decomposition of water to hydrogen. 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://doegenomestolife.org/biofuels/).



Genomics:GTL Roadmap Strategy

GTL aims to understand biological systems well enough to predict their behavior accurately with mechanistic computational models. The ultimate result will be the ability to use the biochemical sophistication of microbes for a broad range of innovative applications, a potential achievable only with huge gains in research performance, productivity, efficiency, cost-effectiveness, and quality.

GTL Hallmarks

- Global, genome-wide view of microbial functions
- Advanced technologies with improved performance to provide comprehensive data sets
- Facilities to dramatically improve throughput, cost, and data quality
- Comparative analysis at all levels enabled by the GTL Knowledgebase (genes → molecules → cell processes → cell and community function)
- Modeling and simulation tools for predictive understanding to enable in silico biology
- Open access to data, protocols, and facilities

*Biotechnology includes applications beyond clinical medicine to those in agriculture and, more recently, industry and the environment. For additional information, see the Biotechnology Industry Organization's web site (www.bio.org).

1.1. Industrial Biotechnology and DOE Missions

Industrial biotechnology* is critical for the future of the nation and will be an engine of economic competitiveness in this century. Biological solutions must contribute to the mix of technologies needed for the key DOE missions of energy security, environmental restoration, and climate change. These missions, shown in Fig. 1. GTL Science and Technology Foundations for DOE Missions, this page, are grand challenges for the development of bioscience and biotechnology, with the potential for trillions of dollars of impact. DOE missions require that we understand biology at every level, from the most detailed molecular processes to vast natural ecosystems. Such knowledge will point the way for countless applications, leading to new industries and economic and social benefits (Brady, Chao, and Clardy 2002; Luengo 2003). Broad interest in the development of the necessary science base will accelerate the pace of discovery and increase the scope of its

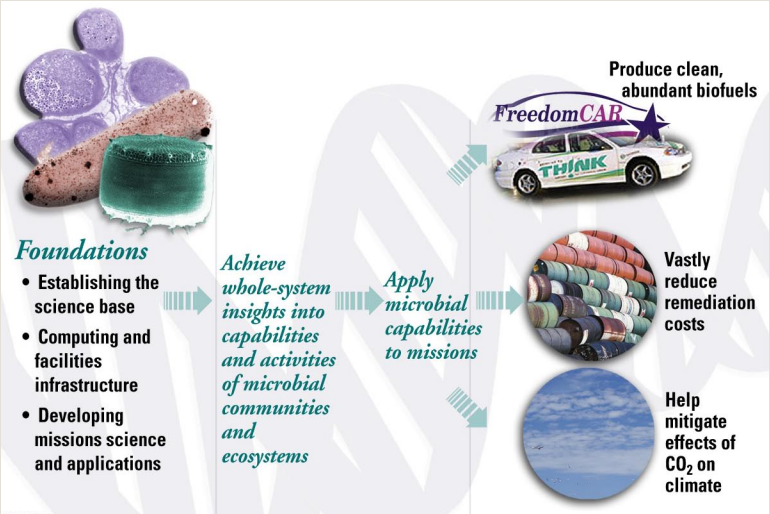


Fig. 1. GTL Science and Technology Foundations for DOE Missions.

impact. While the Human Genome Project opened the door for improvements in human health, GTL is the gateway to biotechnological solutions to DOE mission problems and to stimulating new generations of industrial biotechnology (Herrera 2004; Littlehales 2004). DOE's mission challenges are the following:

- Develop biofuels as a major secure energy source.
- Develop biological solutions for intractable environmental problems.
- Understand relationships between climate change and earth's microbial systems; assess options for sequestration.

See 2.0. Missions Overview, p. 21, and DOE mission appendices: A. DOE Mission: Energy Security, p. 197; B. DOE Mission: Environmental Remediation, p. 215; and C. DOE Mission: Carbon Cycling and Sequestration, p. 227.

The National Research Council, in its 1999 report, *Biobased Industrial Products: Research and Commercialization Priorities*, acknowledged the great potential economic impact of industrial biotechnology, noting:

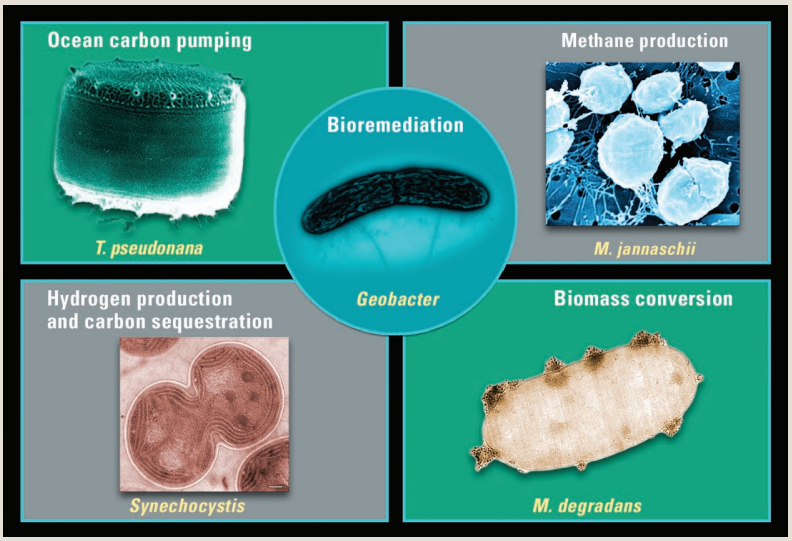
> "Biological sciences are likely to make the same impact on the formation of new industries in the next century as the physical and chemical sciences have had on industrial development throughout the century now coming to a close."

The Economist also championed the promise of biotechnology:

> "At the moment, biotech's main uses are in medicine and agriculture. But its biggest long-term impact may be industrial [Editorial, "Saving the World in Comfort" (March 27, 2003)]."

Microbes can provide the basis for a revolution in industrial biotechnology. These untapped natural treasures are the foundation of the biosphere and sustain all life on earth. Extreme genetic diversity and the ability to function in complex communities give microbes extraordinary biochemical capabilities and adaptability (see Fig. 2. Using the Natural Diversity of Microbes to Create Biotechnology Solutions for DOE Missions, this page). These single-celled organisms are masters at living in almost every environment and harvesting energy in almost any form, from solar radiation to mineral chemistry, and transforming it into chemical compounds that power life. By understanding how microbes function in their many environments, we can reveal their contributions to earth ecosystems and their relationships to climate change. We also can understand how they can provide the basis for environmental remediation and for creating new sources of renewable, less-polluting energy sources and new generations of processes for industrial application (see Fig. 1, p. 2, and sidebar section, The Microbial World, beginning on p. 13).

Fig. 2. Using the Natural Diversity of Microbes to Create Biotechnology Solutions for DOE Missions. Microbes, which have been adapting to countless environments for some 3.5 billion years, offer an untapped reservoir of sophisticated chemistries. Harnessing their functionalities can result in novel and highly efficient strategies for producing hydrogen and other fuels, cleaning up toxic waste at contaminated DOE sites, and capturing and storing carbon from the atmosphere to help mitigate global climate change.



T. pseudonana: B. Palenik and D. Lee, Scripps Inst. of Oceanography; *M. degradans*: R. Weiner, Univ. of Md., College Park; *M. jannaschii*: B. Boonyaratanakornkit, D. S. Clark, G. Vrdoljak, Univ. of Calif., Berkeley

1.2. Genomics and Systems Biology

Every organism's genome encodes its ability to create and sustain life. Genome data provide the foundation for studying biological processes rather than examining isolated parts. The longstanding successful approach to biological research—variously described as "single gene," "reductionist," or "linear"—is piecemeal and, while productive, is less efficient at expeditiously addressing questions of biological complexity and integration.

The new approach for exploration—systems biology—will allow us to envision the microbe as a complete set of intersecting processes and to create models for simulating how microbes operate and respond. This is a major first step toward illuminating the most fundamental principles of living cells and achieving a predictive understanding of the scale and complexity of natural systems.

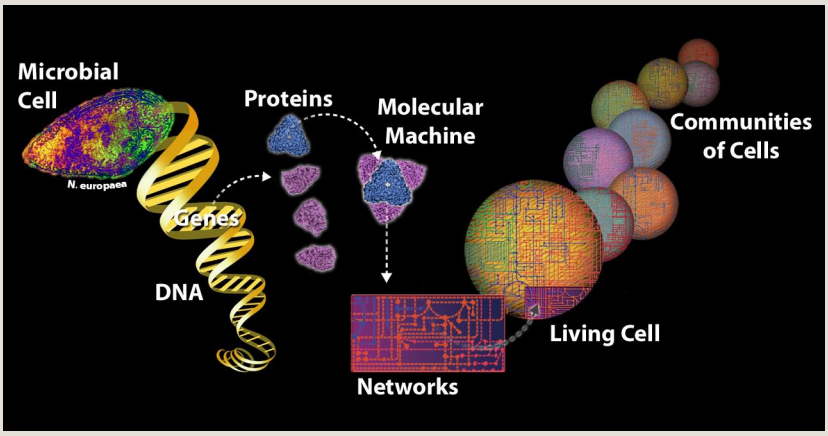
A comprehensive approach to understanding biology encompasses many cellular components. While a genome is a fixed catalogue of information, it dynamically creates the machinery of a cell in response to the changing environment. Thousands of genes encode even greater numbers of proteins that mediate biology in a "just-in-time" strategy by associating in myriad ways (protein "machines") to form intricate pathways and networks within a cell (see sidebar, The Basics: From DNA to Living Cells to Communities, this page). Demonstrating the power of these finely tuned systems, microbes rapidly respond to environmental cues by adjusting their entire cellular operation (O'Toole 2003). (See 3.2.2. Science and Technology Milestones, p. 44, and 5.0. Facilities Overview, p. 101.)

Microbes, in their adaptability, also have the best of both the single- and multicellular worlds. Using mechanisms we are only beginning to understand, microbes carry on a dialogue that establishes community and environmental awareness and enables individual microbes to function together as multicellular organisms (e.g., "biofilms") in complex geochemical environments, with the many benefits that can provide. Functions include sensing the environment; assembling appropriate cells or communities of cells as environmental conditions change; regulating and carrying out cellular function, including critical energy capture and manipulation; and providing for reproduction, sporulation, or senescence as conditions dictate (Check 2002; see sidebar, Life in a Biofilm, p. 18).

We can now rapidly and accurately decode the genomes of microbes and microbial communities in complex natural ecosystems (metagenomes). While more than 99% of microbes historically have been hidden from study because they could not be cultured (Handelsman et al. 1998), genomics allows the assessment of these microbial systems to determine who's there and what some of their functionalities are. Other emerging technologies such as imaging will

The Basics: From DNA to Living Cells to Communities

- **Cells** contain DNA—the hereditary material of all living systems.
- The **genome** is an organism's complete set of DNA.
- **DNA** contains genes whose sequences specify how and when to build proteins.
- **Proteins** perform most essential life functions, often working together as molecular machines. In addition, they form most cell structures.
- **Molecular machines** interact through complex, interconnected pathways and **networks** to make the working cell come alive.
- **Communities of cells** are associations of microbes (each a single cell) working together in a particular environmental niche.



N. europaea: M. A. Bruns, Center for Microbial Ecology, Michigan State University

allow us to track molecules and cells in complex living systems to add a functional perspective without the need for classic culturing.

Not only has genomics enabled early insight into these complex systems, it also has revealed a vastly greater microbial diversity and presence than was previously appreciated (Stein et al. 1996; Meyer 2004; Schaechter, Kolter, and Maloy 2005). Recent genomic studies of microbial communities already have led to the discovery of millions of genes and proteins, thousands of species, and innumerable variations in critical functionalities (Venter et al. 2004), establishing the globe's vast microbial communities as a potentially rich resource for understanding biology and for catalyzing industrial biotechnology (Schloss and Handelsman 2003; Riesenfeld et al. 2004).

Functions of about 40% of sequenced genes, however, remain unknown or poorly characterized—a challenge that will be prevalent in most DOE-relevant systems. As a further objective, this program aims to achieve a holistic, mechanistic understanding of biology. Genes provide life's potential list of components, but complete mechanisms of function can be revealed only after additional experiments and analyses are performed. The research community (Buckley 2004b; Roberts 2004) recognizes that we need new approaches for studying microbes and other systems efficiently (Aebersold and Watts 2002; Roberts et al. 2004) to understand gene and systems function.

1.3. Genomes to Life: Achieving a Predictive Understanding of Microbial Function

GTL aims to understand biological systems well enough to predict their behavior accurately with sophisticated computational models. GTL analyzes critical microbial properties and processes on three fundamental systems levels (see 3.0. GTL Research Program, p. 41, and 4.0. Creating an Integrated Computational Environment for Biology, p. 81).

- **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:** Exploring how diverse microbes interact to carry out coordinated complex processes enabling microbes to both respond to and alter their environments.

Conceptually, genomes contain all the information needed to deduce function, yet the intricately detailed biology underlying life creates a huge barrier to a facile connection between genome sequence and function. The GTL strategy is to provide the technologies, computing infrastructure, and comprehensive knowledge-base to break through the barrier of complexity that prevents the direct translation of genome sequences into predictive understanding of function. This new opportunity grows out of rapid advances in instrumentation for the biosciences, exponential improvements in computing speeds and modeling capabilities, and a growing interest by physical and information scientists in applying these methods to biological problems. The sequences furnish the blueprint for exploration, technologies can produce the data, and computing can relate these enormous data sets to models of process and function (Ellis et al. 2004; Kitano 2002).

Understanding biological systems at these three levels, however, is a daunting task: While the genome of a microbe is a fixed code that represents megabytes of information, a full description of all the dynamic processes involved in making a living cell operate and respond to its environment will be a mixture of complex data sets potentially in the petabyte range (i.e., a billion megabytes) (see 5.3. Facility for Whole Proteome Analysis, Table 1. GTL Data, p. 159). Gathering these data will require huge gains in performance, data quality, productivity, and cost-efficiency. New generations of computing and information capabilities thus are needed to manage, analyze, and transform the information into accurate models. In addition, for biology to have timely impacts on national needs, the time required to achieve fundamental understanding of a system must be accelerated from many years to months, compounding the challenge (see The Framework for DOE Missions, p. 24).

GTL ROADMAP STRATEGY

A national priority is to achieve an unprecedented understanding of the natural world:

> "The application of DNA sequence and other data allows the development of new biotechnological tools, such as microarrays to decipher the functional implications of gene expression, which are helping to unravel longstanding questions in biology. Agencies should target investments toward the development of a deeper understanding of complex biological systems through collaborations among physical, computational, behavioral, social, and biological researchers and engineers. These collaborations will yield new ways of collecting and analyzing data allowing for the exploration of the living world across all levels of biological organization both spatially and temporally. . . . Federal agencies should continue to invest in obtaining additional sequence data and in the development of genomics tools and resources (Marburger and Bolton 2004, www.ostp.gov/html/m04-23.pdf)."

GTL aims to bring biologists, physical scientists, and computing scientists together to establish a new biology for meeting energy, environmental, and climate needs (Frazier et al. 2003a; 2003b).

1.3.1. GTL User Facilities: Performance, Throughput, and Cost

Attaining the ability of the scientific community to analyze microbial systems on a timetable that supports applications in years rather than decades requires that we begin generating and assimilating materials and data on a scale that far exceeds today's capacities. Genome-sequencing projects have shown that pursuing transformational production goals in dedicated facilities can result in such gains (see sidebar, High-Throughput Model Guides Future Facilities, this page). Consequently, to advance its research goals, DOE plans to develop four cost-effective, high-throughput user facilities for systems microbiology research, in addition to supporting a broad range of research projects in GTL. Achieving economies of scale, the GTL suite of four facilities is designed to provide a phased and integrated set of analytical and production capabilities to determine microbiological structure and function from the genomic through the ecosystem levels. This is depicted in the sidebar, GTL Facilities: Accelerating Scientific Discovery and Applications Research for Energy and Environment, p. 7, and in the Facilities Overview and subsequent descriptions beginning on p. 101.

Each of these facilities is important in its own right, but all are intricately linked in their long-term goals, targets, technologies, capabilities, and capacities. They will provide production and analytical resources for scientists to collect and use the information needed to put microbes and their capabilities to work. While vital for GTL progress, the facilities also will help accelerate biological research supported by other agencies.

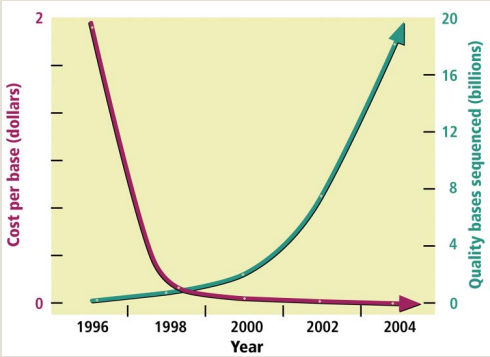
High-Throughput Model Guides Future Facilities

The dramatically increased productivity and reduced costs achieved in the Human Genome Project via high-throughput production environments (e.g., the DOE Joint Genome Institute) provide the paradigm for dedicated industrial-scale facilities envisioned for GTL systems biology research. These resources democratize cutting-edge science, enabling even the smallest research laboratory to participate.

A growing mandate from the scientific community echoes the need for systems biology facilities:

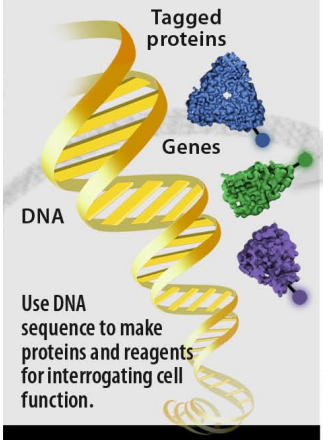
"To make progress, science should not accept the limitations placed on discovery by traditional methods, conventional approaches, or existing infrastructure. Powerful, but expensive, modern equipment should be housed in community facilities, open to researchers who might not otherwise have access to these technologies." [Source: *Microbiology in the 21st Century: Where Are We and Where Are We Going?* American Society for Microbiology (2004)]

Large-Scale Genome Sequencing Facilities Spur Cost, Productivity Improvements



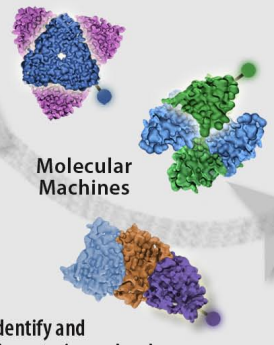
GTL Facilities: Accelerating Scientific Discovery and Applications Research for Energy and Environment

Protein Production and Characterization



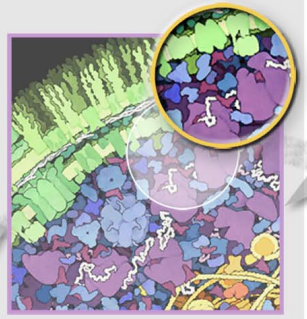
Use DNA sequence to make proteins and reagents for interrogating cell function.

Molecular Machines



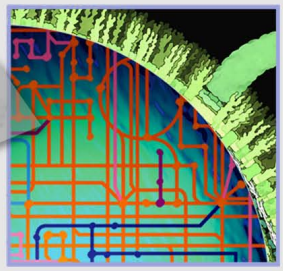
Identify and characterize molecular complexes and other interactions.

Proteomics



Identify proteins and other molecules produced by cells in response to environmental cues.

Cellular Systems



Achieve an in silico, predictive understanding of microbes in their natural environments.

Production and Characterization of Proteins and Molecular Tags

- Produce proteins encoded in the genome.
- Create affinity reagents that allow each protein to be identified, located, and manipulated in living cells.
- Perform biophysical and biochemical characterizations of proteins produced to gain insights into function.

Characterization and Imaging of Molecular Machines

- Isolate and analyze molecular machines from microbial cells.
- Image structure and cellular location of molecular machines.
- Generate dynamic models and simulations of molecular machines.

Whole Proteome Analysis

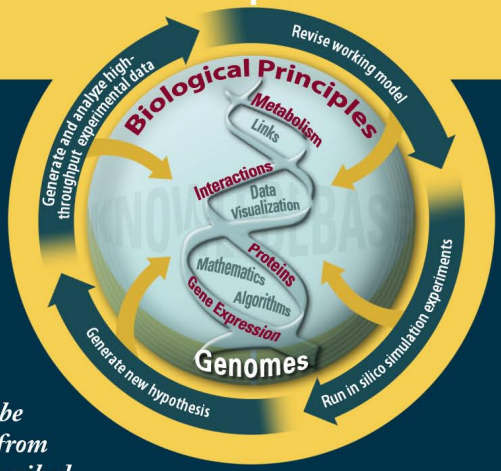
- Measure molecular profiles and their temporal relationships.
- Identify and model key pathways and other processes to gain insights into functions of cellular systems.

Modeling and Analysis of Cellular Systems

- Integrate knowledge and models to understand the structure and functions of cellular systems, from single cells to complex communities.
- Integrate imaging and other technologies to analyze molecular species from subcellular to ecosystem levels as they perform their functions.

Understanding how the information in a genome dictates cellular functions requires knowledge of a cell's molecular complement, interactions, and regulation. These studies must be carried out on a scale far exceeding today's capacities.

Microbial genome sequences will be the foundation on which all data from the large-scale GTL facilities (described above) are related.



- Comprehensive integration of GTL and research community databases
- Transparent and intuitive access to computational tools
- Simulations of microbial behavior using genome sequences as input
- Information and support for research, policy, and applications

Systems Microbiology Knowledgebase to Enable a Predictive Understanding of Microbes and Communities

YGG 04-0172R4

1.3.2. Computing and Information Science

GTL is as much a computing program as a biology program. Computational modeling is at the heart of the research described in this roadmap. The new biology must tightly integrate computational analysis and experimental characterization of biological systems and combine linked measurements using dozens of sophisticated new analytical instruments to create huge multivariate data sets and relate them to models at the appropriate systems level.

The vision for the GTL computing environment encompasses the creation of a seamless enterprise with support mechanisms for collaboration via transparent scientific-community access to data and tools. Features include the following.

- **Open access to data.** Furnish easily accessible, large-scale data archives and community databases containing enabling data, knowledge, and models of biological systems.
- **Open access to tools.** Provide powerful suites of analysis, data-mining, modeling, and simulation tools that enable GTL facilities, projects, and end users to interpret, understand, and predict the behavior of biological systems.
- **Accessible infrastructure**. Provide computing hardware, operating software, data storage, and network capabilities to support large-scale systems biology conducted by a diverse research community.

The GTL vision puts biology on the same path that much of research and industry have followed, with computation, modeling, and simulation as an integral part of the research process. As the performance of computing increases and costs decline, modeling and simulation are critical to experimentation that is becoming ever more complex, time consuming, and costly. Computation provides the insights needed to make experimentation more focused and informative. Transforming biology into a quantitative and predictive science based on models and data will accelerate discovery and ultimately shorten the technology-development cycle, more rapidly yielding practical national benefits (see 3.2.2.3. Milestone 3, p. 51; 4.0. Computing, p. 81; and computing roadmaps for each facility in 5.0. Facilities Overview, p. 101).

1.3.3. Power of the GTL Knowledgebase: Economies Provided by Nature

Analysis of each new microbe benefits from knowledge about all other microbes and life forms because of nature's simplifying principles. 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 rules. Once successful machines, pathways, and networks arise, they tend to be preserved, subtly modified and optimized, and then reused as variations on enduring themes throughout many species. Thus, accumulating detailed information on numerous microbes across a wide range of functionalities will provide the insight needed to interpret these principles. To take full advantage of this phenomenon, relating all known information generated computationally or experimentally to the genome (annotation) is an important task that must be performed continually as new genomes are sequenced or new experiments performed.

In this new era of systems biology, all-against-all comparisons of extensive microbial data amassed in the GTL Knowledgebase will accelerate and sharpen our research strategies. Along with high-throughput facilities and computing, this capability is a key element of our approach to reducing the analysis time for a microbial system. 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. Comparative genomics, founded on these principles, ultimately will allow us to predict the functions of unknown microbes by deriving a working model of a cell from its genetic code.

This paradigm combines (1) discovery science as we navigate huge unexplored data sets that can reveal unforeseen properties and phenomena and (2) computationally driven hypothesis science to derive insights into previously unfathomable complexity (see conceptual diagram of GTL Knowledgebase in sidebar at bottom of p. 7; 3.2.2.3.2. GTL Knowledgebase, p. 52; and 4.2.1.Theory, Modeling, and Simulation Coupled to Experimentation of Complex Biological Systems, p. 85).

1.3.4. Bridging the Gap Between Big and Small Science—The Need for a Third Model

The biology community faces a complex, yet critical, challenge: Preserve the creativity and entrepreneurial spirit of the single investigator in light of the increasingly sophisticated and costly resource requirements of leading-edge biological research. Making the most advanced technologies and computing resources available to the research community will democratize access to the tools needed for systems biology. The GTL research program, its integrated computational environment, and user facilities are designed for DOE science needs but also will help bridge the capability gap between large and small labs (Relman and Strauss 2000). Already, only a minority of even large laboratories can afford to be adequately equipped. Individual investigators need the capabilities of big science, and scientific progress requires diverse contributions from the whole scientific community. One benefit to DOE in providing these capabilities will be the involvement of a larger scientific community in the important mission challenges we face. Another benefit to the research community is the availability from industrial vendors of instrumentation and processes developed to meet GTL's requirements. A key point is that GTL facilities will not supplant or compete with investigator-initiated science but rather will complement and enhance it, just as high-throughput sequencing facilities (e.g., the DOE Joint Genome Institute) do today.

The facilities will open new avenues of inquiry, fundamentally changing the course of biological research and greatly accelerating the pace of discovery. Daunting technical, time, and cost barriers to a concerted and comprehensive approach to biology will be removed, allowing scientists to aspire to a higher-level perspective and higher-value research.

> "The ecosystem predictive capability will come from detailed work that links an understanding of the genome to an understanding of gene expression, protein function, and complex metabolic networks. We must create centers that facilitate research community access to postgenomic analytical capabilities. [*The Global Genome Question: Microbes as the Key to Understanding Evolution and Ecology*, American Society for Microbiology, 2004]"

1.3.5. Department of Energy: Experienced at Large-Scale Projects

The Department of Energy's Office of Science (SC), in addition to providing major facilities for the scientific community, conducts research on problems in fundamental science, energy, the environment, and climate that involve large interdisciplinary teams from many institutions pursuing strategic science goals. To make the necessary advances in biology over the coming decades, this research model will be critical to the comprehensive study of biology, ultimately impacting DOE mission problems (see sidebar, GTL User Facilities Leverage DOE Experience and Skills, this page).

Multidisciplinary Teams. Organized and focused teams are needed because meeting these great challenges will require many skills and capabilities in complementary and supportive roles to foster new thinking and approaches (Nass and Stillman 2003). The teams will take advantage of the unique convergence of disciplines that has occurred over the past decade;

GTL User Facilities Leverage DOE Experience and Skills

DOE is the leading funder of physical sciences in the nation. For more than half a century, its Office of Science (SC) has envisioned, designed, constructed, and operated many of the world's premier research facilities. These facilities continue to grow in importance to biology and today serve 20 times as many users from the life sciences community as in 1990.

DOE now seeks to bring its experience and skills in physical sciences to finding biological solutions to mission challenges. GTL facilities are among those featured in SC's 20-year plan (*Facilities for the Future of Science: A Twenty-Year Outlook*, 2003, www.science.doe.gov), with the director noting that "Investment in these [GTL] facilities will yield extraordinary scientific breakthroughs and vital societal and economic benefits." This plan was developed through discussions with and assistance from the scientific community.

several fields have arrived simultaneously at the same scale and complexity levels from different directions. This meeting of biology, computing, computational chemistry, materials science, synthetic and analytical sciences, microtechnologies, and, most recently, nanoscience and nanotechnology allows us to explore the amazing nanoworld of microbes with rich and powerful probes and methods.

Through this strategic approach to research and in the proposed GTL facilities, DOE's strengths in the biological, physical, and computational sciences will bring breakthrough technologies to bear on biology. GTL program sponsors, the Office of Biological and Environmental Research (BER) and the Office of Advanced Scientific Computing Research (OASCR), have made this commitment.

DOE supports a wide range of applied research and technology development at the national laboratories and in academia through its offices of Energy Efficiency and Renewable Energy (EERE), Fossil Energy (FE), and Nuclear Energy (NE). Through advanced technology development and commercialization programs, DOE works with industry to bring new tools and processes to the marketplace. GTL will be coordinated with these programs and will provide its research resources, supporting facilities, and computing infrastructure to enable the timely migration of new science and scientific capabilities into impactful technologies for DOE missions.

This coordinated approach is consistent with the DOE research tradition.

> "This capacity to deal with both the scale and complexity of these efforts is especially important in today's rapidly changing world. As we continue to gain the ability to work at very small scales and to probe the dynamic, three-dimensional structure of molecules, the interface between physical science and life science is of critical importance. Work at the interface of frontier disciplines like bioinformatics, genomics, proteomics, and nanotechnology is greatly enhanced by DOE's capacity in large-scale computation and research tools based upon physical science. [*Critical Choices: Science, Energy, and Security: Final Report of the Secretary of Energy Advisory Board's Task Force on the Future of Science Programs at the Department of Energy,* October 13, 2003]"

1.3.6. Implementation: What is the Time Frame, Who will be Involved, and How will Decisions be Made?

1.3.6.1. Three-Phase Implementation of the GTL Program

GTL is following a roadmap (see Table 1. GTL Science, Technologies, and Applications Roadmap, p. 11) to do the critical science to establish systems biology; to develop the necessary tools, resources, and facilities; and to complete the transformation to true systems biology aimed at mission applications. The strategy is divided into three distinct phases:

- **Phase I (First 8 years): Genomics to Systems Biology.** Transition from genomics to systems biology will include key proof-of-principle experiments in systems biology and technology prototyping and piloting that will create the science and technology base and begin to train a community of scientists in systems biology. The design, R&D, and early deployment of critical computing and information tools and infrastructure will provide the necessary foundation for data management, analysis, and modeling. Intense planning activities involving scientists, technologists, and mission programs will support the conceptualization, design, R&D, and construction of GTL production facilities.
- **Phase II (9 to 16 years): Technology Integration and Scaleup.** Production facilities will provide an engine for dramatically accelerating the study of microbial systems and for discovering and developing predictive systems understanding—ultimately reducing the time for analysis of a microbial system from years to months.
- **Phase III (16+ years): Biological Systems Knowledge for DOE Applications.** Knowledge and capabilities developed in GTL will be provided for rapid and complete systems studies of important problems in science and for useful applications. This knowledge and these capabilities will position GTL to rapidly transform new science into revolutionary new processes and products to help meet critical DOE national needs.

1.3.6.2. Integrated Management and Development

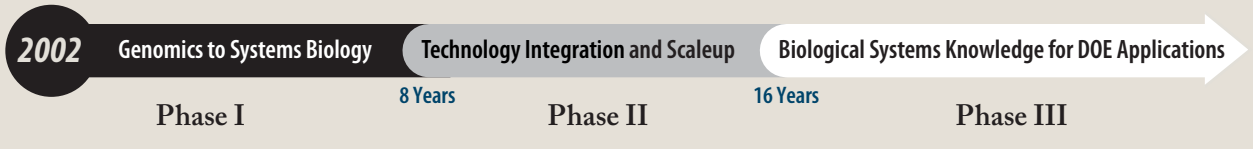
A concerted planning and management process will allow a more aggressive technical and scientific strategy and optimize the use of research program resources and R&D investments. Key issues include:

- Establishing program and facilities governance to ensure the best science and most useful technologies.
- Coordinating GTL program needs with facility design and development.
- Developing an integrated computational infrastructure for the facilities and program.
- Coordinating development activities across facilities and within the program to accommodate interdependencies and to optimize the use of resources.
- Establishing and coordinating workshops and working groups for solving problems and developing technologies.
- Understanding the needs of DOE mission applications groups to support use of GTL facilities for biotechnology development.
- Identifying promising technological areas that have global applicability or require long lead times.
- Building a communication strategy for public and community outreach and information gathering.
- Making GTL facilities accessible to the broader research community on a peer-reviewed basis.

Specific examples of development needs are outlined in 6.0. GTL Development Summary, p. 191.

Table 1. GTL Science, Technologies, and Applications Roadmap

Science Base		
Genomics to systems biology Molecular, cellular, and community studies Key insights and strategies for study	High-throughput study of key systems and processes Comparative analyses, systems modeling, and simulation Fully integrated systems for experimentation and computing	Integrated knowledgebase for designing mission solutions Science and technology development and application Science of innovation and next-generation concepts
Technologies, Computing, and Facilities		
Advanced technology development and testing Pilot studies, computing, and technology scaleup Facilities researched, developed, designed, and built	Facilities' operations: Comprehensive data Integrated data and computing capabilities operational Full biological systems data available in months	Facilities applied to engineered systems Tested, evaluated, monitored, and verified New functions and advanced concepts engineered
Applications of Missions Science		
Key systems chosen for mission interest Insights into cellular and community processes and interactions Systems data needs and strategies	Mission model systems analysis begun Understanding leading to engineering strategies for missions Application-specific strategies set	Fully engineered systems designed and developed Engineered systems tested and evaluated First generation fielded, next generation developed



1.3.6.2.1. GTL Program and Facility Governance

DOE will establish a governance process to ensure advancement of DOE, GTL, and research-community objectives; excellence in science; optimized facility access and operations; and continuous facility and equipment enhancement. Governance will include appropriate scientific and technological advisory groups and peer review for access and resource allocations (see 3.4. GTL Program and Facility Governance, p. 77, and 5.0. Facilities Overview, p. 101).

1.3.6.2.2. Facility Development and Acquisition Process

The GTL user facilities will be developed and acquired using a process based on the robust criteria described in DOE Order 413.3 (www.science.doe.gov/opa/PDF/O4133.pdf), which ensures the successful design, implementation, and management of DOE facilities. This construction-project management process includes a rigorous assessment of the science objectives, technology requirements and development needs, facility requirements, and staffing. The process encourages a sound technology-baseline definition and supports research, design, development, testing, and evaluation as needed to ensure that the facilities and all associated equipment and information systems perform at specifications. The scientific and technological community is involved extensively through workshops, working groups, and other mechanisms to refine objectives and designs and to ensure that new developments are incorporated into eventual facility technical capabilities and that scientific goals and functions take advantage of progress.

1.3.6.2.3. Community Involvement in GTL Technology Development

The GTL program and supporting facilities require extensive and continuing technology development. While GTL facilities will have a mission to keep capabilities fresh and relevant to emerging science and progress, a major element will be a distributed model for development of new methods and instruments that can be incorporated into the facilities. GTL will place key technology development where it can best take advantage of skills, capabilities, and infrastructure in universities, national laboratories, and industry.

1.3.6.2.4. Communication in a Multidisciplinary Environment

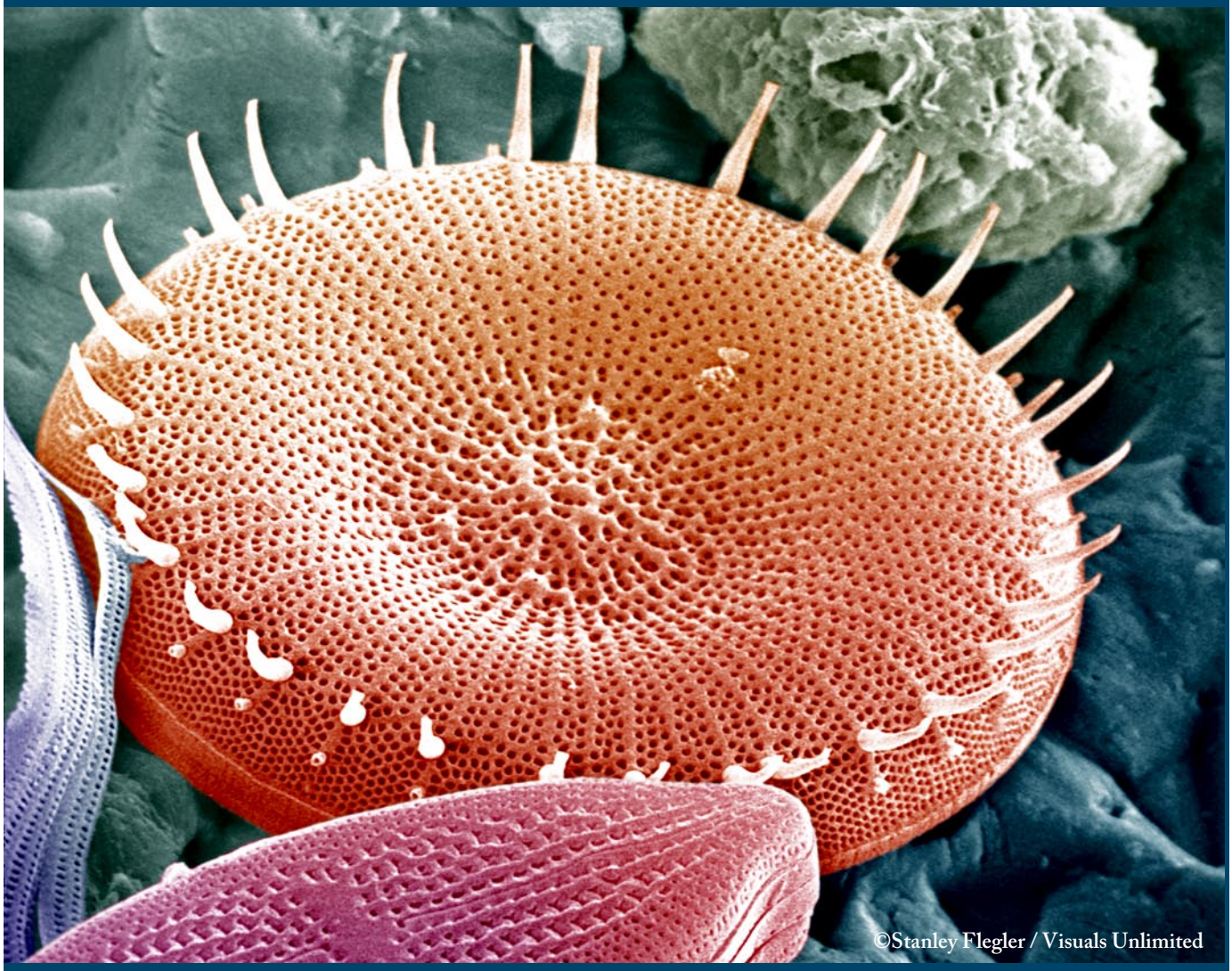
GTL communication strategies are to provide information to stimulate discourse and consensus, resulting in faster evolution of program design and content and more effective plans for achieving program aims. This communication will aid coordination and integration, particularly needed by a program requiring contributions from many disciplines, each with its own culture and vernacular. GTL also will foster cooperation between fundamental research and the development of technologies and applications. Communication will be prospective in conveying strategies and plans to generate new ideas. It will be retrospective in relating progress, results, and impacts. Effective communication, community involvement, and outreach will be prerequisites to establishing the GTL user facilities and a constituency to make best use of them. GTL will involve scientific societies and their ongoing meetings and other functions in these activities.

1.3.6.2.5. Facility User Access

GTL's dedicated user facilities will provide the broader scientific community with technologies, research resources, data and data-analysis tools, and computing and information infrastructure to perform systems microbiology studies (see 5.0. Facilities Overview, p. 101).

Access to GTL user facilities will be 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 science inventiveness, relevance to solving energy- and environmental-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. 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 missions.

The Microbial World: A Challenging Frontier



The physical diversity of microbes reflects a commensurate underlying genetic and functional diversity—yielding a broad range of biochemical capabilities that sustain the planet. Diatoms (pictured above) are photosynthetic microorganisms that play a role in global carbon cycling and sequestration. Famous for their wide variety of intricately shaped silica walls, these organisms are abundant in plankton and in marine and freshwater sediments, often being found in fossil deposits.

The Microbial World:

A Vast and Genetically Rich Resource

Microbes and their communities make up the foundation of the biosphere and sustain all life on earth. These single-celled organisms are masters at living in almost every environment and harvesting energy in almost any form, from solar radiation to photosynthesis-generated organic chemicals to minerals in the deep subsurface.

Microbes have evolved over 3.5 billion years, transforming the atmosphere with oxygen (a by-product of photosynthesis) more than a billion years ago to create the environment for life as we know it. Some microbes can thrive in either aerobic (with oxygen) or anaerobic (without oxygen) conditions. Microbes also capture nitrogen from the atmosphere, make it available to plants (and other life forms), and carry out processes responsible for soil fertility. Most do not cause disease. The unique microbial biochemistries amassed over eons in every niche on the planet now offer a deep and virtually limitless resource of capabilities that can be applied to national needs, including DOE energy and environmental missions.

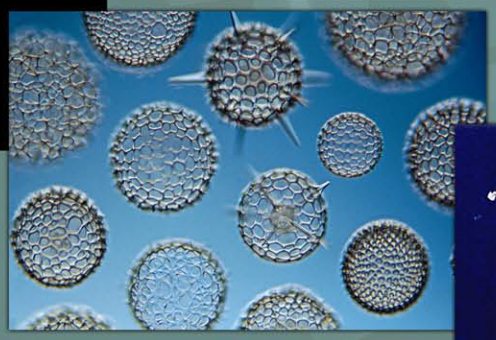
Although immense, the microbial world remains largely unexplored, a frontier of truly astronomical dimensions: The estimated nonillion or 10^{30} individual bacteria on earth are 10^9 times more than the number of stars in the universe. The vast majority, however, cannot be studied using standard techniques. While 2000 to 3000 species are estimated to be present in a single gram of soil, we can cultivate for study only some 0.1 to 1% of the species in that or any other environment. About 5700 species have been described thus far.[1–3]

Investigators now are beginning to apply the tools of genomics to studying this enormous untapped natural treasure. Because microbes have modest-sized genomes (averaging 4 to 5 million bases compared with 3 billion bases in the human and other mammalian genomes), they represent a tractable life form we can use to explore and understand life processes at a whole-system level. Already, limited environmental sampling of microbes and their communities has led to the discovery of millions of previously unknown genes and proteins, thousands of species, and innumerable variations in critical functionalities. As scientists begin to scratch the surface of the microbial world, they are finding analysis an enormous challenge.

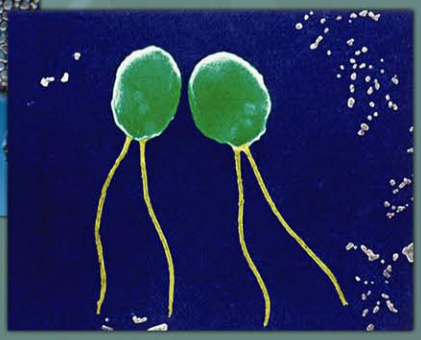
Recent discoveries from projects funded by DOE's Biological and Environmental Research program highlight the ubiquitous presence and critical importance of microbes in all ecosystems. For example:

- The cyanobacteria *Prochlorococcus* and *Synechococcus*, along with other ocean phytoplankton, account for about half of global photosynthesis.[4]

***Micrasterias rotata*, a Desmid Undergoing Cell Division or Cytokinesis.** Desmids are attractive unicellular freshwater green algae that have a distinct basic symmetry consisting of two semicells with the nucleus situated in the narrow center. When they divide, two new semicells are formed.
[© Wim van Egmund / Visuals Unlimited]



***Spumellarian radiolarian*, Skeletons from the Ocean Bottom.** Radiolarians are unicellular protists with strikingly beautiful siliceous skeletons showing radial symmetry.
[© Wim van Egmund / Visuals Unlimited]



***Chlamydomonas*, Green Algae with Two Flagella for Movement.** These microbes can generate hydrogen from light, water, and basic nutrients.
[Elias Greenbaum, Oak Ridge National Laboratory]

- Diatoms, ancient and intricately shaped ocean microbes, store an amount of carbon comparable to that in all the earth's rainforests combined. Over geological time, diatoms may have influenced the earth's climate.[5]
- More than a million previously undiscovered genes, possibly representing new biochemical functions, were the surprising find in sequencing DNA fragments from the Sargasso Sea—a region heretofore thought to sustain little life.[6] This discovery also was named one of *Science* magazine's "Breakthroughs of the Year."[7]

A Challenging Frontier

- Microbes thrive deep within the earth's subsurface and at extremes previously thought to extinguish life.[8]

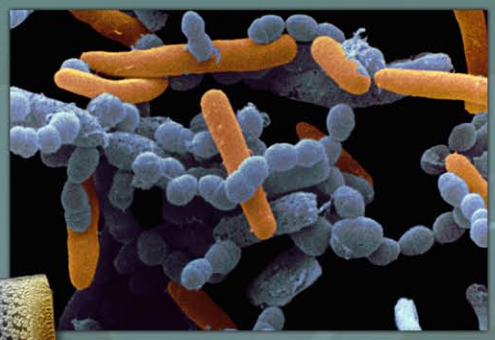
Growing recognition of microbial capabilities and potential applications has made a compelling case for further investigations by DOE and other agencies and institutions.

Before we can harness their capabilities, microbes must be understood in far greater detail and in the realistic context of whole living systems—whether as individuals or communities of interacting microbes—rather than as isolated components such as single genes and proteins. Microbes already can be manipulated at the molecular, cellular, and system levels, but understanding and taking advantage of their complexities and surmounting the technical challenges of whole-systems biology is a daunting prospect.

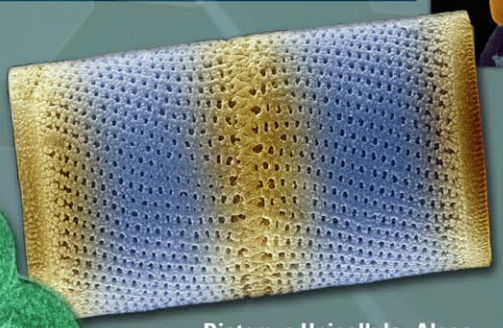
Understanding MICROBES and Their Communities

Most microbes live in highly organized and interactive communities that are versatile, complex, and difficult to analyze from many perspectives. Some of these challenges are outlined below.

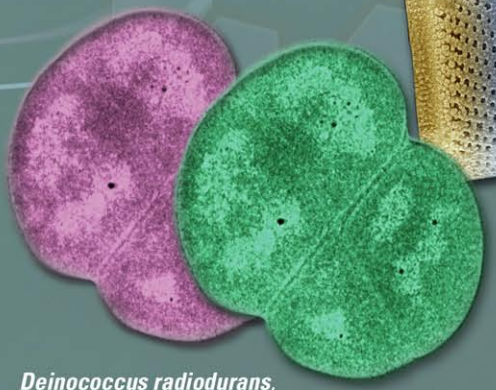
- Microbes are exceedingly small—only 1/8000th the volume of a human cell and spanning about 1/100th the diameter of a human hair. Investigating processes within this size range is challenging.
- The microbial world encompasses millions of genes from thousands of species, with hundreds of thousands of proteins and multimolecular machines operating in a web of hundreds of interacting processes in response to numerous physical and chemical environmental variables. Gene control is complex, with groups or "cassettes" of genes (operons) directing coordinated transcription and translation of genes into interacting proteins.
- Microbes adapt rapidly in response to environmental change, an ability that underlies their survival for billions of years. For example, various species of "extremophile" microbes have adapted to great extremes of pressure, temperature, pH, salinity, and radiation. Their high surface-to-volume ratio enhances interactions and supports adaptation. Unlike animal cells, they have no protective nucleus for their DNA, which leaves it more vulnerable to alteration. Genes move easily among species. Moreover, microbial communities are awash in genetic material from viruses that confer additional genetic properties and expand their range of adaptability.



Rod-Shaped (Bacilli) and Spherical (Cocci) Bacteria Found in Compost. Decomposition of organic matter is an extremely important process in nature and a part of the global carbon and nutrient cycle. [© Simko / Visuals Unlimited]



Diatom, a Unicellular Algae. The cell walls of diatoms are made of silica and come in a variety of shapes. These microscopic algae may be either fresh or saltwater, are photosynthetic, and play a role in carbon cycling. [© Stanley Flegler / Visuals Unlimited]



***Deinococcus radiodurans*, the Most Radiation-Resistant Microbe Known.** [Michael Daly, Uniformed Services University of the Health Sciences]

- Microbial communities can extend in size from cubic millimeters (or smaller) to cubic kilometers. Even relatively simple communities can have millions of genes, giving them a genetic diversity substantially greater than that of higher life forms, even humans. Recent investigations have focused on collecting DNA fragments from environmental samples in the sea and other natural ecosystems. These "metagenomics" studies have given us a glimpse into the intricacies of these natural ecosystems and their diverse functions.

References noted on these pages are listed on the last page of this section.

Microbes on the Move

Chemotaxis: Sensing and Moving in a Chemical Gradient

Motile bacteria use sophisticated information-processing devices to detect and respond to changes in their chemical environments. *Escherichia coli* cells, for example, use a signaling cascade of protein phosphorylation and dephosphorylation reactions to control the stiff flagellar filaments responsible for cell motility. The filaments, made up of several individual molecular motors, rotate to propel the cell in favorable directions (called chemotaxis) or allow it to tumble randomly. Cells make motility decisions by comparing current conditions to those occurring previously. Chemical changes as minute as 1 part per 1000 can be detected.

The methylation state of transmembrane chemoreceptors (methyl-accepting chemotaxis proteins) encodes the memory of its chemical environment and controls the flux of phosphates through a signaling cascade. Two signals are produced: A feed-forward signal that alters the motor rotation and a feedback signal that updates the methylation record. Motor responses occur in a few hundred milliseconds, whereas sensory-adaptation machinery updating the methylation record takes several seconds. New in vivo experimental approaches are needed to better understand the functional anatomy of bacterial receptor clusters.

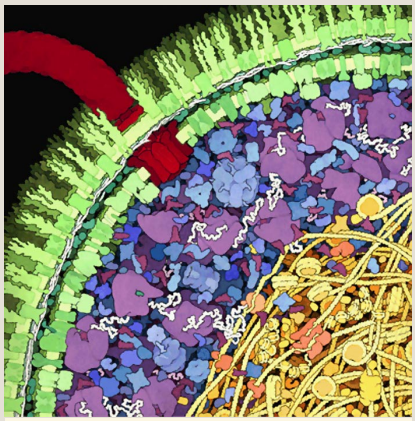


Illustration of *E. coli* Cell with Protruding Flagellum.

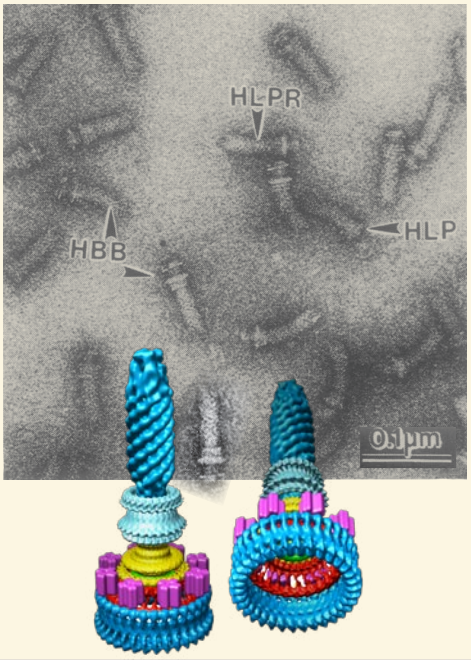
© 1999 D. Goodsell, Scripps Research Inst.

Reference

J. S. Parkinson, "Signal Amplification in Bacterial Chemotaxis Through Receptor Teamwork," *ASM News* (2004).

Built for Motility: A Complex Molecular Motor

The bacterial flagellum consists of multiple copies of at least 13 different proteins. This multiprotein machine contains an axial structure running through the length and a set of ring structures within the basal body embedded in the cell envelope. At right is a conventional electron microscopic image of uranyl acetate-stained flagellar complexes and subcomplexes prepared from *Salmonella typhimurium*. The hook-basal body (HBB in photo) contains a hook and rings of proteins; HL-PR refers to the hook and L and P rings, while HLP refers to the complex containing hook, L and P rings, and distal portion of the rod.



Flagellar Complexes from *S. typhimurium*.

Inset: Surface Representation of a 3D Map of the Bacterial Hook-Basal Body Flagellar Complex of *S. typhimurium*.

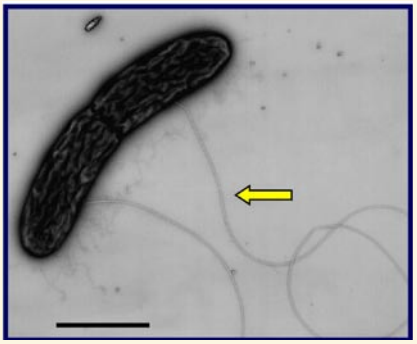
Inset illustration: D. Thomas, N. Francis, and D. DeRosier, Brandeis Univ.
Photomicrograph: C. Anderson, Brookhaven National Laboratory

Reference

G. E. Sosinsky et al., "Mass Determination and Estimation of Subunit Stoichiometry of the Bacterial Hook–Basal Body Flagellar Complex of *Salmonella typhimurium* by Scanning Transmission Electron Microscopy." *Proc. Natl. Acad. Sci. USA* 89(11), 4801–5 (1992).

Growing Flagella in a Pinch

Some microbes grow flagella only when they need to find their way to nutrients essential to metabolic processes. Such is the case with *Geobacter metallireducens*, which produces energy for biochemical reactions by transferring electrons to metals. Although originally thought to be nonmotile, *G. metallireducens* genome analysis turned up genes encoding flagella. Further investigations showed that this microbe produces flagella when faced with insoluble sources of iron or manganese [Fe(III) or Mn(IV)]. Genes for pili (fine hair-like structures on the microbe's surface) also are present and expressed during growth on insoluble oxides; studies indicate their role as facilitating movement toward and aiding attachment to iron oxides (see sidebar, Bacteria Use "Nanowires" to Facilitate Extracellular Electron Transfer, p. 73). Additional genes for chemotaxis also were apparent in the genome, leading to the discovery of a novel mechanism for chemotaxis to iron. Understanding *Geobacter's* physiology is important for optimizing strategies to use this organism to bioremediate metals such as uranium in contaminated subsurface environments. Global gene-expression studies are helping to identify regulatory circuits, specifically those involved in bioremediation pathways and electricity production. [Source: Derek Lovley, University of Massachusetts, Amherst]



Flagella (arrow) Produced by *Geobacter* in the Presence of Insoluble Sources of Iron or Manganese.

D. Lovley, University of Massachusetts, Amherst

References

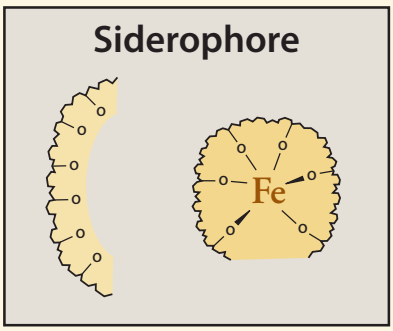
S. E. Childers, S. Ciufo, and D. R. Lovley, "*Geobacter metallireducens* Accesses Insoluble Fe (III) Oxide by Chemotaxis," *Nature* **416**, 767–69 (2002).

D. R. Lovley, "Cleaning Up with Genomics: Applying Molecular Biology to Bioremediation," *Natl. Rev. Microbiol.* **1**, 35–44 (2003).

Siderophores: Sending out Shuttles to Scout for Iron

Another efficient mechanism evolved by many microbes to obtain iron in limited environments such as marine surfaces is the production and secretion of siderophores. These low-molecular-weight chelating agents act as shuttles to bind insoluble iron [Fe(III)] and transport it back to the microbe, where it enters the cell by recognizing specific membrane receptor proteins and transport systems. More than 500 types of siderophores are known to exist. In addition to supplying essential nutrients, siderophores of one organism can lock up iron to achieve an advantage over their competitors. However, some microbes were discovered recently to have receptors for the siderophores of other organisms.

Until the genetic sequence of the ocean diatom *T. pseudonana* was determined in the DOE Microbial Genome Program and compared with sequences of other organisms, researchers were unaware that these organisms possessed siderophores. Diatoms, along with other ocean microbes, contribute to absorbing CO_2 in amounts comparable to that absorbed by all the world's tropical rain forests combined. Obtaining more detailed knowledge of their life processes will help us better understand their vital role in global carbon cycling.



Schematic Representation of a Siderophore Before and After Iron Acquisition. [Adapted from H. Boukhalfa and A. L. Crumbliss, "Chemical Aspects of Siderophore-Mediated Iron Transport," *Biometals* **15**, 25–39 (2002).]

Reference

E. V. Armbrust et al., "The Genome of the Diatom *Thalassiosira pseudonana*: Ecology, Evolution, and Metabolism," *Science* **306**, 79–86 (2004).

Group Living and Communicating

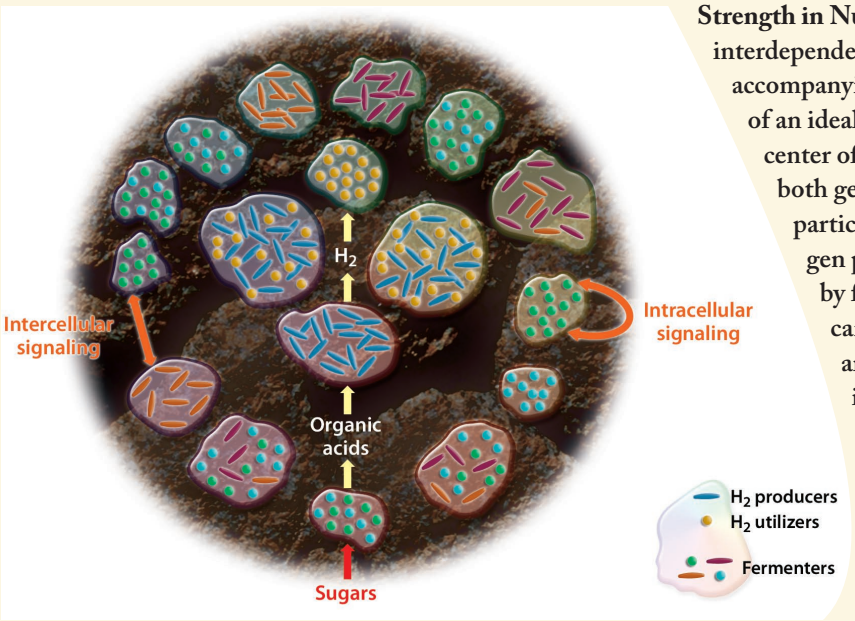
Life in a Biofilm

Most microbes live attached to solid surfaces (biotic or abiotic) within highly organized and functionally interactive communities called biofilms. These biofilms can be composed of populations that developed from a single species or a community derived from multiple species. All exhibit collective and interdependent behavior, with different genes rapidly brought into play as conditions dictate (see figure below). Among the many advantages of biofilm living are nutrient availability with metabolic cooperation, acquisition of new genetic traits, and protection from the environment.

Researchers are only beginning to realize the prevalence and significance of biofilms. These communities probably play major roles in such complex natural processes as the cycling of nitrogen and sulfur and the degradation of environmental pollutants and organic matter, activities that require a range of metabolic capabilities. Recent metagenomic and metaproteomic studies focused on biofilm members in an acid mine drainage environment (see sidebar, Metagenomics, p. 62).

Reference

M. E. Davey and G. A. O'Toole, "Microbial Biofilms: From Ecology to Molecular Genetics," *Microbiol. Mol. Biol.* **64**(4), 847–67 (2000).



Strength in Numbers. Microbes in biofilms live an interdependent, community-based existence (see accompanying text above). In this overhead view of an idealized biofilm, four microcolonies in the center of the figure represent organisms that both generate and consume hydrogen. Two participate in syntrophism, in which hydrogen producers use organic acids generated by fermenting organisms that gain their carbon and energy by using various sugars. In addition to potential metabolic interactions, signaling molecules may aid in inter- and intraspecies communication. These genetic factors and environmental influences contribute to the biofilms' spatial organization. [Figure and caption adapted from Davey and O'Toole, 2000.]

Quorum Sensing

Microbes communicate with each other by sending and detecting a wide variety of chemical signals (autoinducers). These molecules trigger group behaviors, including the formation and persistence of biofilms, symbiosis, and other processes. Many of these activities are density dependent, that is, when a threshold concentration of chemicals is detected (reflecting a certain number of cells), microbes respond with a change in gene expression. This process, called quorum sensing, facilitates coordination of gene expression by the entire community, in essence enabling it to behave like a multicellular organism. Quorum sensing allows microbial communities to adapt rapidly to environmental changes and reap benefits that would be unattainable as individuals.

Quorum sensing first was described in the bioluminescent marine bacterium *Vibrio fischeri*. This microbe lives in symbiotic association with several marine animal hosts, who use the light it produces to attract prey, avoid predators, or find a mate. In exchange, *V. fischeri* obtains a nutrient-rich home environment. *V. fischeri* emits light only inside a specialized light organ of the host, where the concentration of these organisms becomes dense; it does not give off light when free living in the ocean. Light production depends on producing, accumulating, and responding to a minimum-threshold concentration of an autoinducer (acylated homoserine lactone). Only under the nutrient-rich conditions of the light organ can *V. fischeri* grow to such high populations. Also, trapping the diffusible autoinducer molecule in the light organ with the bacterial cells allows it to accumulate to a concentration sufficient for *V. fisheri* to detect it.

Recent studies have revealed diverse chemical languages that enable bacterial communication both within and between species (the latter called cross talk). The extracellular matrix surrounding mature biofilms (composed of glycans and other components) plays a crucial role in transmitting these chemical signals into and between cells. Biotechnological researchers are developing molecules structurally related to autoinducers to exploit quorum-sensing capabilities and possibly improve industrial production of natural products.

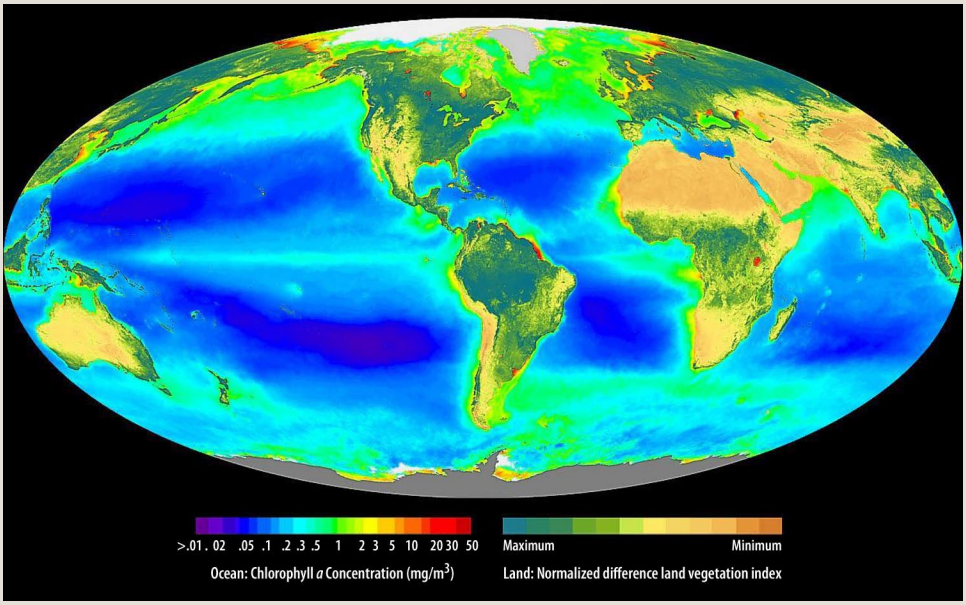
References

J. W. Hastings and K. H. Nealson, "Bacterial Bioluminescence," *Annu. Rev. Microbiol.* **31**, 549–95 (1977).

S. Schauder and B. L. Bassler, "The Language of Bacteria," *Genes Devel.* **15**, 1468–80 (2001).

Photosynthetic Microbes—Major Contributors to Earth's Life-Support System

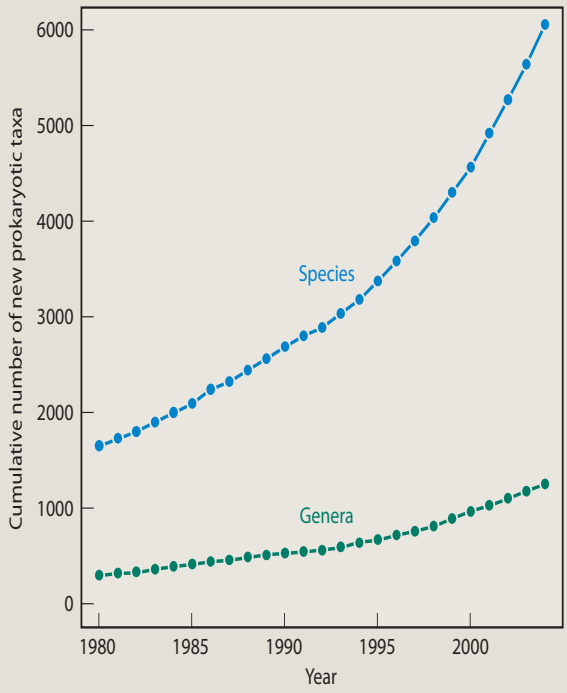
When photosynthetic microbes, fungi, and plants convert light energy from the sun into glucose, they establish the foundation for the food chain on which all life, including human, depends. This NASA SeaWiFS image of the global biosphere shows the density of photosynthetic organisms on land and in the oceans. On land, the dark greens represent areas of abundant vegetation, with tans showing relatively sparse plant cover. In the oceans, red, yellow, and green pixels depict dense blooms of phytoplankton (photosynthetic microbes), while blues and purples show regions of lower productivity.



Ocean's Long-Term Average Phytoplankton Chlorophyll Concentration. Image combines September 1997 through August 2000 concentration data with the SeaWiFS-derived Normalized Difference Vegetation Index over land; http://oceancolor.gsfc.nasa.gov/SeaWiFS/.

Cataloging Microbial Diversity

What's in a Name? The Challenges of Tracking Microbial Species



Prokaryotic systematics is a dynamic field. The rate at which new species, genera, and higher taxa are described in the literature has increased dramatically since the 1990s (see figure at left), driven largely by advances in sequencing technology. Even with all this progress, scientists believe that 99% of the microbial world has yet to be discovered.

Just how many species (and genera) of *Bacteria* and *Archaea* are listed in *Bergey's Manual*, a widely used international reference for taxonomy? It seems a simple question, but, at present, the number of named species actually exceeds the true number of species having official standing in the nomenclatural record by about 22%. An explanation follows.

Many taxa bear two or more names, because when species or higher taxa are reassigned to existing or newly created taxa, both the new name and the old name are valid (in the original published context). Other types of nomenclatural synonyms exist as well. As a result, about 6900 species are listed in *Bergey's*, with information and data published in accordance with *The International Code of Bacterial Nomenclature*. The true number of named prokaryotes, however, is closer to 5700. [Source: George Garrity, *Bergey's Manual*]

References for pp. 14–15

1. T. P. Curtis and W. T. Sloan, "Prokaryotic Diversity and Its Limits: Microbial Community Structure in Nature and Implications for Microbial Ecology," *Curr. Opin. Microbiol.* **7**, 221–26 (2004).

2. V. Torsvik and L. Øvreås, "Microbial Diversity and Function in Soil: From Genes to Ecosystems," *Curr. Opin. Microbiol.* 5(3), 240–45 (2002).

3. *Bergey's Manual*, 2nd edition (2001).

4. T. S. Bibby et al., "Low-Light–Adapted *Prochlorococcus* Species Possess Specific Antennae for each Photosystem," *Nature* **424**, 1051–4 (2003).

5. E. V. Armbrust et al., "The Genome of the Diatom *Thalassiosira Pseudonana*: Ecology, Evolution, and Metabolism," *Science* **306**, 79–86 (2004).

6. J. C. Venter et al., "Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science* **304**, 66–74 (2004).

7. The News Staff, "Breakthrough of the Year: The Runners Up," *Science* **306**, 2013–17 (2004).

8. T. C. Onstott et al., "Indigenous and Contaminant Microbes in Ultradeep Mines," *Environ. Microb.* 5(11), 1168–91 (2003); K. Kashefi and D. R. Lovley, "Extending the Upper Temperature for Life," *Science* **301**, 934 (2003).

2.0. Missions Overview: The Role of Microbial Systems in Energy Production, Environmental Remediation, and Carbon Cycling and Sequestration

2.1. Introduction to GTL Goals for DOE Missions 22

2.2. GTL Research Analyzing Mission-Relevant Systems 22

2.2.1. Engineered Systems 23

2.2.2. Natural Ecosystems 23

2.2.3. Shortening the Missions Technology Cycle 23

2.3. Basic Energy Research: Develop Biofuels as a Major Secure Energy Source 24

2.3.1. Ethanol Production from Cellulose 24

2.3.1.1. Science and Technology Objectives 25

2.3.1.2. Other Commercial Products 28

2.3.2. Biophotolytic Hydrogen Production 29

2.3.2.1. Science and Technology Objectives 29

2.4. Environmental Remediation: Develop Biological Solutions for Intractable Environmental Problems 31

2.4.1. Science and Technology Objectives 32

2.5. Microbial Roles in Carbon Cycling and Sequestration: Understand Biosystems' Climate Impacts and Assess Sequestration Strategies 33

2.5.1. Science and Technology Objectives 35

2.5.2. Marine Microbial Communities 35

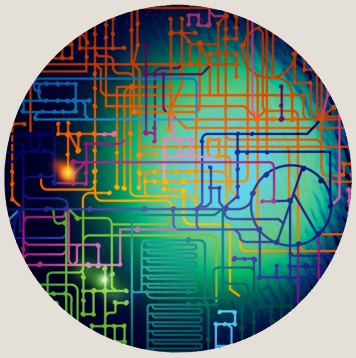
2.5.2.1. Specific Scientific Needs for Marine Microbial Communities 36

2.5.3. Terrestrial Microbial Communities 36

2.5.3.1. Specific Scientific Needs for Terrestrial Microbial Communities 36

2.6. Summing Up the Challenges 37

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, the Bioenergy Research Center, will concentrate on overcoming biological barriers to the industrial production of biofuels from biomass and on other potential energy sources such as biophotolytic decomposition of water to hydrogen. 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://doegenomestolife.org/biofuels/).



Missions Overview: The Role of Microbial Systems in Energy Production, Environmental Remediation, and Carbon Cycling and Sequestration

The Department of Energy's overarching missions are to advance the national, economic, and energy security of the United States; promote scientific and technological innovation in support of that mission; and ensure environmental cleanup of the national nuclear weapons complex (*The Department of Energy Strategic Plan,* 2003).

2.1. Introduction to GTL Goals for DOE Missions

The complexity of DOE's missions requires groundbreaking research and integration across multiple disciplines to create new generations of technologies. In the coming decades, bioscience and biotechnology must play an increasing role in informing policy and decision making and providing innovative solutions (see Fig. 1. Grand Challenges for Biology, Payoffs for the Nation, p. 23). The earth's microbial systems are the foundation for life and a potential source of capabilities that we can put to use to meet national challenges; their study forms the core of the GTL program. This chapter links the scope of DOE missions, some potential microbial contributions, and science and technology objectives to achieve timely impacts.

The ultimate GTL scientific goal is to attain a predictive systems-level understanding of microbes. Each mission area has a distinct technical endpoint and set of subsidiary science goals as described in this chapter. These goals define a unique set of research challenges that collectively will require new capabilities and a large body of integrated knowledge on every aspect of microbial systems behavior.

2.2. GTL Research Analyzing Mission-Relevant Systems

In the first phase of the GTL program, research projects are focusing on basic biological studies relating to mission-relevant systems. The goals are to understand scientific issues and challenges, begin to use new generations of research technologies, learn how to apply computation and modeling, and work in a multidisciplinary team environment (see 3.3. Highlights of Research in Progress to Accomplish Milestones, p. 55, and Appendix E. GTL-Funded Projects, p. 245). DOE BER has sequenced the genomes of nearly 200 microbes with wide-ranging biochemical capabilities (see Appendix G. Microbial Genomes Sequenced or in Process by DOE, p. 253). Some of the microbes and microbial communities being studied in GTL have potential for stabilizing toxic metals and radionuclides, degrading organic pollutants, producing energy feedstocks including biofuels and hydrogen, sequestering carbon, and playing a critical role in cycling ocean carbon and other elements.

2.2.1. Engineered Systems

Biology will be transformed into a quantitative and model-based science to allow the kind of systems engineering that has typified material- and chemical-based mission technologies. We need to understand the molecular mechanisms of microbial processes well enough to reliably redesign systems for new applications in unique engineered environments. This scientific research requires multidisciplinary teams focused on substantial goals.

2.2.2. Natural Ecosystems

For climate and environmental applications, GTL will develop the capability to understand the molecular mechanistic processes and the global responses of microbes in ecosystems. Metagenomics, a new field of culture-independent genomic analysis of microbial communities (Schloss and Handelsman 2003), is revealing that microbes in oceans and soils are substantially more genetically and potentially more biochemically diverse than expected. One recent experiment in the Sargasso Sea has uncovered more than a million genes, resulting in doubling the total amassed set of sequenced genes in the world (Venter et al. 2004; Meyer 2004). Since fully 40% of genes encountered are of unknown function and even homology-assigned functions often are uncertain, we must devise means to analyze very large numbers of genes in the laboratory without expressing them in their native hosts. In addition, the wide genetic diversity of hydrogenases and bacteriorhodopsins, for example, calls for analyses to determine the functional significance of gene variations. Specifically, working from genomic sequence, we can create and characterize proteins to estimate function. Ultimately, we will measure the molecular responses of cultured cells or cells in their natural environments (e.g., transcriptomes, proteomes, metabolomes). Mining the global gene pool also presents an opportunity to discover new genes, processes, and species that could point the way for biotechnology applications for DOE missions.

2.2.3. Shortening the Missions Technology Cycle

GTL knowledge, experimental capabilities, and facilities coupled with the GTL computational biology environment will form a bridge between science and applications. Comprehensive data and models will allow scientific discoveries at molecular-level time and spatial scales to be incorporated into larger models and simulations. These will cover the large process, spatial, and time scales used in mission applications for systems engineering of application technologies (e.g., biofuel production and bioremediation) and policy-support products (e.g., climate models, economic models, and integrated assessments). These capabilities will contribute to a dramatic shortening of the technology cycle, allowing frontier science to be incorporated more directly into useful systems and reducing time and costs between discovery and use.

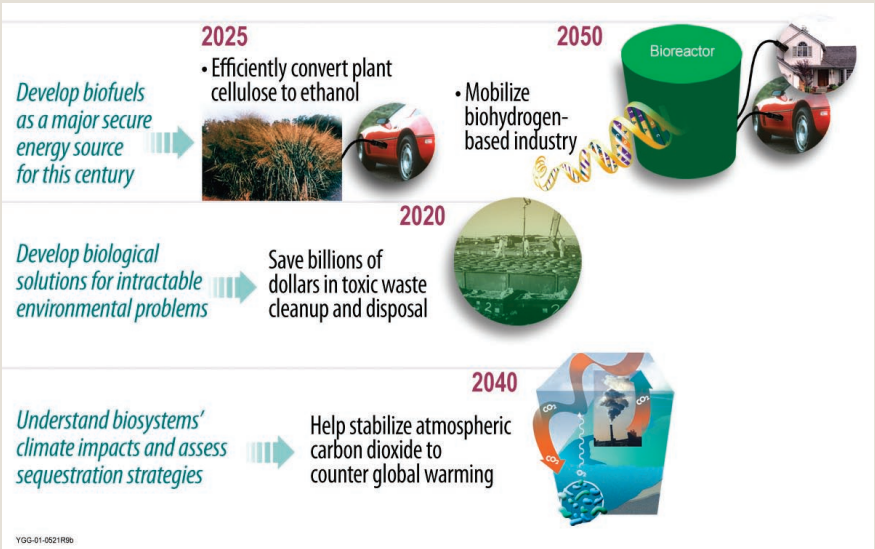


Fig. 1. Grand Challenges for Biology, Payoffs for the Nation.

2.3. Basic Energy Research: Develop Biofuels as a Major Secure Energy Source

Global energy demand is projected to rise rapidly in this century due to population growth and increasing worldwide gross domestic products, standards of living, and the energy intensity of developing economies (see sidebar, The Framework for DOE Missions, this page, and Appendix H. Programs Complementary to GTL Research, p. 265).

The national energy strategy's central tenet is that technology development will enable deployment of necessary energy resources and greenhouse gas (GHG) abatement as world economies build out the energy infrastructure to meet increased demand. Only a small part of the global energy infrastructure required by the end of the century exists today, and much of that will require replacement in the next 50 years. Without new energy technologies and sources, this situation will result in raising GHG emission levels significantly and increasing the strain on global energy supply lines and their security and on economic growth. Numerous technology options are required, and biotechnology is projected to have a substantial role in this buildout (Abraham 2004; Pacala and Socolow 2004; Socolow 2005).

The Framework for DOE Missions

DOE's Energy-Production and Climate-Change Strategies. DOE pursues its energy technology and climate science goals under the multiagency framework defined by the Climate Change Science Program (CCSP) and the Climate Change Technology Program (CCTP)—both founded on a technology-development strategy. The programs seek to understand climate change, reduce greenhouse gas (GHG) emissions, provide growing global economies with adequate and inexpensive energy, and improve GHG emission monitoring (see Appendix F. Strategic Planning for CCSP and CCTP, p. 249).

DOE Environmental Remediation Commitment. Agreements among DOE, the Environmental Protection Agency, and affected states have committed them to cleaning up the legacy of defense-related nuclear activities (i.e., large volumes of soil, sediments, and groundwater contaminated with metals, radionuclides, and a variety of organics). DOE estimates that, without major technical breakthroughs, cleanup will take about 35 years at a cost up to $142 billion (Closure Planning Guidance 2004).

By 2100, biotechnology-based energy use could equal all global fossil energy use today (see Fig. 2. Filling the Technology Gap, p. 25). Biologically derived fuels are renewable and expandable to meet the growing demand. They are domestically and globally available for energy security, with most being carbon neutral—or potentially carbon negative (if coupled with sequestration)—and supportable within the current agricultural infrastructure.

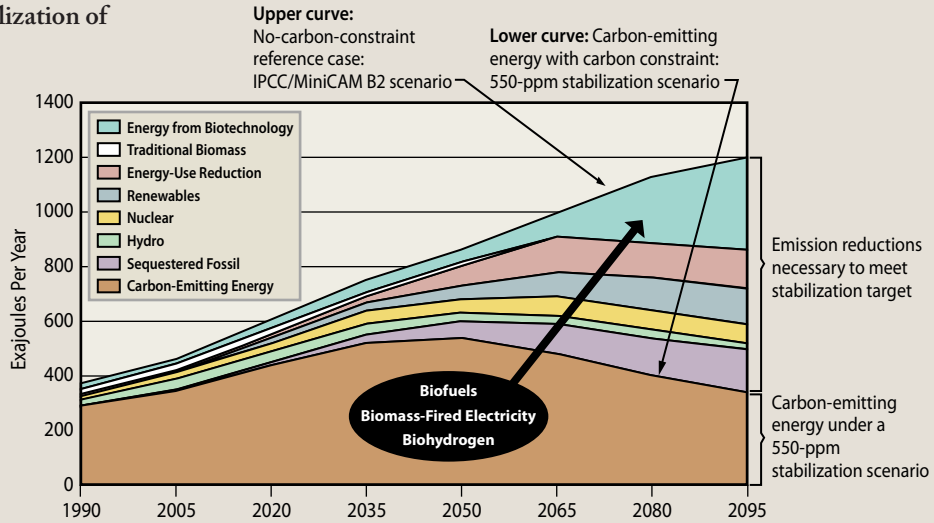
Two example biofuels discussed here are cellulose-derived ethanol and biophotolytic hydrogen. Cellulosic ethanol, a carbon-neutral fuel, is usable with the existing energy infrastructure. Hydrogen is the ultimate carbon-free energy carrier that can be converted efficiently to energy in fuel cells, with water as the only chemical by-product. Other potential biofuels include lipids, biodiesel, ammonia, methane, and methanol, each with multiple production options. Other future energy systems might include fuel cells based on biological processes.

The technology endpoint for energy systems is global deployment of engineered biological or biobased processes. This application requires a science base for molecular and systems redesign of numerous proteins, pathways, and full cellular systems. Biofuels could be produced using plants, microbes, and enzymatic solutions. Understanding the entities involved in these processes and the principles that govern biological mechanisms will allow scientists and technologists to design novel biofuel-production strategies such as engineered nanobiostructures (see sidebar, Synthetic Nanostructures, p. 25).

2.3.1. Ethanol Production from Cellulose

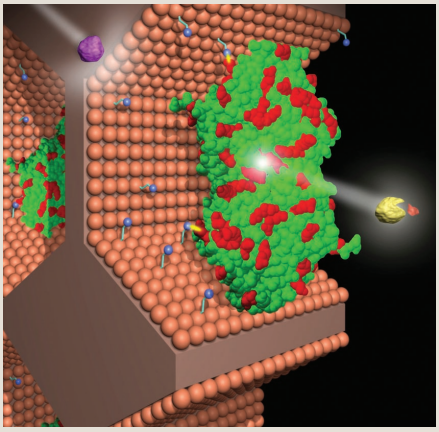
Biofuels such as cellulosic ethanol can provide alternatives to oil, displacing it as a transportation fuel with security, economic, and environmental benefits. Cellulosic ethanol can be cost competitive with oil-based gasoline and can reduce net CO_2 emissions from the transportation sector (roughly one-third of U.S. emissions) by more than 80% at no extra cost (Greene et al. 2004; Mann 2004). The cellulosic ethanol option would allow us to invest our energy dollars domestically, providing a profitable crop for farmers. In addition to

Fig. 2. Filling the Technology Gap. Based on technical and economic analyses, this figure compares two hypothetical scenarios for energy demand and supply growth over this century: (1) The IPCC/MiniCAM B2 scenario (upper curve), which assumes a relatively unchanged energy mix, is not carbon constrained; (2) A carbon-constrained scenario that stabilizes atmospheric concentrations of CO_2 at 550 ppm (family of curves below) was chosen to illustrate the types of changes in energy mix that might occur. An acceptable level of atmospheric CO_2 is still to be determined. U.S. energy strategy is based on technology development to provide multiple options to fit various national and global contingencies, market forces, and the ultimate stabilization of carbon emissions to near zero. This combination of technologies includes energy-use reduction, new and expanded nonemitting energy sources, and carbon sequestration. To meet these goals in the scenario illustrated here, analysis indicates that by the end of the century biotechnology sources of energy must grow to roughly equal today's fossil-fuel usage.



Synthetic Nanostructures: Putting Microbial Capabilities to Work

Understanding the sophisticated biochemistries of microbes can lead to the discovery of ways to isolate and use their components to carry out some of the functions of living cells. An example in this figure shows the enzyme organophosphorus hydrolase (OPH), which has been embedded in a synthetic nanomembrane (mesoporous silica) that enhances its activity and stability [*J. Am. Chem. Soc.* 124, 11242–43 (2002)]. The OPH transforms toxic substances (purple molecule at left of OPH) to harmless by-products (yellow and red molecules at right). Applications such as this could optimize the functionality of countless enzymes for efficient production of energy, removal or inactivation of contaminants, and sequestration of carbon to mitigate global climate change. The knowledge gained from GTL also could be highly useful in food processing, pharmaceuticals, separations, and the production of industrial chemicals.



E. Ackerman, Pacific Northwest National Laboratory

reducing GHGs, these crops improve air and soil quality, reduce soil erosion, and expand wildlife habitat. GTL research can contribute to making cellulosic ethanol more economical and practical by decreasing the complexity and cost of processing cellulose to ethanol.

2.3.1.1. Science and Technology Objectives

The first step in increasing the economic viability of biofuels and biochemicals, including ethanol, is to use cellulose and other such biomass constituents as hemicellulose and lignin instead of the ultimately limited food starch that predominates today. Ethanol from cornstarch has a 14% energy yield (i.e., net energy content of the feedstock converted to energy in ethanol), whereas cellulose can have a 37% yield (see Table 1, Cellulosic Ethanol Goals and Impacts, p. 26) resulting from improved process efficiencies

(Smith et al. 2004). Use of waste cellulose can provide an energy source equal to 10% of current gasoline usage; to achieve greater impacts, energy crops must be used (Mann 2004). Cellulose, a carbohydrate polymer that makes up plant cell walls, is the most abundant biological material (see sidebar, Cellulose: Microbes Process it into Ethanol-Convertible Sugars, p. 27). The strong, rigid, water-insoluble nature of cellulose and the cell's other structural materials, however, makes them resistant to degradation into sugars and difficult to process into ethanol. The complex multistep process for commercially converting cellulose into ethanol currently combines thermochemical and biological methods in large centralized processing plants. With improved enzyme systems, we can replace expensive thermochemical processes. The ultimate innovation, integrated processing—combining all key hydrolytic and fermentative steps in one process using either a single microbe or stable mixed culture—would enable smaller-scale and more cost-effective and energy-efficient distributed processing plants.

Biofuel Development

- **Mission Science Goals: Understand the principles underlying the structural and functional design 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.**
- **Challenges: Analyze thousands of natural and modified variants of such processes as cellulose degradation, fermentative production of ethanol or other liquid fuels, and biophotolytic hydrogen production.**

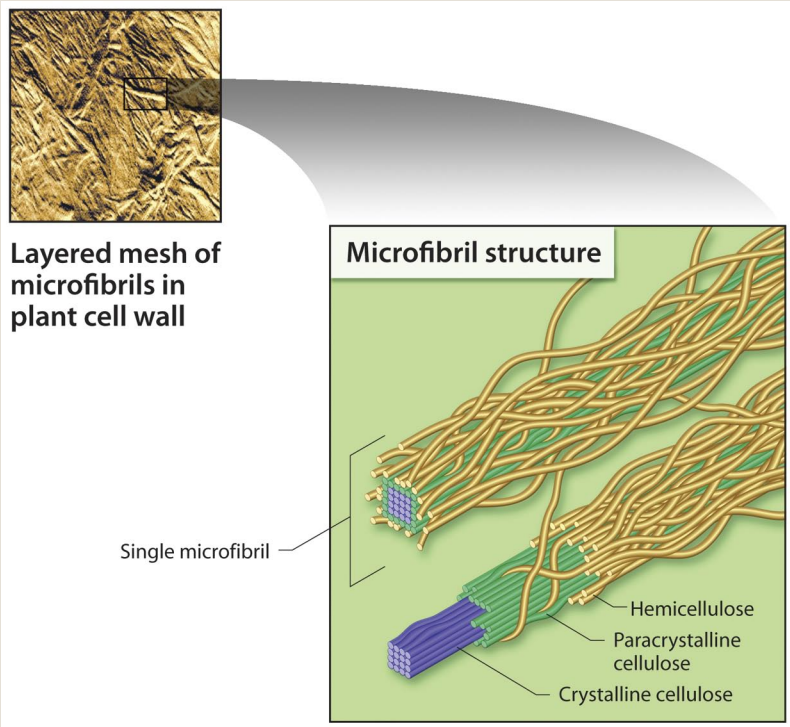
Biomass-degrading microbes and fungi are sources of enzymes that can improve wood preprocessing and cellulase enzymes that break down cellulose into fermentable sugars (see Fig. 3. Converting Cellulose to Sugars, p. 27). Large numbers of cellulase-producing organisms and potentially thousands of bacteria and yeast species can convert simple sugars to ethanol. An important part of GTL science will be to analyze and screen different microbes, fungi, and natural microbial communities to increase the number of enzymes that can be examined. The DOE Joint Genome Institute has determined the genome sequence of white-rot

Table 1. Cellulosic Ethanol Goals and Impacts

Factors	Today	Interim	Long-Term*
Billion gallons Fossil fuel displaced** CO_2 reduced	4 2% 1.8%	20 10% 9%	30 to 200 15 to 100%*** 14 to 90%
Feedstock****	Starch (14% energy yield)	Waste cellulose	Cellulosic energy crops (>37% energy yield)
Process	Starch fermentation Little cellulose processing	Acid decrystallization: Transition to enzymes Cellulases Single-sugar metabolism Multiple microbes Some energy crops	Enzyme decrystallization and depolymerization Cellulase and other glycosyl hydrolases Sugar transporters High-temperature functioning Multisugar metabolism Integrated processing Designer cellulosic energy crops Carbon sequestration through plant partitioning
Deployment	Large, central processing	Large, central processing	Distributed or centralized, efficient processing plants
Other impacts: Energy dollars spent at home, third crop for agriculture, land revitalization and stabilization, habitat, soil carbon sequestration, yield per acre roughly tripled (cellulose over corn starch).			

*Enabled by GTL.
**Current U.S. consumption of gasoline is about 137 billion gallons per year, which corresponds to about 200 billion gallons of ethanol (Greene et al. 2004) because a gallon of ethanol has 2/3 the energy content of a gallon of gasoline.
***Assumes improvements in feedstocks, processes, and vehicle fuel efficiency.
****Adapted from Smith et al. 2004.

Cellulose: Microbes Process It into Ethanol-Convertible Sugars



Cellulose, the main structural component of plant cell walls, is a linear polymer consisting of thousands of glucose residues arranged in a rigid, crystalline structure. Layers upon layers of cellulose-containing microfibrils give plant cell walls their remarkable strength. Each microfibril consists of a crystalline cellulose core encased within a complex outer layer of amorphous polysaccharides known as hemicellulose. The crystallinity of cellulose and its association with hemicellulose and other structural polymers such as lignin are two key challenges that prevent the efficient breakdown of cellulose into glucose molecules that can be converted to ethanol. Adding to the difficulty is the diverse mix of simple sugar molecules generated from the hydrolysis of cellulose and hemicellulose. Fermentative microorganisms prefer to use six-carbon sugars (e.g., glucose) as substrates for producing ethanol; however, hemicellulose is composed of a variety of five-carbon sugars that are not efficiently converted into ethanol by microorganisms. [Microfibril structure adapted from J. K. C. Rose and A. B. Bennett, "Cooperative Disassembly of the Cellulose-Xyloglucan Network of Plant Cell Walls: Parallels Between Cell Expansion and Fruit Ripening," *Trends Plant Sci.* 4, 176–83 (1999).]

Fig. 3. Converting Cellulose to Sugars. Cellulases include a mix of enzymes that break down cellulose into simple sugars that can be fermented by microorganisms to ethanol. Three general classes of cellulases—endoglucanases, exoglucanases, and cellobiases—work together in a coordinated fashion to hydrolyze cellulose. Endoglucanases internally cleave a cellulose chain, and exoglucanases bind the cleaved ends of the cellulose chain and feed the chain into its active site where it is broken down into double glucose molecules called cellobiose. Cellobiases split cellobiose to yield two glucose molecules. The cellulase pictured is an exoglucanase whose binding domain on the right extracts a cellulose chain. At the active site in the larger catalytic domain on the left, the cellulose chain is hydrolyzed to yield cellobiose subunits. [Image from M. Himmel et al., "Cellulase Animation," run time 11 min., National Renewable Energy Laboratory (2000).]

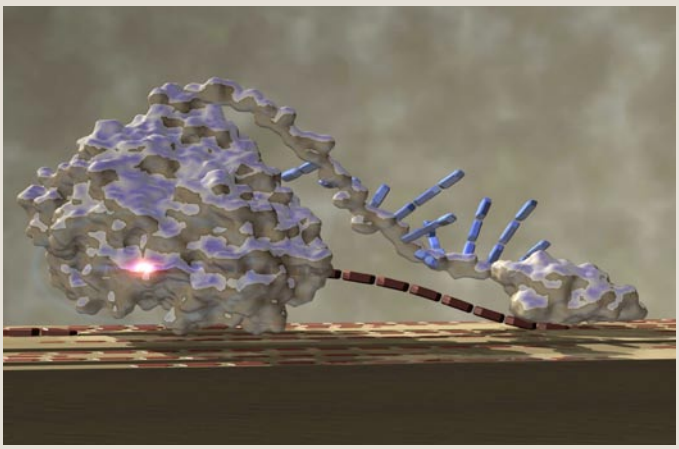


Fig. 4. White-Rot Fungus, this page). Such enzymes as cellulase and other glycosyl hydrolases are capable of hydrolyzing biomass polymers. Because these hydrolases are much slower in their intrinsic turnover rates (the number of molecules hydrolyzed per second) than most other enzymes, one goal is to improve their efficiency. To increase understanding of these multisubunit complexes and derive the principles of their function, large numbers of natural and modified cellulases and other molecular machines must be analyzed (see Table 2. Cellulosic Ethanol Challenges, Scale, and Complexity, this page).

GTL research can help by producing and characterizing complex cellulase structures and their functions, by analyzing naturally occurring and modified protein and molecular machine variants of essential pathways, understanding the synergistic activity of multiple cellulases, and resolving temperature-sensitivity issues that prevent optimal functioning of cellulase enzymes at fermentative temperatures. Other challenges include characterizing the structures and functions of membrane-bound molecular machines that deliver sugars to the metabolic pathways of fermentative organisms; understanding the inefficiencies in conversion of different sugars to ethanol; maintaining large-scale mixed cultures; and improving disease resistance. Finally, understanding cell regulatory processes will be central to incorporating multiple functionalities into a single organism or a microbial consortium and enabling optimized overexpression in many related processes.

For more information, refer to Summary Table. GTL Science Roadmap for DOE Missions, p. 40, and 5.0. Facilities Overview, p. 101.

2.3.1.2. Other Commercial Products

These grand challenges to biology posed by DOE missions will provide the foundation for countless new commercial bioproducts and bioprocesses. A strategy for introducing these technologies to the marketplace could be integrated biorefineries capable of producing a suite of products as substitutes for chemical-based fossil feedstocks.

Fig. 4. White-Rot Fungus. In this image of a longitudinal section of *Phanerochaete chrysosporium* colonizing aspen, hyphae are visible throughout and in the vessel pit on the right. This June 2004 issue of *Nature Biotechnology* reports the full genome sequence of the white-rot fungus *P. chrysosporium* (Martinez et al. 2004). [Cover and caption used by permission from *Nat. Biotechnol.*, www.nature.com/nbt/]

Table 2. Cellulosic Ethanol Challenges, Scale, and Complexity

Research and Analytical Challenges	Scale and Complexity
• Screening of databases for natural variants of cellulases (generally glycosyl hydrolases) and other enzymes or molecular machines in metabolic networks; and characterization of variants • Analysis of modified variants to establish design principles and functional optimization • Modeling and simulation of cellulase, sugar transport, and multiple sugar-fermentation processes and systems • Integration of processing steps into single microbes or stable cultures	• Thousands of variants of all enzymes; screening of millions of genes, thousands of unique species and functions • Production and functional analysis of potentially thousands of modified enzymes, hundreds of regulatory processes and interactions • Models at the molecular, cellular, and community levels incorporating signaling, sensing, regulation metabolism, transport, biofilm, and other phenomenology and using massive databases in GTL Knowledgebase • Incorporation of complete cellulose-degradation and sugar-fermentation processes into microbes or consortia—hundreds of metabolic, regulatory, and other interconnected pathways

Market demand for these high-value alternatives can generate financial returns that make biorefineries commercially competitive, providing a viable base for lower-value products such as transportation fuels. An example is the polylactic acid now being produced by the Dow-Cargill venture, providing a biodegradable polymer that can be used in a variety of applications from carpets to clothes (Littlehales 2004; www.natureworkspla.com). After meeting market opportunities for high-value products, other lower-value products now derived from fossil fuel would be produced, ultimately leading to mass marketing of fuels from biomass resources. In developing microbial systems for supporting energy applications, a useful consideration would be systems that can produce, for example, ethylene, benzene, vinyl chloride, adipic acid, and, ultimately, the full suite of industrial chemicals derived from fossil fuels.

For more information, see www.bioproducts-bioenergy.gov/pdfs/BioProductsOpportunitiesReportFinal.pdf, www.metabolix.com, www.eere.energy.gov, and www.bio.org.

2.3.2. Biophotolytic Hydrogen Production

The importance of microbial systems in developing biological solutions to the energy challenge has been recognized by the National Research Council and the National Academy of Engineering (NAE). In a report on the hydrogen economy, the NAE Committee on Alternatives and Strategies for Future Hydrogen Production and Use recommended that DOE "refocus its biobased program on more fundamental research on photosynthetic microbial systems to produce hydrogen from water at high rate and efficiency" and "make use of important breakthroughs in molecular, genomic, and bioengineering research." [*The Hydrogen Economy: Opportunities, Costs, Barriers, and R&D Needs,* NAE (2004)] (See also Fig. 1, p. 23; Fig. 5. Two Approaches of GTL Research, this page; and Table 3. Biophotolytic Hydrogen: Goals and Impacts, p. 30.)

2.3.2.1. Science and Technology Objectives

As the planet's dominant photosynthetic organisms, microbes are capable of using solar energy to drive the direct conversion of water to hydrogen and oxygen (biophotolysis). One technology option for deploying biophotolytic hydrogen-production systems would involve the use of living organisms. Extensive farms of sealed enclosures (photobioreactors) containing photosynthetic microbes would split water to produce hydrogen for collection; oxygen would be released as the only by-product. Another deployment option would involve engineering artificial systems that would use natural or designed enzyme catalysts to yield

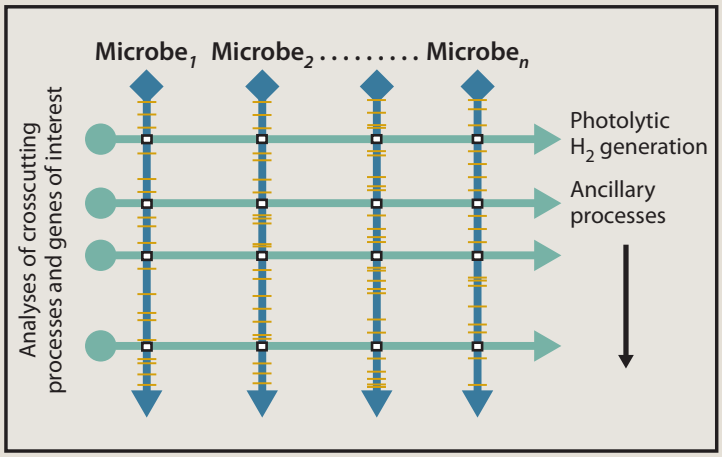


Fig. 5. Two Approaches of GTL Research. As outlined in Tables 2, 4, 6, and 8, research on GTL problems will take two general forms. The example illustrated above depicts analysis of the functionality of microbes and microbial communities. A vertically integrated approach will examine all relevant processes with the goal of consolidating the data to achieve a predictive whole-systems understanding.

To analyze such processes as hydrogen production, biomass conversion, or contaminant transformation, families of individual molecular assemblies and pathways will be studied horizontally across a wide set of natural variations. This approach will achieve insights into design principles and the range of variations. In addition to the primary process, all relevant ancillary processes (e.g., signaling, transport, and regulation) will require similar analyses. The eventual goal is to optimize these processes, either in living systems or captured in, for example, inorganic nanosystems.

hydrogen in vitro. Enzymes that split water are fixed to a synthetic nanostructure that maintains optimal conditions for hydrogen production. In addition to biophotolytic hydrogen production, other microbial processes that generate hydrogen (e.g., fermentation of biomass and nitrogen fixation) should be developed.

Photolytic production of hydrogen uses metabolic steps that are part of the photosynthetic process of microbes. Biological systems are designed for growth, functioning, and survival—producing biomass for structure, energy, and function. We must learn how to convert such a process to one that generates hydrogen in continuous water splitting—thus subverting the microbe's natural goal—to achieve our energy objective. To accomplish this, molecular and process models will be developed for enzyme, pathway, and whole-systems design (see Table 4. Biophotolytic Hydrogen Production Challenges, Scale, and Complexity, this page).

Screening of natural environments, particularly those under extreme conditions where hydrogen production might contribute to the success of microbial communities, may identify novel enzymes with desirable properties and new metabolic pathways that generate hydrogen. We must learn what makes biophotolysis possible in oxygenic photosynthetic microbes (e.g., algae and cyanobacteria that split water to generate oxygen); understand the principles underlying the natural range of hydrogenase properties, relevant metabolic processes and pathways involved in hydrogen production, regulatory processes that inhibit hydrogen overproduction, electron transfer-rate limitations, and competing pathways. Reducing the oxygen sensitivity of hydrogenases is a needed breakthrough that will require the redesign of multiple metabolic-network elements. Other phenomena to be understood include reverse reactions, efficiency of light utilization, and ways in which the organism or its component processes can be manipulated to increase the efficiency and yield of hydrogen production. Thousands of natural and modified hydrogenases and supporting pathways will be analyzed for relevant mechanisms, desirable properties, and insight into design principles. These capabilities will enable production systems engineering and allow frontier science to be imported quickly into technologies.

For more information, refer to Summary Table, p. 40; 5.0. Facilities Overview, p. 101; and Appendix A. DOE Mission: Energy Security, p. 197.

Table 3. Biophotolytic Hydrogen: Goals and Impacts

- Sunlight and water, two resources in virtually limitless supply, can be used to produce the ultimate fuel and energy carrier, hydrogen. High-efficiency use of hydrogen in fuel cells can generate electricity directly with water as the by-product.
- This energy cycle is carbon free and can be developed as the complement to the electric grid for all energy applications—industrial, transportation, and residential.
- Development of biological photolytic processes to produce hydrogen at high rates and efficiency will enable the establishment of a hydrogen-economy strategy based on a renewable source.

Table 4. Biophotolytic Hydrogen Production Challenges, Scale, and Complexity

Research and Analytical Challenges	Scale and Complexity
• Database screening for and characterizing of natural variants of hydrogenases and other enzymes and molecular machines in the entire set of pathways that underlie this process • Analysis of modified variants to establish design principles for functional optimization of the overall process including oxygen sensitivity, reverse reactions, transport, light capture, and conversion efficiency • Modeling and simulation of photolytic systems to support systems design and optimization	• Screening of millions of genes, thousands of unique species and functions, and thousands of variants of all enzymes • Production and functional analysis of potentially thousands of modified enzymes, hundreds of regulatory processes and interactions • Models at the molecular, cellular, and community levels incorporating signaling, sensing, regulation, metabolism, transport, and other phenomenology and using massive databases in GTL Knowledgebase

2.4. Environmental Remediation: Develop Biological Solutions for Intractable Environmental Problems

DOE is committed to remediating the large volumes of soil, sediments, and groundwater contaminated with metals, radionuclides, and a variety of organics at diverse defense facilities and sites across the nation (see Table 5. Bioremediation: Goals and Impacts, this page). As an example of the problem's scope, about 5700 individual contaminant plumes, some quite extensive, are known to be present at DOE sites (Linking Legacies 1997). One plume at Savannah River extends over 7.8 km^2, and an 18-km^2 plume exists at Hanford. Examples of the volume of contaminated soils and sediments at the Nevada Test Site and Fernald alone are 1.5 and 0.71 million m^3, respectively. Additionally, unknown quantities of waste are buried at numerous places. Projected costs for restoring these sites and disposing of wastes is $142B (Closure Planning Guidance 2004). Although DOE has the goal of completing the remediation of 108 of 114 contaminated sites by 2025 (DOE Strategic Plan 2003), the 6 remaining to be addressed after 2025 are the most challenging, and successful remediation will require development and deployment of innovative methods (see Fig. 6. A Legacy of Hazardous Waste, this page).

Although comparisons of the cost and effectiveness of metal and radionuclide bioremediation (the focus of the DOE effort) with traditional methods are not available, costs savings for bioremediation of organics are

Table 5. Bioremediation: Goals and Impacts

- Understand and incorporate the effects of biological processes into computer models describing the fate and transport of contaminants in the environment. This knowledge could result in savings in the billions of dollars by supporting decisions to take advantage of natural attenuation alternatives, use bioremediation for previously intractable problems, or improve the efficiency of conventional technologies.
- Develop new or improved bioremediation strategies and technologies to save potentially billions of dollars over traditional treatments. Bioremediation may offer solutions in previously intractable cases (i.e., where there was no solution at any price).
- Develop new suites of biosensors and performance assessment and monitoring techniques to track progress of environmental cleanup strategies and optimize operation of current cleanup techniques.

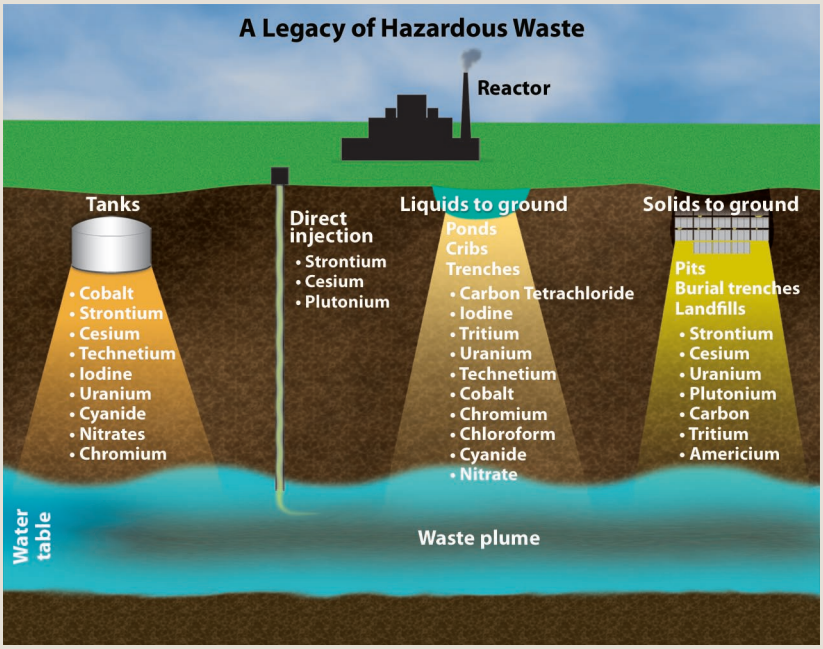


Fig. 6. A Legacy of Hazardous Waste. For more than 50 years, the United States created a vast network of facilities for research and development, manufacture, and testing of nuclear weapons and materials. The result is subsurface contamination on more than 7000 sites at over 100 facilities across the nation, more than half of which contain metals or radionuclides and most with chlorinated hydrocarbons. Biologically based techniques can provide cost-effective restoration strategies for many of these sites.

estimated to range from 30 to 95%. In addition, in situ bioremediation, taking advantage of natural microbial populations in the subsurface, has the potential for reducing costs and increasing the efficiency of groundwater treatment as compared to conventional pump-and-treat technology. Given that over 1 billion m^3 of water and 55 million m^3 of solid media at DOE sites in 29 states are contaminated with radionuclides (Linking Legacies 1997), potential savings accrued by use of innovative biotechnologies are likely to amount to billions of dollars (Bioventing Performance 1996; Patrinos 2005; Scott 1998).

2.4.1. Science and Technology Objectives

In conjunction with the capabilities of other science programs, GTL science will facilitate detailed, large-scale discovery and investigation of microbes with important contaminant-transformation capabilities. DOE bioremediation strategies and biogeochemistry research focus on using natural microbial communities to reduce the mobility and toxicity of metals and radionuclides. The interdependent metabolic survival strategies used by microbial communities can directly or indirectly remove contaminants from groundwater or transform toxic contaminants into benign chemical products. For example, *Shewanella* and *Geobacter*, two model microbes currently being studied in GTL and BER's Environmental Research Sciences Division (ERSD) projects, can enzymatically reduce certain toxic contaminants. This capability transforms, for example, Uranium(VI), which is soluble and moves in groundwater, to Uranium(IV), which is insoluble and precipitates out of the groundwater as a biologically unavailable solid (see 3.3.4. Sidebars Illustrating Details of Specific Research, p. 58). Studying these model microbes is a first step in expanding our understanding of the structure, function, metabolic activity, and dynamic nature of microbial communities and their role in influencing subsurface geochemistry. This knowledge is needed to predict microbe-mediated contaminant fate and transport and to develop efficient bioremediation strategies (Fredrickson and Balkwill 2005; Croal et al. 2004; Madsen 2005; Ben-Ari 2002; Gold 1992; Spear et al. 2005; Nealson 2005).

A biotreatment technique that works well at one site may perform poorly at another because microbial communities, geochemical properties, and flow regimes frequently differ markedly between sites. We often lack understanding of how microbial processes are coupled to other processes influential in contaminant behavior and are scaled in heterogeneous environments. In addition, we need new tools for measuring key microbial, geochemical, hydrological, and geological properties and processes in these systems. Less than 1% of all microorganisms collected at only a few sites have been cultured and characterized in any great detail, and only a small fraction of those have been sequenced. Even less is known regarding the interactions of microorganisms in communities.

Subsurface microbial communities can be quite distinct from soil and ocean communities, with far lower microbial densities and unique genetic traits. The metabolic processes observed in the subsurface are often the result of unique interactions—in these "geologically powered dark ecosystems"—between the microbial community and subsurface geochemistry (Nealson 2005). We have only begun to appreciate the existence of such systems, let alone understand them so that we can take advantage of their diverse capabilities (Gold 1992; see sidebar, The Microbial World, p. 13).

In this complex venue, we first must define the genomic potential of microbial communities (see Table 6. Bioremediation Challenges, Scale, and Complexity, p. 33). Whereas historically our studies have been limited to microbes that can be cultured in the laboratory, the combination of metagenomics with the production and characterization of proteins from genes now allows culture-independent insights into microbial function. Though difficult to obtain, if environmental or cultured samples are available, then functional determination can include information on microbial responses to environmental stimuli as captured in, for example, gene and protein expression and metabolite analyses. In addition, new imaging techniques augmented by fluorescent probes with molecular resolving power will allow the analysis of individual cells and processes in complicated community and geochemical venues.

These results will form the basis for evaluating and modeling pathways of such cellular processes as signaling, growth, and response to contaminants. Other processes of importance in modifying contaminant trans-

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

To accomplish this linkage, we first need a cohort of trained scientists to determine the makeup of subsurface microbial communities and their interactions with the geochemical environment. Methods also must be developed for incorporating GTL genome-based microbial knowledge into meaningful field-scale models. Some of these techniques already are being generated within BER ERSD environmental-restoration programs and GTL (www.science.doe.gov/ober/ERSD_top.html). For more information, refer to Table 5, p. 31; Table 6, this page; Summary Table, p. 40; 5.0. Facilities Overview, p. 101; and Appendix B. DOE Mission: Environmental Remediation, p. 215.

Environmental Remediation

- **Mission Science Goals: 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 (see 3.0. GTL Research Program, p. 41).**
- **Challenges: Bioremediation will require understanding biogeochemical processes from the fundamental-molecular to community levels to describe contaminant-transformation processes coinciding with simulated changes in microbial-community composition and structure.**

2.5. Microbial Roles in Carbon Cycling and Sequestration: Understand Biosystems' Climate Impacts and Assess Sequestration Strategies

Marine and terrestrial ecosystems each play major roles in the global cycling of carbon. Microbes are essential to maintaining the planet's ability to sustain life, including recycling most of earth's biomass, both assimilating and respiring large amounts of carbon dioxide—many times that of anthropogenic CO_2 emissions (Doney et al. 2004; Falkowski et al. 2000; Field et al. 1998; Johnston et al. 2004; Hess 2004; see Fig. 7. Simplified Global Carbon Cycle, p. 34, and 2.6. Summing Up the Challenges, p. 37). Small changes in these natural fluxes induced by climate change or natural processes could overwhelm any attempts at mitigation we might make within global energy systems, some of which might be very costly.

Table 6. Bioremediation Challenges, Scale, and Complexity

Research and Analytical Challenges	Scale and Complexity
• Analysis of microbial communities and their metabolic activities that impact the fate and transport of contaminants • Analysis of geochemical changes in subsurface environments due to microbial or chemical activity • Accurate conceptual and quantitative models for coupling and scaling microbial processes to complex heterogeneous environments	• Hundreds of different sites, millions of genes, thousands of unique species and functions • Functional analysis of potentially thousands of enzymes involved in microbe-mineral interactions; hundreds of regulatory processes and interactions; spatially resolved community formation, structure, and function; influence on contaminant fate • Models at the molecular, cellular, and community levels incorporating signaling, sensing, metabolism, transport, biofilm, cell-mineral interactions; incorporated into macromodels for fate and transport

These global ecosystems are affected by climate change (Climate Change Science Program 2003). Ascertaining relationships among such natural and managed carbon pools as agricultural soils, forests, and atmospheric CO_2 and determining the resultant climate effects will make important new contributions to climate models. Such models will aid in understanding the long-term sequestration capacities of these pools. Microbes as primary mediators of earth's elemental cycling also can serve as indicators of ecosystem health and change as we monitor the impacts of climate (see Table 7. Carbon Cycling and Sequestration: Goals and Impacts, this page; sidebar Ocean Monitors, p. 234; and Tringe et al. 2005).

Table 7. Carbon Cycling and Sequestration: Goals and Impacts

- Improved understanding of key feedbacks and sensitivities of biological and ecological systems and accelerated incorporation into climate models will reduce uncertainties in assessments of climate change.
- Knowledge of the carbon cycle will allow evaluation of carbon-sequestration strategies and alternative response options.
- Development of sensors and monitoring techniques and protocols will allow use of these sensitive ecosystems as sentinels for the effects of climate change.

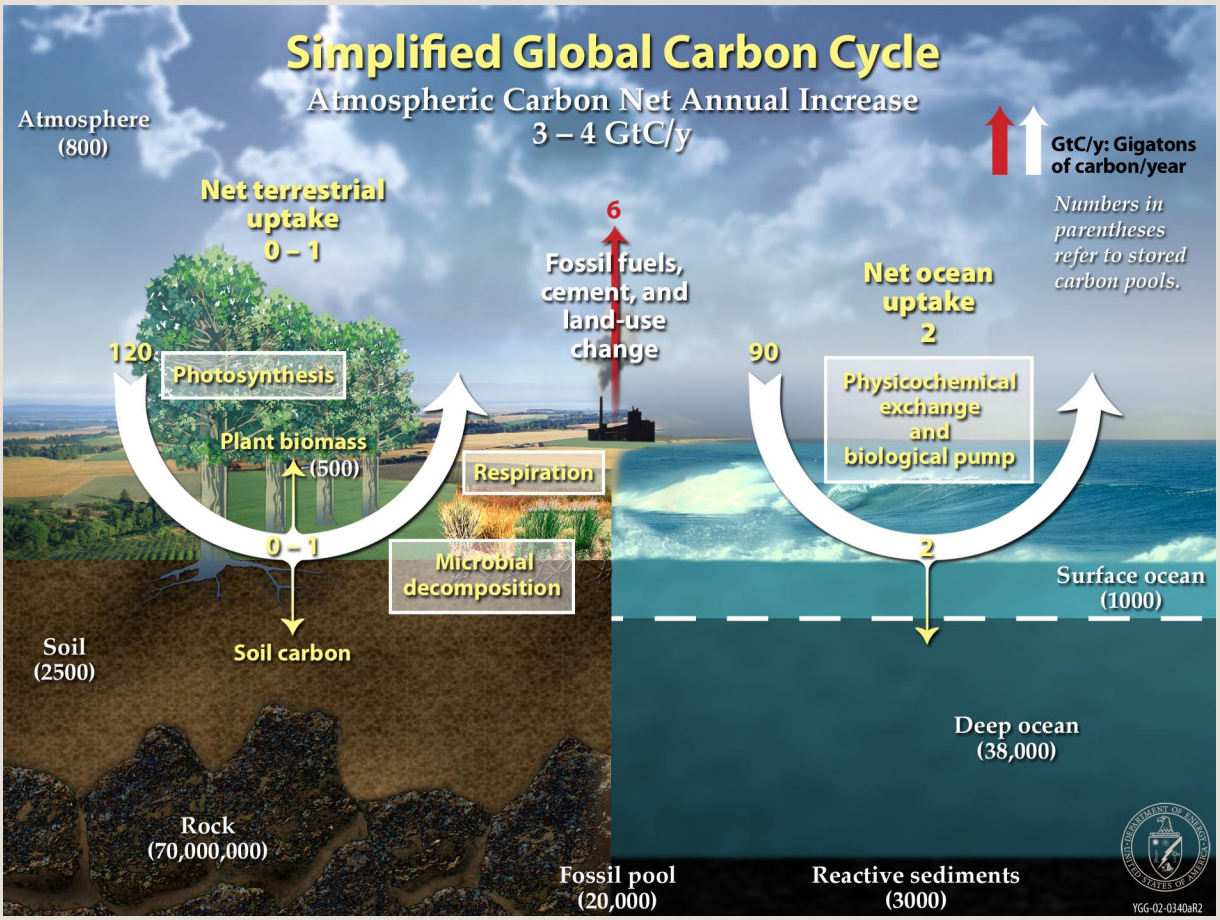


Fig. 7. Simplified Representation of the Global Carbon Cycle. The illustration depicts human-induced changes relative to the total cycle. [Graphic adapted from *Carbon Sequestration Research and Development* (1999).]

2.5.1. Science and Technology Objectives

The science undertaken by the GTL program will provide a systems-level understanding of microbial processes essential to carbon cycling in ocean and terrestrial environments, as well as the cycling of such other elements as nitrogen, phosphorous, sulfur, oxygen, and metals. Detailed knowledge revealed by GTL about the mechanistics and functions of microbial processes and communities will be incorporated into global climate-change models to provide a robust science base for evaluating potential impacts of proposed carbon-management strategies. Moreover, GTL ultimately will help determine ways in which microbial environments might be used or manipulated to enhance carbon residence time in ocean and soil ecosystems without harming those ecosystems.

Carbon Cycling and Sequestration

- **Mission Science Goals: 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 (see 3.0. GTL Research Program, p. 41, and Appendix C. DOE Mission: Carbon Cycling and Sequestration, p. 227).**
- **Challenges: We are just beginning to understand the genetic and functional diversity of ocean and terrestrial ecosystems. They potentially contain millions of microbial species organized in extensive communities. We must understand both the global and molecular mechanistic behaviors of these large systems.**

Because they have been transforming the planet for some 3.5 billion years, microbes play a dominant role in ecosystem processes and are key mediators of energy transfer and materials cycling in the biosphere. Microbes cycle immense volumes of carbon: they can fix CO_2 by light-driven (photoautotrophy) and geochemically driven (lithoautotrophy) reactions, generate methane, produce CO_2 during the decomposition of organic matter, precipitate carbonate minerals, and catalyze the polymerization of plant polymers into recalcitrant pools of soil organic matter. Microbes perform all their activities in dynamic communities—in the upper ocean layer where photosynthesis occurs, removing carbon from and largely returning it to the atmosphere; and in terrestrial communities where microbes and fungi cycle nutrients containing carbon and other elements and decompose biomass. In these systems, we do not envision any redesign of microbes; rather, we seek an understanding of natural processes to enable predictions about microbial evolution and to support wise management practices, including the addition of nutrients where appropriate. As with the subsurface modeling of contaminant transport and fate, this problem involves the very complex linking of the time and length scales of molecular and microbial processes to the millennial and global scales important for carbon sequestration and climate models.

2.5.2. Marine Microbial Communities

In the oceans, microbes are the primary photosynthetic organisms, producing most oceanic organic materials and constituting the foundation of the marine food chain. The photosynthesis of such phytoplankton as diatoms, dinoflagellates, and cyanobacteria converts about as much atmospheric carbon to organic carbon as does plant photosynthesis on land (Fuhrman 2003). Large oscillations in phytoplankton abundance, therefore, can impact greatly the ocean's ability to take up atmospheric carbon. Although most organic matter produced in surface waters is consumed by other microorganisms and returned rapidly to the atmosphere as carbon dioxide, diatoms are capable of synthesizing organically complexed carbon that can be carried to the ocean depths. Because carbon cycling is considerably slower in the deep oceans than in surface waters (thousands to millions of years), this carbon is effectively sequestered. In this way, diatoms may sequester more carbon than all the earth's rainforests combined (Armbrust 2004). In addition to understanding photosynthetic assimilation of atmospheric carbon carried out by diatoms and other phytoplankton, GTL and other BER carbon-cycle research will provide understanding of microbial processes that degrade organic matter in the ocean's depths and ultimately return carbon to the atmosphere.

2.5.2.1. Specific Scientific Needs for Marine Microbial Communities

GTL studies, based on genomic and metagenomic sequences of a broad range of microbes from diverse marine environments, must ascertain the molecular design principles of photosynthetic and related nutrient and metabolic systems. Paramount to understanding the dynamics of carbon-assimilation pathways, these principles will provide the foundation for new hypotheses about the types and diversity of pathways and capabilities of individual species. A key element of these analyses is to determine whether the very broad genetic diversity translates into a commensurate range in function and the environmental significance of that range.

The fact that many environmental microbes remain uncultivated and that we lack good experimental models for most microbes requires the development of high-throughput capabilities for ultimately gaining this information from genome sequence alone. Investigating the natural dynamics of relationships among microbial, biogeochemical, and physical processes requires new high-throughput sampling and analysis tools. Capabilities for explorations of critical biological reactions, including photosynthetic modes and various metabolic pathways, also must be established. The resultant molecular understanding can lead to the development of increasingly sophisticated microsensors that can detect changes in the levels of biomolecules (DNA, mRNA, proteins, metabolites) and serve as indicators of microbial-community response to environmental stressors. Detailed mathematical models that encompass full-systems phenomenology will form the basis for including these processes in climate and integrated-assessment models to inform policy and technology decisions (see sidebar, Integrated Assessment Program, p. 236).

2.5.3. Terrestrial Microbial Communities

Terrestrial ecosystems fix CO_2 directly from the atmosphere into biomass, mainly via plant photosynthesis. Carbon in terrestrial ecosystems can be stored in plant biomass or soils. Microbial communities can influence both areas of terrestrial carbon storage in a variety of ways. In the narrow zone of soil surrounding the root (the rhizosphere), microbial interactions with sugars, amino acids, enzymes, fatty acids, and other organic compounds exuded from roots can significantly impact plant growth and development. Microbes can enhance plant growth by providing nutrients such as phosphorous and nitrogen or by suppressing plant pathogens in the soil. Some microbial populations are beneficial to plant growth, while others have neutral and even harmful effects. Identifying metabolic requirements and environmental factors that can give an advantage to beneficial microbes, therefore, is important. A better understanding is needed of molecular mechanisms that enable microbes to colonize root surfaces, interact with plant exudates, and compete or cooperate with other soil organisms.

In addition to promoting plant growth, microbes can return CO_2 rapidly to the atmosphere as carbon dioxide or transform root exudates and decaying plant materials into humic acids with varying degrees of recalcitrance to degradation. Some of the most stable soil organic compounds have carbon-turnover times of hundreds of years.

Shifts in the microbial decomposition of organic matter to CO_2 can occur during environmental stresses such as climate change (King et al. 2001). For example, plant death due to temperature shift, water stress, or disease can promote microbial decomposition, seriously complicating the use of standing biomass or soil organic matter for reducing the amount of atmospheric CO_2. Although forest clearing and farmland tillage have significantly reduced soil carbon content, the amount currently in soils accounts for 75% of the carbon contained in the terrestrial biosphere. Carbon-depleted soils represent a large potential reservoir (50-Gt one-time gain) that could be used to mitigate atmospheric carbon emissions (Rosenberg, Izaurralde, and Malone 1999).

2.5.3.1. Specific Scientific Needs for Terrestrial Microbial Communities

Systems biology will support a biological understanding of interactions among terrestrial ecosystems and changes in atmospheric composition and the climate system (see Table 8. Carbon Cycling and Sequestration

Table 8. Carbon Cycling and Sequestration Challenges, Scale, and Complexity

Research and Analytical Challenges	Scale and Complexity
• Analysis of ocean and terrestrial microbial-community makeup and genomic potential • Analysis of carbon and other cycling processes » Photosynthesis and respiration in oceans » Storage and decomposition in soil: Microbial, fungal, and plant communities • Modeling and simulation of microbe biogeochemical systems	• Thousands of samples from different sites consisting of millions of genes, thousands of unique species and functions • Functional analysis of enzymes involved—potentially tens of thousands; hundreds of regulatory processes and interactions; spatially resolved community formation, structure, and function • Models at the molecular, cellular, and community levels incorporating signaling, sensing, metabolism, transport, biofilm, and other phenomenology into macroecosystem models

Challenges, Scale, and Complexity, this page). We must elucidate microbial contributions to carbon transformation in soils and assess the potential for sequestering meaningful amounts of carbon (gigatons per year) in more stable forms. We seek to understand the genomic-mechanistic basis for adaptations made by microbial species to climate change. We also must determine biological feedbacks to the climatic system brought about through the terrestrial carbon cycle and microbial contributions to carbon transformation in soils. This knowledge will enable us to design and assess sequestration strategies and make crucial contributions to overall climate-modeling efforts (King et al. 2001). GTL technical requirements for analyses of ocean and terrestrial systems are very similar.

- Defining communities and their collective genetic functional potential requires single-cell and community sequencing (in situ and in vitro), systems biology studies, and the ability to relate microbial activities to soil processes.
- We need to understand processes that impact production of GHGs (carbon dioxide, methane, nitrous oxide, dimethyl sulfide); correlate biomolecular inventories with environmental conditions; characterize microbial system interactions with soils, rhizosphere, and plants; and functionally image microbial activities (e.g., proteome and metabolome) at cellular and community levels.
- We need the ability to predict microbial responses to manipulations of plant inputs to the carbon cycle, to human inputs to soils, and to other environmental changes.
- GTL seeks to employ microbes as sentinels of change in environmental conditions; measures of change include microbial responses (biomarkers) including an array of biological components (e.g., RNAs, proteins, metabolites, and signaling); community genomic makeup; functional assays; and environmental conditions including carbon and nutrient inventories.

For more details see Appendix C. DOE Mission: Carbon Cycling and Sequestration, p. 227.

2.6. Summing Up the Challenges

Each of these mission areas poses grand scientific and technical challenges for biology. Each is confronted with the tremendous genetic and functional diversity of microbial systems in widely varying environments (note scale and complexity in Tables 2, 4, 6, 8; see sidebar, Deciphering the Scale and Complexity of Global Microbial Communities, p. 39, including Table 9. Microbial Community Characteristics in Diverse Earth Environments, p. 39). Each also requires an unprecedented level of understanding to be able to predict systems behaviors or to engineer systems for energy or other applications. The Summary Table, p. 40, presents a capsule summary of systems being studied, mission goals that drive the analysis, generalized science roadmaps, and outputs to DOE missions. Science capabilities that GTL will use to implement these roadmaps and achieve a systems-level understanding of microbial processes are described in the following chapters and in the three mission appendices beginning on p. 197.

MISSIONS OVERVIEW

Genomics has opened the door to the study of these complex natural systems and processes, but the science also has revealed an unforeseen diversity, bringing the realization that we know the functions of only a small fraction of newly discovered genes. Clearly, we must develop the capabilities to move beyond the parts list provided by genomics to understand microbes so well that we can predict their behavior. While each of the systems presents a daunting challenge, the fact that they all obey the same set of principles and share enduring genetic and functional traits means that studying them together will have a synergistic effect. Researchers can use the same set of tools and concepts to study all these systems, and the knowledgebase amassed by GTL will inform new investigations in all realms. But to achieve timely impacts in mission problems, to deal with the complexities of microbial systems, and to practice "systems biology," we must develop technological and computing capabilities with dramatically improved performance, throughput, quality, and cost. The following chapters in this roadmap describe the science and technology goals and milestones to achieve these gains including computing, technologies, and facilities.

Deciphering the Scale and Complexity of Global Microbial Communities

At the frontier of modern biology, microbes in their vast natural communities are immensely important to our planet's future. DOE missions lead us to the study of ocean, terrestrial, and subsurface microbial communities whose existence in disparate environments, genetic diversity, and complexity of function have been appreciated only recently. With the advent of genomics and systems biology, the myriad capabilities of these organisms now can be revealed by deciphering the interactions of millions of genes with physicochemical variables. Understanding their specialized biochemical capabilities and their contributions to nutrient cycling, to the global carbon cycle, and to overall ecosystem function could lead to new energy sources and cost-effective strategies for remediating the environment and sequestrating carbon.

Although the three environments and resulting microbial biochemistries differ greatly, the basic processes by which microbes operate (based on DNA) are the same and can be investigated using the same set of core principles and technologies. Studying a broad range of microbial systems in one program at a scale commensurate with practical national and global needs will enhance the richness of the science and its impact on our understanding of all living systems. The table below compares and contrasts these microbial communities in their different ecosystems, their metabolic processes and structures, and other characteristics that make studying microbes one of the great challenges and opportunities of 21st Century science.

Table 9. Microbial Community Characteristics in Diverse Earth Environments

Topic	Oceans (water column)	Terrestrial (surface soils)	Deep Subsurface (ocean, terrestrial)
Energy Sources and Ecologies	Primary photosynthesis (i.e., microbes at base of food chain)	Plant photosynthesis, plant-rhizosphere-microbe symbiosis (microbes are decomposers)	Reduced inorganic compounds, simple communities, lack of predators, minimal gene exchange
Energy and Materials Storage	Rapid carbon and nutrient turnover, carbon exported to depths	High resident carbon (plant roots, microbes, soil organic matter)	Minerals, microbial biomass, and fossil organic carbon
Key Processes	Ocean carbon biological pump, C, N, P, O cycles	Soil carbon cycle, symbiosis, C, N, P, O cycles	Redox manipulation of subsurface to transform contaminants
Rates of Cell Division (and Metabolism)	Relatively rapid (hours to weeks)	Relatively long (months to 100s of yrs)	Very long (decades to 100s of yrs)
Microbial Population Size	Medium density	High density	Low density
Key Environmental Variables	pH, light, nutrients (e.g. P, Fe), dissolved organic matter, mixing, temperature, currents	Plant community, organic matter content and composition, mineral composition, temperature, moisture	Mineralogy, hydrology, trace elements, gas composition, groundwater geochemistry, temperature
Science Goals	Understand microbes and communities and flow of carbon and nutrients (C, N, P, S, O, metals)	Understand microbe-rhizosphere symbioses, carbon flows and lifetimes, and nutrient cycles (C, N, P, S, O, metals)	Mechanistic understanding of electron-transfer processes, utilization of solid-phase nutrients (minerals, rocks)
Application Goals	Predict responses of ocean C cycling communities to climate scenarios	Predict responses of soil C cycling communities to perturbations and manipulations	Prediction of geomicrobial fate of metals and radionuclides in the subsurface

Summary Table. GTL Science Roadmap for DOE Missions

GTL Science Roadmap for DOE Missions

		DOE Mission Goals	GTL Science Roadmaps
Selected Processes	Biofuels	**Processes to convert cellulose to fuels** • Understanding and improving cellulase activity • Improving sugar transportation and fermentation to alcohols • Integrated processing **Microbial processes to convert sunlight to hydrogen fuels** • Understanding photolytic fuel production • Designing photosynthetic biofuel systems	*Science Objectives* **Characterize genes, proteins, machines, pathways, and systems** • Conducting genomic surveys and comparisons • Mining natural systems for new functions • Producing and characterizing proteins • Analyzing interactions, complexes, and machines **Understand functions and regulation** • Measuring molecular responses: Inventories • Performing functional assays **Develop predictive mechanistic models** • Conducting experimental design • Designing and manipulating molecules • Using cellular and cell-free systems
Selected Processes	Environmental Remediation	**Microbial processes to reduce toxic metals** • Understanding microbe-mineral interactions • Devising restoration processes	*Mission Outputs* **Systems engineering** • System-design strategies for deployment • Living and extracellular systems • Validation and verification analyses
Natural Systems' Behavior	Environmental Remediation	**Subsurface microbial communities' role in transport and fate of contaminants** • Understanding fate and effects • Supporting remediation decisions	*Science Objectives* **Analyze communities and their genomic potential** • Sequencing and comparing genomes • Screening natural systems for processes • Producing and characterizing proteins **Understand community responses, regulation** • Comparing CO_2, nutrients, biogeochemistry cycles • Producing cellular and community molecular inventories • Performing community functional assays **Predict responses and impacts** • Building interactive and predictive models • Applying natural and manipulated scenarios
Natural Systems' Behavior	Carbon Cycling and Sequestration	**Ocean microbial communities' role in the biological CO_2 pump** • Understanding C, N, P, O, and S cycles • Predicting climate responses • Assessing impacts of sequestration **Terrestrial microbial communities' role in global carbon cycle** • Understanding C, N, P, O, and S cycles • Predicting carbon inventories and climate responses • Assessing sequestration concepts	*Mission Outputs* **Robust science base for policy and engineering** • Model ecosystem response to natural events • Efficacy and impacts of intervention strategies **Sensor development**

A capsule summary of systems being studied, mission goals that drive the analysis, generalized science roadmaps, and outputs to DOE missions. To elucidate design principles, each of these goals entails the examination of thousands of natural primary and ancillary pathways, variants, and functions, as well as large numbers of experimental mutations.

3.0. GTL Research Program

3.1. Background and Approach 42

3.1.1. Phase I Implementation: Current GTL and Related Projects 43

3.2. Scientific Goals and Milestones 43

3.2.1. Missions Science Goals 44

3.2.2. Science and Technology Milestones 44

3.2.2.1. Milestone 1: Develop Techniques to Determine the Genome Structure and Functional Potential of Microbes and Microbial Communities 44

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 47

3.2.2.3. Milestone 3: Develop the Knowledgebase, Computational Methods, and Capabilities to Advance Understanding and Prediction of Complex Biological Systems 51

3.2.2.4. Milestone 4: Design and Build User Facilities to Accelerate GTL Microbial Systems Biology 55

3.3. Highlights of Research in Progress to Accomplish Milestones 55

3.3.1. Research Highlights for Milestone 1: Sequences, Proteins, Molecular Complexes 56

3.3.2. Research Highlights for Milestone 2: Cell and Community Function, Regulation, and Dynamics 56

3.3.3. Research Highlights for Milestone 3: Computing 58

3.3.4. Sidebars Illustrating Details of Specific Research 58

3.4. GTL Program and Facility Governance 77

3.4.1. Facility User Access 77

3.4.2. Collaborative Environment 77

3.4.3. Facility Governance 78

3.5. Training 78

3.6. Ethical, Legal, and Social Issues (ELSI) 79

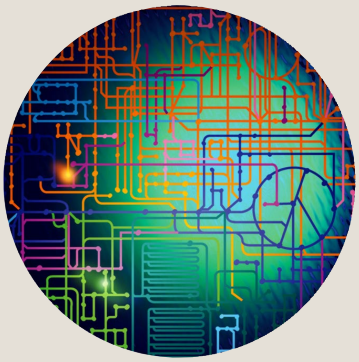
3.6.1. GTL Commitment to Explore ELSI Impacts 79

3.6.1.1. Examples of Potential GTL ELSI Issues 79

3.6.1.2. Using Microbial Diversity for Practical Applications 79

3.6.2. The Path Forward 80

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, the Bioenergy Research Center, will concentrate on overcoming biological barriers to the industrial production of biofuels from biomass and on other potential energy sources such as biophotolytic decomposition of water to hydrogen. 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://doegenomestolife.org/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 knowledge-base 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.

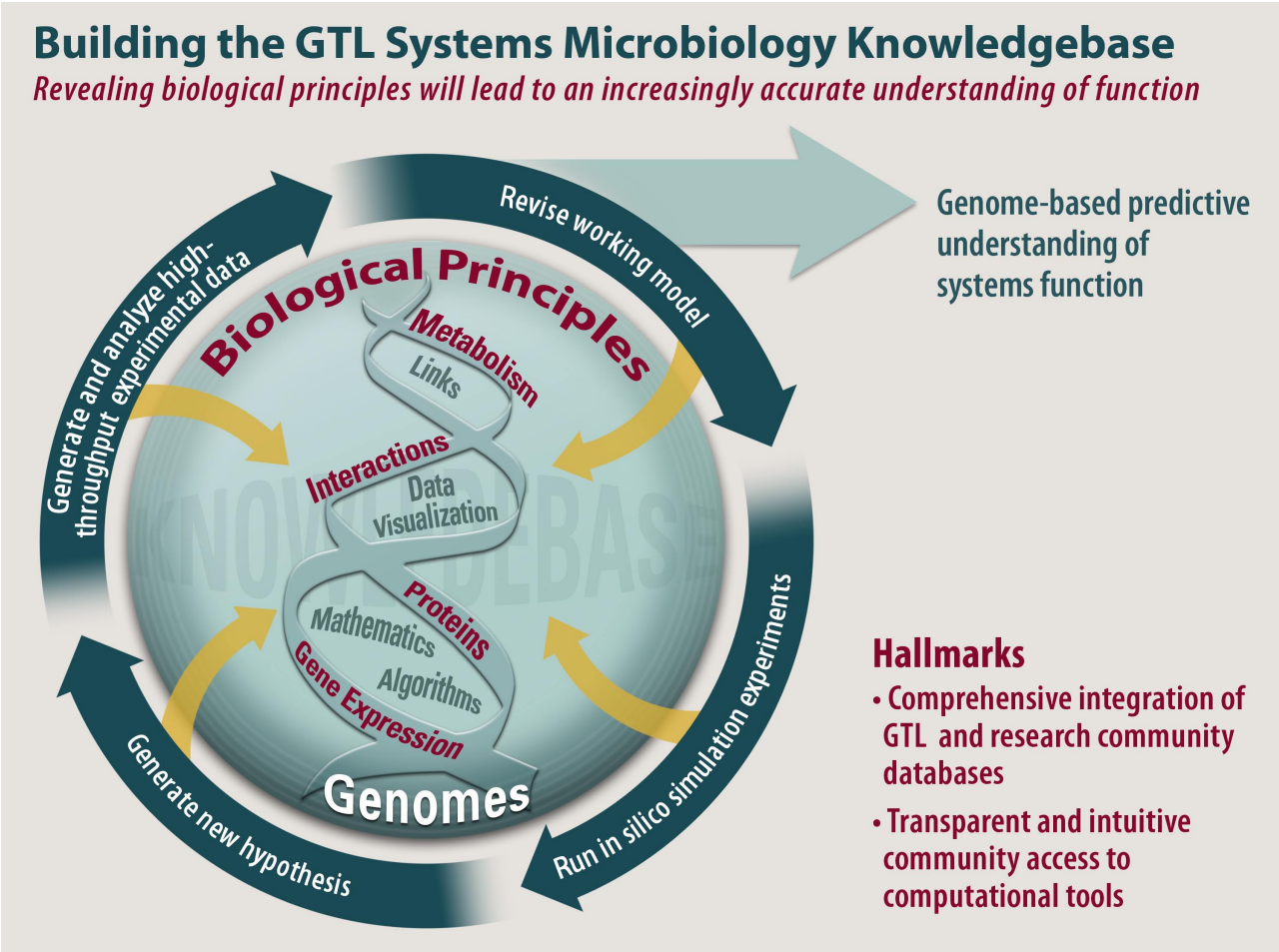
Models can form the foundation for understanding complex systems. They can be applied to such useful ends as developing biological sources of clean energy, cleaning up toxic wastes, and understanding the roles of microbial communities in ocean and terrestrial carbon cycling (i.e., how they sequester 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



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.

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/

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

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

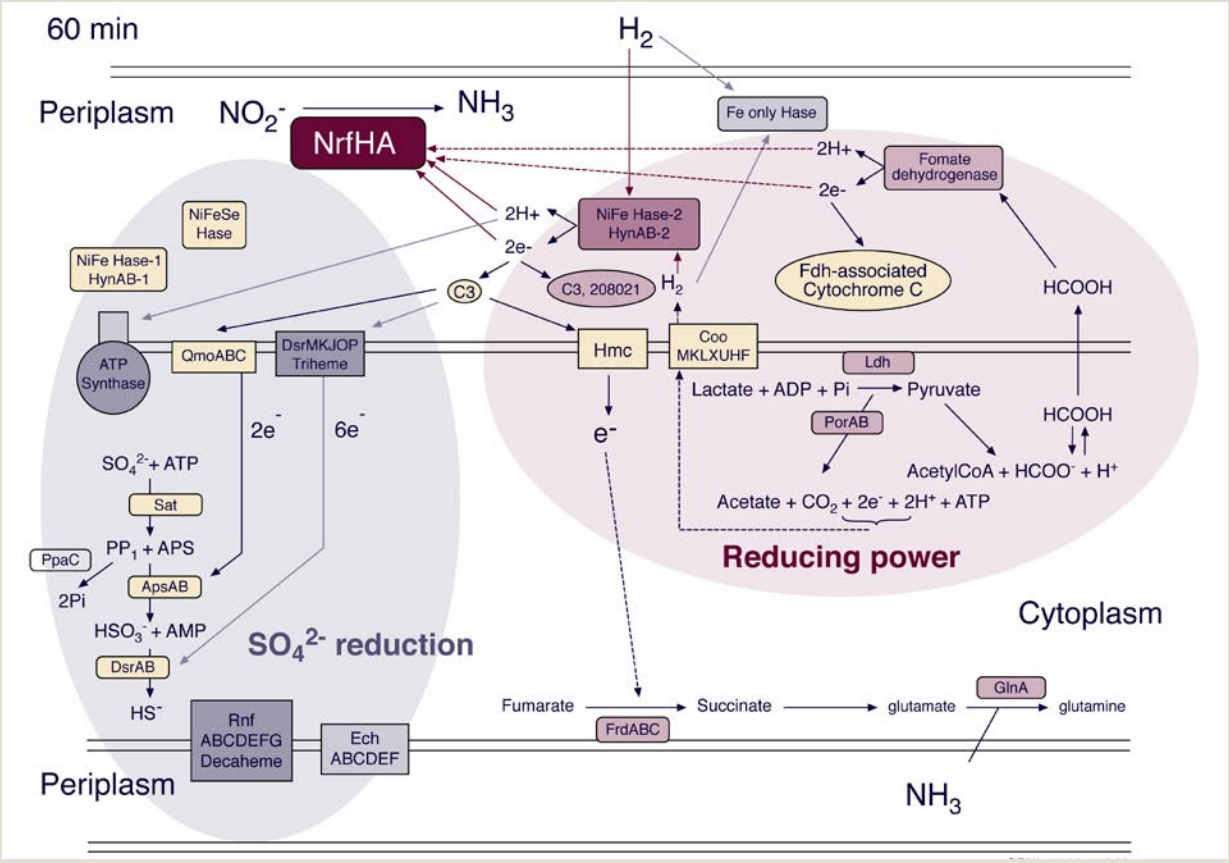


Fig. 1. Schematic Depiction of the Nitrate Response of *D. vulgaris*. Red systems are upregulated and blue systems downregulated.

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.

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

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]

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).

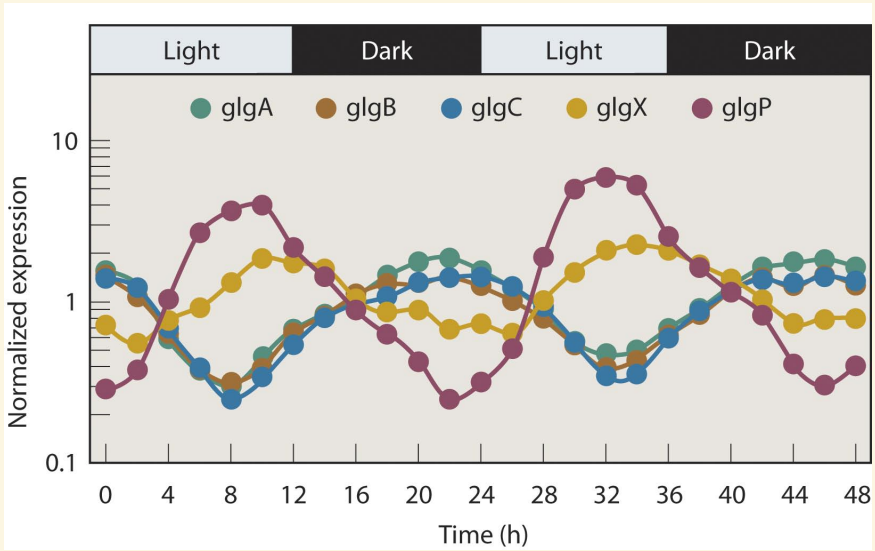
6. Whole-genome shotgun sequencing project data from Iron Mountain and the Sargasso Sea available on the web (www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/sargasso.html).

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.



E. R. Zinser, Massachusetts Institute of Technology

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]

Transfer of Photosynthetic Genes Between Bacteria and Phages

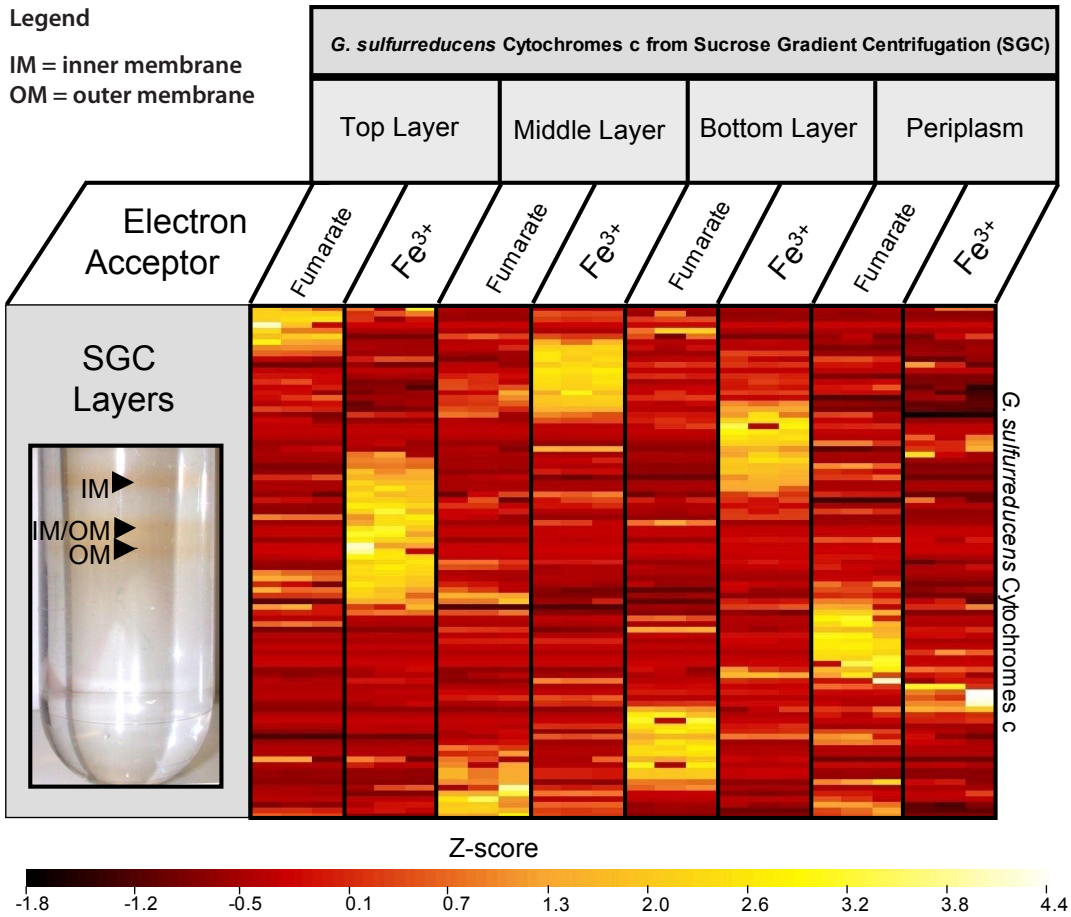
Viruses (phages) infecting the oceanic cyanobacteria *Prochlorococcus* are thought to mediate population sizes and affect the evolutionary paths of their hosts. GTL researchers analyzed genomes from three *Prochlorococcus* phages: a podovirus and two myoviruses. They appear to be variations of two well-known 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.

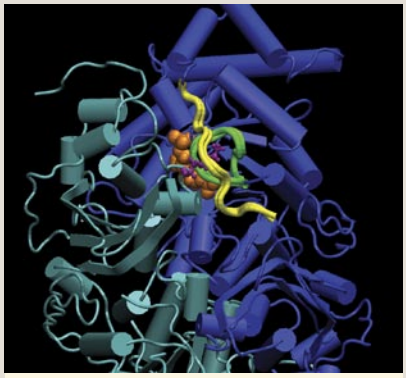
New Imaging and Computational Tools Enable Investigations of Carbon Cycling in Marine Cyanobacteria

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.



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.

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]

Reference

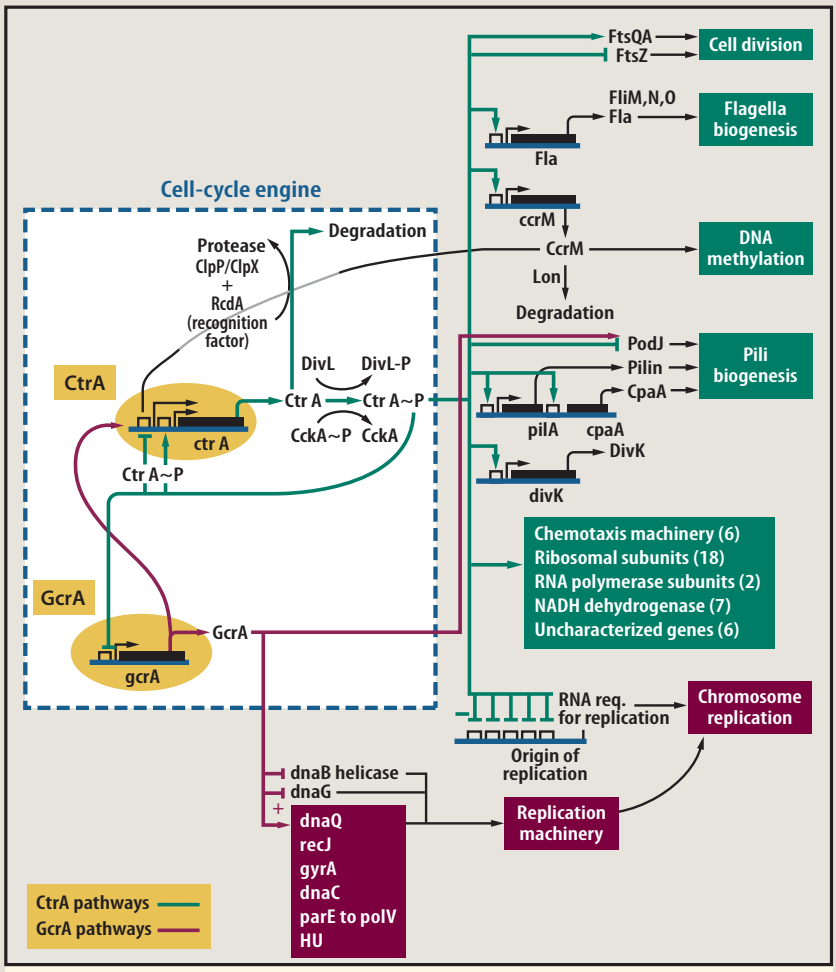
M. B. Sinclair et al., "Design, Construction, Characterization, and Application of a Hyperspectral Microarray Scanner," *Appl. Optics* 43, 2079–88 (2004).

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

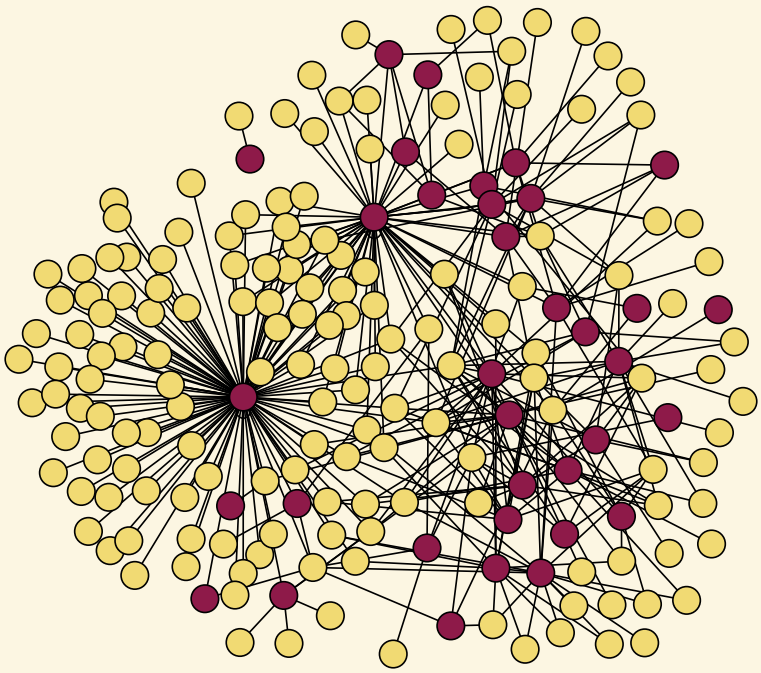
1. J. Holtzendorff et al., "Oscillating Global Regulators Control the Genetic Circuit Driving a Bacterial Cell Cycle," *Science* **304**, 983–7 (2004).

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

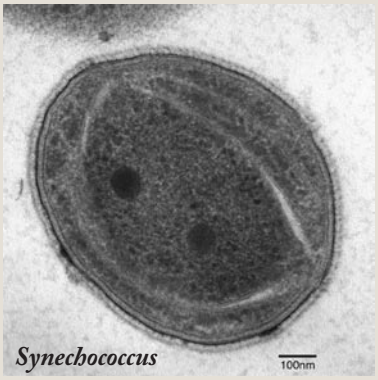
1. F. W. Larimer et al., "Complete Genome Sequence of the Metabolically Versatile Photosynthetic Bacterium *Rhodopseudomonas palustris*," *Nat. Biotechnol.* **22**, 55–61 (2004).

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.



J. McCarren, Scripps Institution of Oceanography

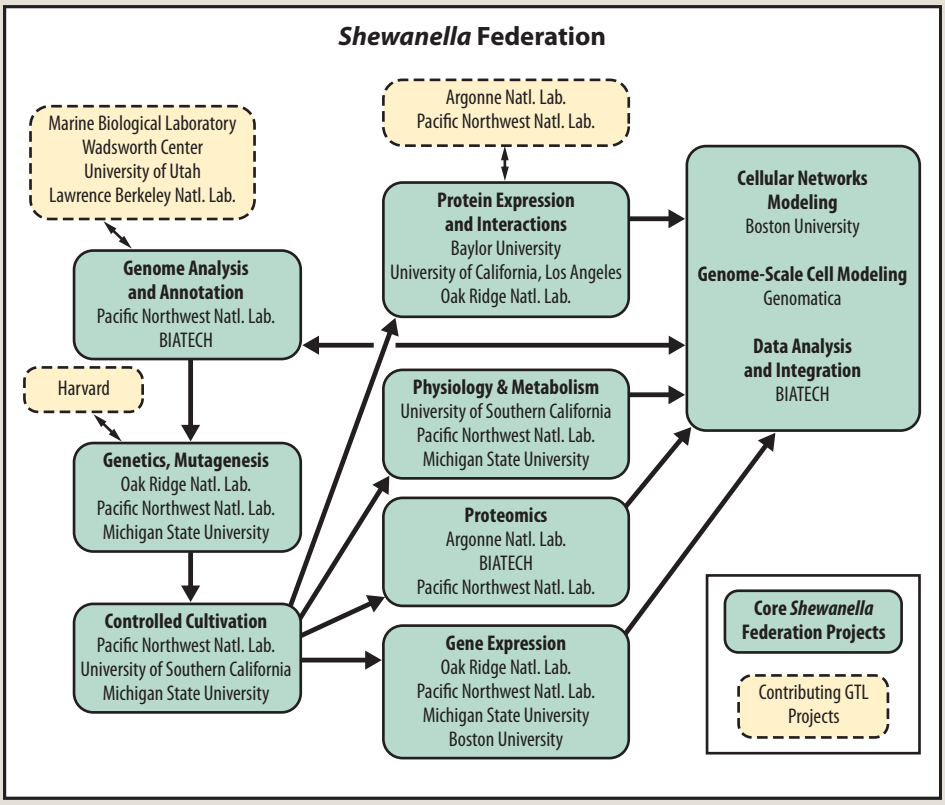
These resources are enabling biologists to combine knowledge and see relationships that previously were obscured by the distributed nature and diverse data types present in biological databases. GTL researchers are using the tools to create knowledgebases for other organisms as well (e.g., *R. palustris*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

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.

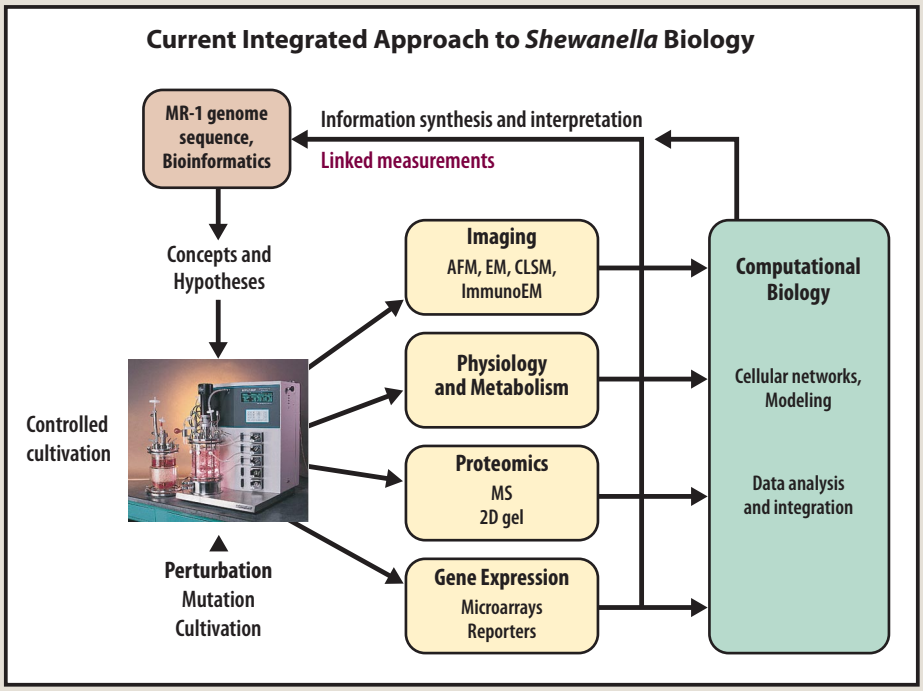
Combining Computational and Experimental Approaches to Enhance *Shewanella* Genome Annotation

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.



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.

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

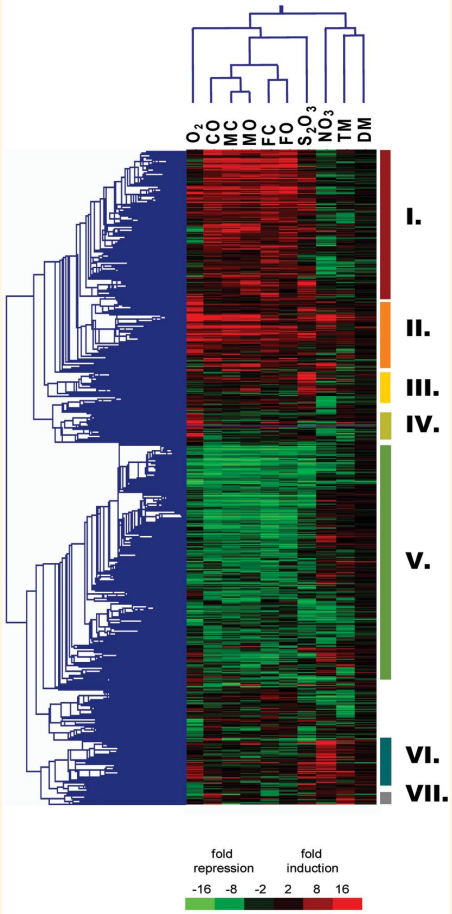
E. Kolker et al., "Global Profiling of *Shewanella oneidensis* MR-1: Expression of Hypothetical Genes and Improved Functional Annotations," *Proc. Natl. Acad. Sci. USA* **102**, 2099-2104 (2005).

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

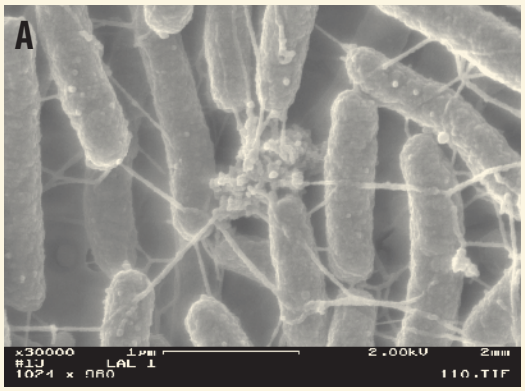


Fig. A. SEM MR1 Grown in a Bioreactor under EAL Conditions. Sample was prepared by critical-point drying.

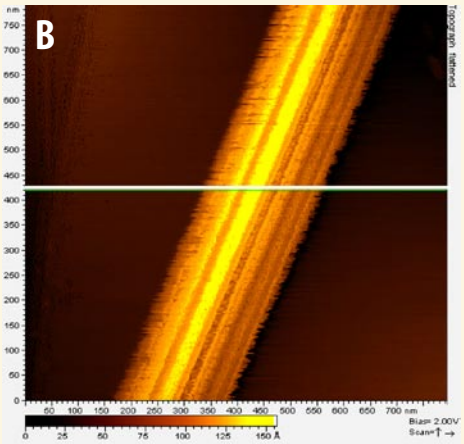


Fig. B. STM Image of Isolated Nanowire from Wild-Type MR-1. Nanowire has lateral diameter of 100 nm and topographic height of 5 to 10 nm. High magnification shows ridges and troughs running along the structure's long axis.

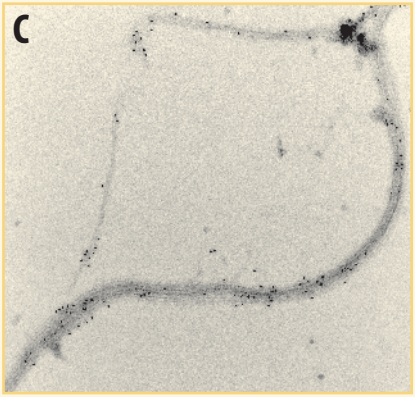


Fig. C. Immunogold Labeling of MtrC on Nanowires. TEM images of whole mounts of MR-1 nanowires from cells grown in continuous culture under EAL conditions reveal that MtrC is localized specifically to the nanowires.

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]

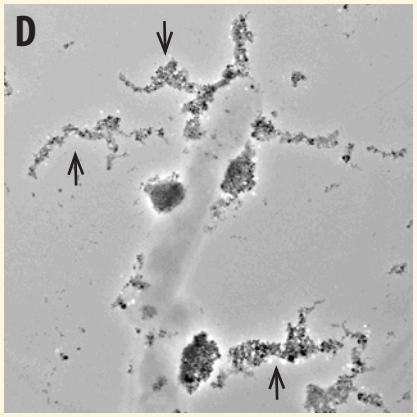


Fig. D. Magnetite Associated with Nanowires. TEM images of whole mounts of MR-1 cells incubated with the Fe(III) mineral ferrihydrite revealed the formation of nanocrystalline magnetite along the nanowires (indicated by arrows).

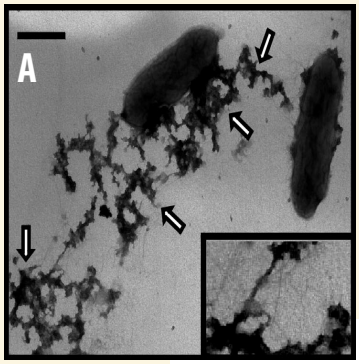
Reference

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.



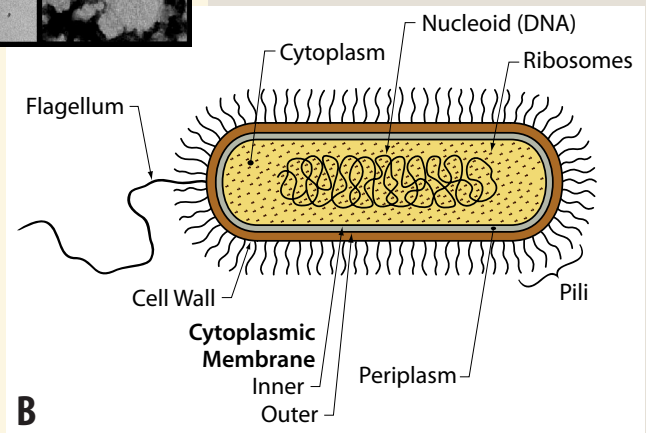


Fig. B. View of Simplified Microbial Anatomy.

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

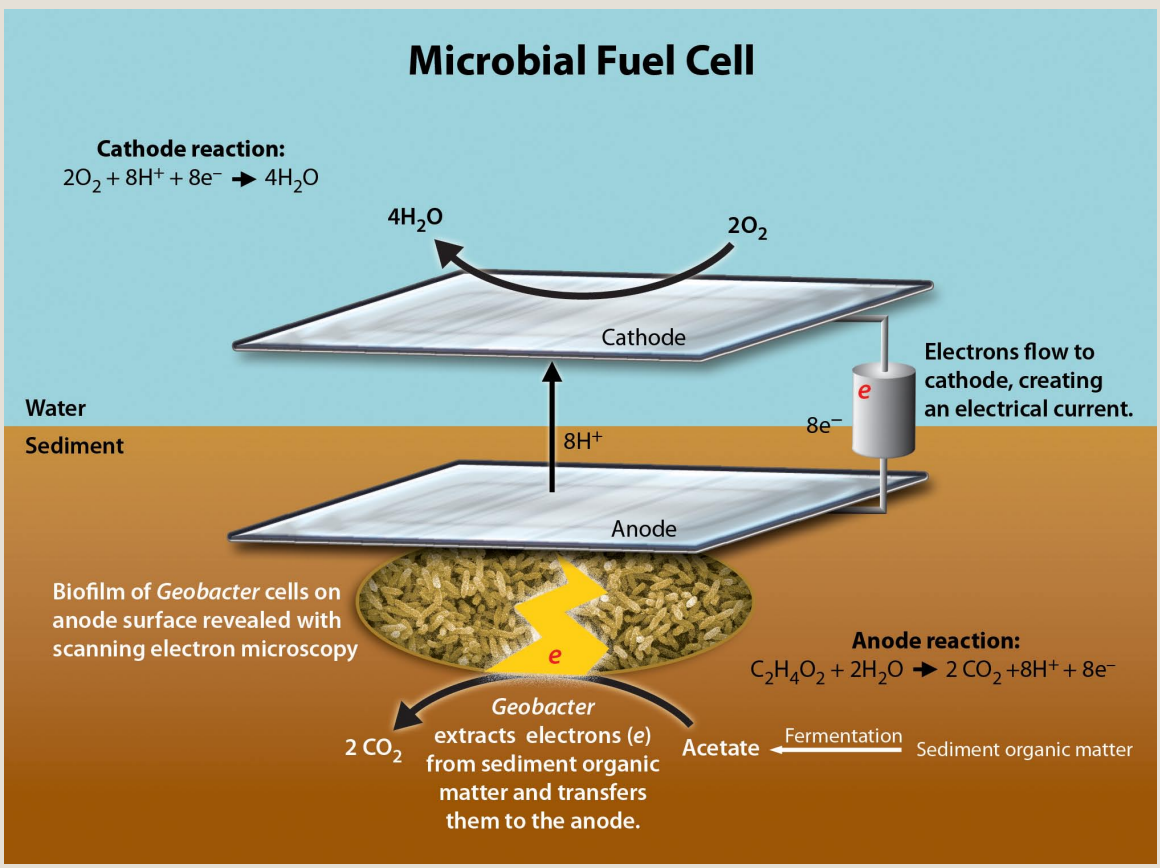
Microorganisms known as "electricigens" can efficiently convert organic wastes, renewable biomass, and even mud into electricity and harmless by-products. This capability offers the potential for using microbes (or their components) to generate electricity at low cost while transforming industrial, domestic,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



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.

4.0. Creating an Integrated Computational Environment for Biology

4.1. An Essential Foundation 82

4.2. Capabilities for an Integrated Computational Environment 85

4.2.1. Theory, Modeling, and Simulation Coupled to Experimentation of Complex Biological Systems 85
4.2.1.1. Microbial Behavior: Modeling at the Molecular Level 87
4.2.1.2. Computer Science and Mathematics Challenges 87
4.2.1.3. Fundamental Questions and Issues 87
4.2.1.4. Chemistry Challenges 87
4.2.1.5. Structure, Interactions, and Function 88
4.2.1.6. Microbial Behavior: Metabolic Network and Kinetic Models of Biochemical Pathways 88
4.2.1.6.1. Current State of Cell-Network Modeling: Moving from Experiment (Real Life) to Simulation (Abstract Systems Model) 88
4.2.1.6.2. Advanced Modeling Capabilities 89
4.2.1.6.3. Crosscutting Research and Development Needs 90

4.2.2. Sample and Experimental Tracking and Documentation: Laboratory Information Management System (LIMS) and Workflow Management 91
4.2.2.1. LIMS Impact 91
4.2.2.2. LIMS Requirements for GTL 91

4.2.3. Data Capture and Archiving 92

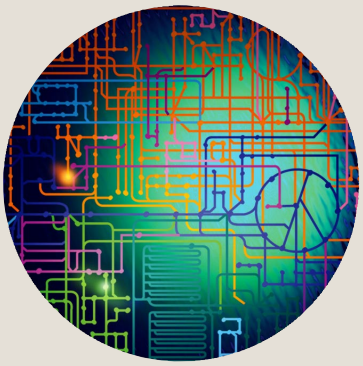
4.2.4. Data Analysis and Reduction 93
4.2.4.1. Infrastructure 94
4.2.4.2. Examples of Analyses and Their R&D Challenges for GTL Science 95

4.2.5. Computing and Information Infrastructure 96

4.2.6. Community Access to Data and Resources 97
4.2.6.1. Capabilities Needed 98
4.2.6.2. Some R&D Challenges 98

4.2.7. Development Requirements 99

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, the Bioenergy Research Center, will concentrate on overcoming biological barriers to the industrial production of biofuels from biomass and on other potential energy sources such as biophotolytic decomposition of water to hydrogen. 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://doegenomestolife.org/biofuels/).



Creating an Integrated Computational Environment for Biology

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

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

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

4.1. An Essential Foundation

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

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

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

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

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

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

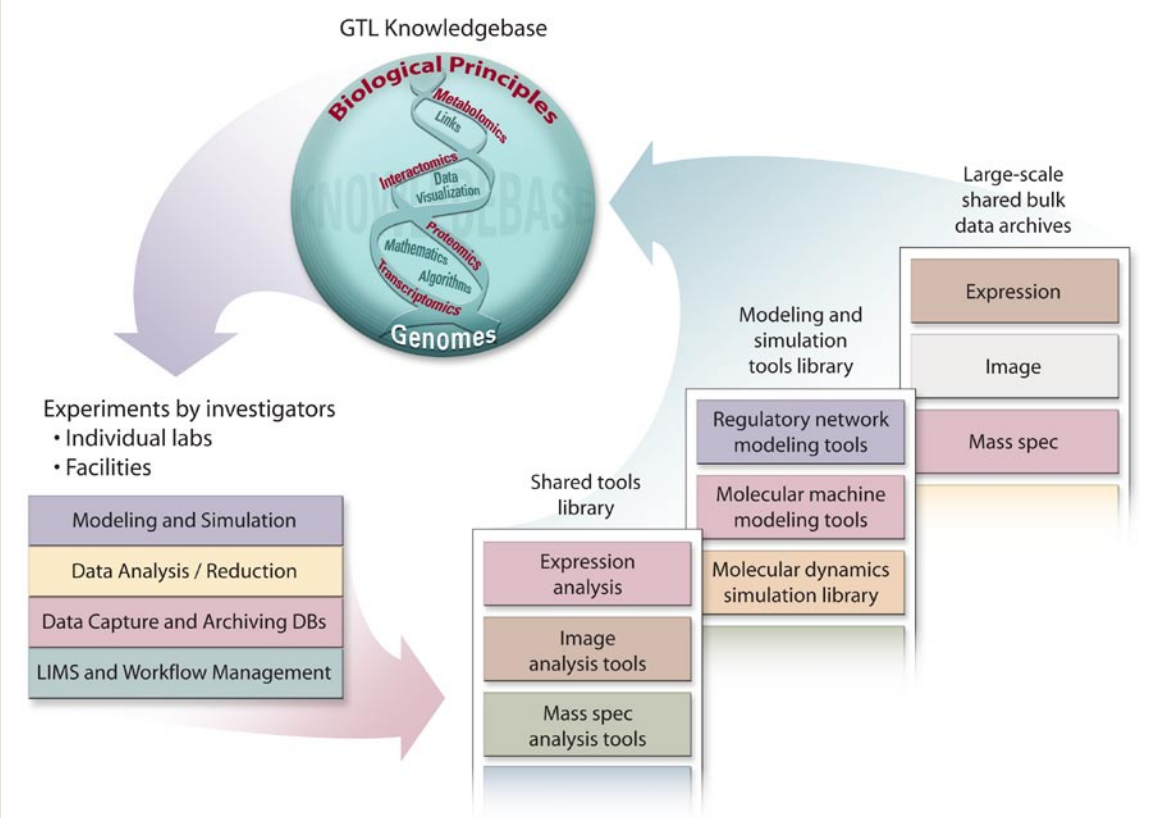


Fig. 1. GTL Integrated Computational Environment for Biology: Using and Experimentally Annotating GTL's Dynamic Knowledgebase. At the heart of this infrastructure is a dynamic, comprehensive knowledgebase with DNA sequence code as its foundation. Offering scientists access to an array of resources, it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

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

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

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

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

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

Scientific Discovery Through Advanced Computing

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

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

Two of the largest and most-successful SciDAC projects are

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

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

4.2. Capabilities for an Integrated Computational Environment

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

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

These capabilities are described more fully in the following text.

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

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

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

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

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

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

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

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

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

Examples of Biological Understanding and Possible Applications Enabled by TMS

Advancing Understanding of Biological Subsystems via Modeling and Simulation

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

Simulation Examples

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

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

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

4.2.1.1. Microbial Behavior: Modeling at the Molecular Level

The starting point for GTL analysis is to decipher microbial processes at the molecular level. The centerpiece of GTL is the ability to analyze, reconstruct, and model the networks of molecular interactions at the core of life processes. Cell networks arise from the series or chains of molecular interactions during metabolism, protein synthesis and degradation, regulation of genetic processes such as transcription and replication, and cell signaling and sensing. In short, cellular molecular networks and pathways are at the center of cell modeling and cellular behavior and, ultimately, of microbial-community modeling and behavior. Such models would predict how a cell's genome and environmental factors combine to yield its phenotype. Models will be powerful tools for scientific discovery as we explore the enormous complexity of microbes and their communities.

4.2.1.2. Computer Science and Mathematics Challenges

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

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

4.2.1.3. Fundamental Questions and Issues

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

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

4.2.1.4. Chemistry Challenges

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

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

4.2.1.5. Structure, Interactions, and Function

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

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

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

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

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

The primary goal of cell-network modeling is to capture in an abstract mathematical model the structure (topology), kinetics, and dynamics necessary to analyze and simulate the behavior of networks present in a particular organism. Models are constructed from a combination of mathematical principles and experimental data (e.g., from annotated genomes, proteomics databases, in vitro experiments, expression, and the historical literature). Models are used to facilitate a general understanding of cellular networks and for simulations that attempt to reproduce or predict a particular experimental result. When attempting to develop a systems understanding of complex biology, investigators will use simulation and modeling as one of a few ways to derive insight from complicated interactions involving numbers of variables and details that cannot be grasped intuitively.

Current state-of-the-art models can be used to make specific quantitative predictions for limited regions of well-characterized metabolic pathways or a limited set of specific regulatory or signaling circuits. Although

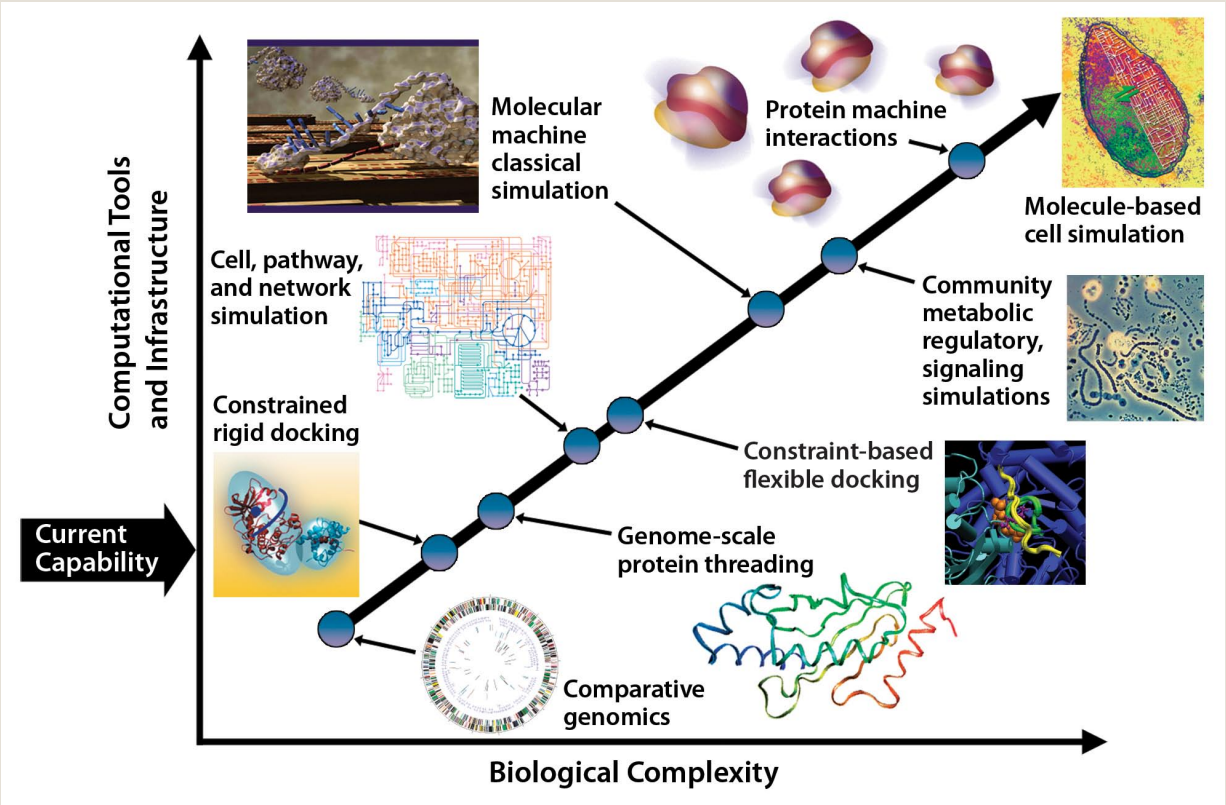
more-general qualitative predictions can be made for larger, more complete networks, the current lack of kinetic constants for most enzymes and of concentration data for intermediate metabolites limits the ability to simulate quantitative results for entire networks including cells and communities. Figure 2 illustrates current capabilities on the path from genome data to full cell simulation. Modeling also is hampered by the incomplete specification of networks due to lack of functional gene assignments, protein complex and association data, and data for regulatory elements and interactions. Bioinformatics techniques are used upstream of modeling and simulation to extract from experimental data the relationships and functions needed for simulation.

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

4.2.1.6.2. Advanced Modeling Capabilities

No dominant formalism, however, has emerged that can satisfactorily represent both the kinetics and dynamics of metabolic networks and the logical structure of signaling and regulation. Much new work is needed in this area. Another critical topic that must be addressed is how best to represent multiple levels of spatial and

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



temporal scales in cellular systems and incorporate them into models. Most models of cellular networks are one dimensional (1D) (e.g., box models that assume a completely mixed environment). To make progress toward the ultimate goal of accurate phenotype prediction, future modeling schemes need to incorporate 3D modeling and intracellular compartmentalization. Multiple modeling and inference techniques can address different classes of problems, each with distinct temporal and spatial scales and each with potentially different computational complexity. All classes of problems have specific data limitations and a diverse set of data sources, as mentioned above. Limitations on the models themselves depend on the levels of abstraction used and the mathematical treatment of the problem.

Compartmentalized models will become increasingly important for depicting distinct types and phases of metabolism in organisms such as cyanobacteria, which have both oxygenic and anoxic pathways separated either spatially or temporally. Compartmentalized models will be needed to fully describe life cycles of prokaryotes, which include mechanisms such as sporulation, heterocyst formation, and differentiation. Models with multiple compartments will have to address coupling of compartments (e.g., data and flux representations and stability and fidelity) in a scalable fashion. Much may be learned from the experiences of the DOE National Nuclear Security Agency's Accelerated Strategic Computing and the climate modeling community. Compartmentalization and coupling also will become an issue in multicellular systems (e.g., bacterial communities and multicellular organisms). A major modeling challenge is the choice and effective exploitation of mathematical abstractions. Biological systems differ from those produced by human engineering in that hierarchies or functional subsystem modules are not necessarily obvious, yet exploiting modularity or lumping the system may be essential for efficient modeling and simulation.

4.2.1.6.3. Crosscutting Research and Development Needs

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

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

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

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

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

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

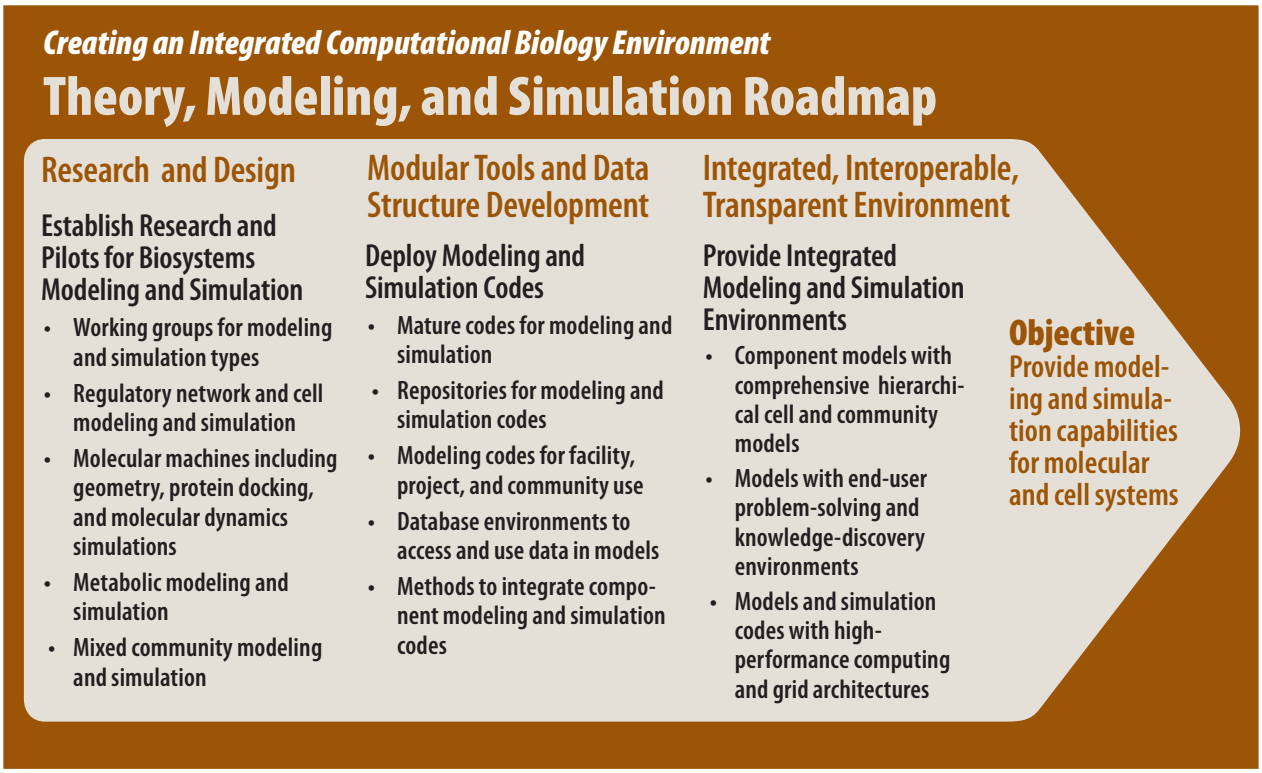
4.2.2.1. LIMS Impact

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

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

4.2.2.2. LIMS Requirements for GTL

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



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

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

4.2.3. Data Capture and Archiving

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

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

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

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

Creating an Integrated Computational Biology Environment

LIMS and Workflow Management Roadmap

Research and Design

Establish LIMS/Workflow Research and Pilots

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

Modular Tools and Data Structure Development

Deploy LIMS and Workflow Systems

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

Integrated, Interoperable, Transparent Environment

Integrate LIMS and Workflow

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

Objective

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

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

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

Key features of databases and structures include:

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

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

4.2.4. Data Analysis and Reduction

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

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

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

4.2.4.1. Infrastructure

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

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

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

Creating an Integrated Computational Biology Environment

Data Capture and Archiving Roadmap

Research and Design

Establish Research and Pilots for Archival Data Storage and Retrieval

- Working groups for data types
- Data representations, standards and ontologies for bulk data types; expression, imaging, mass spec
- Technologies for large-scale storage
- Preliminary design, pilots for storage archives
- Siting of archival storage systems

Modular Tools and Data Structure Development

Deploy Local and Shared Bulk Data Archives

- Mature design for bulk data archives
- Bulk archives for key large-scale data types; expression, image, mass spec
- Processes to link archives to data production
- Local facility data storage
- Archives linked to analysis-tool libraries

Integrated, Interoperable, Transparent Environment

Integrate Data Archives with User Environments

- Bulk archives linked to GTL community databases
- Integration with end-user problem-solving and knowledge-discovery environments
- Integration with high-performance computing analysis, modeling, and simulation environments

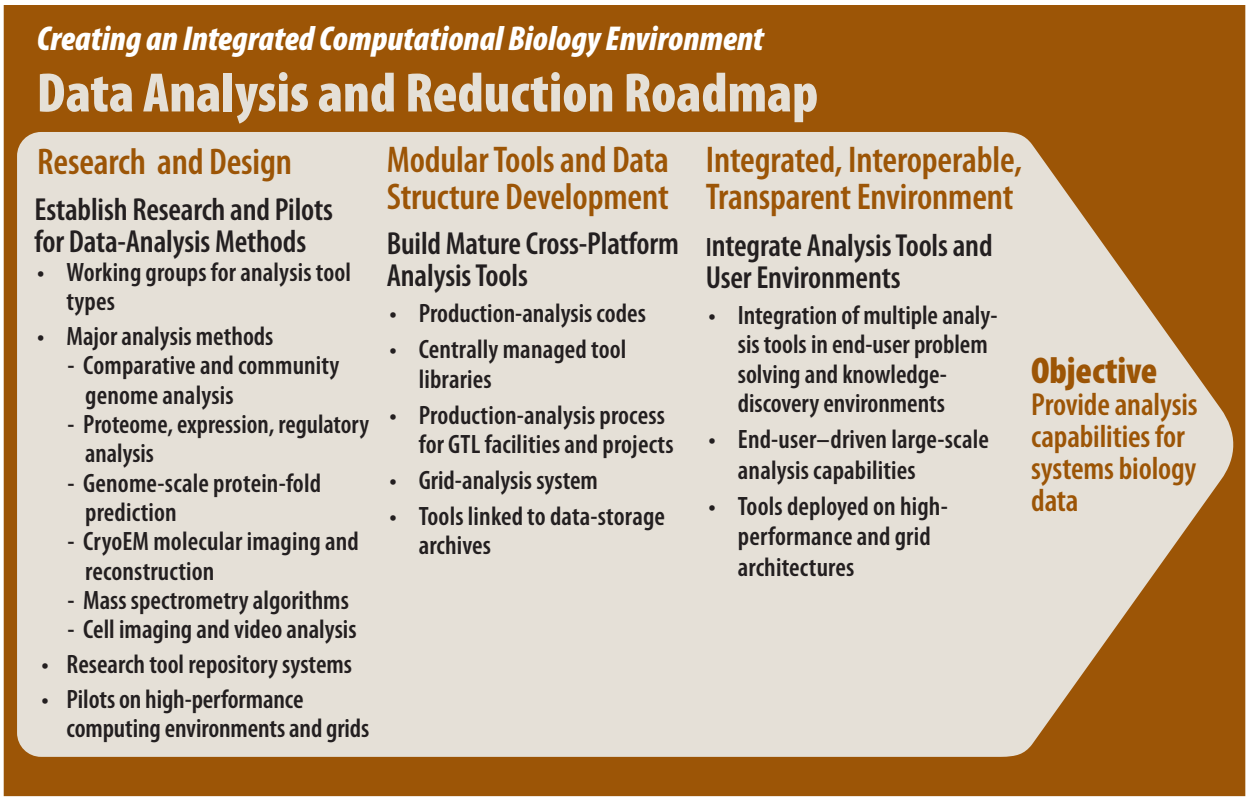
Objective
Provide data capture and large-scale storage and retrieval

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

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

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

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



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

4.2.5. Computing and Information Infrastructure

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

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

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

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

4.2.6. Community Access to Data and Resources

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

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

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

Creating an Integrated Computational Biology Environment

Computing and Information Infrastructure Roadmap

Research and Design

Establish Research and Pilots for Computing-Infrastructure Development

- Working groups to study computing infrastructure
- Storage and network requirements defined for GTL projects, user facilities, and community
- Pilots for large-scale storage facilities; selected sites
- Grid approaches to tools and data-sharing sites
- Computing requirements for analysis, modeling, and simulation on HPC* environments
- Tools and simulation codes evaluated on new computing architectures
- Sites selected for MPP† infrastructure

Modular Tool and Data Structure Development

Establish Large-Scale Grid, MPP,† and Storage Infrastructure

- Production storage archives
- Integration with GTL facility and project data production
- Production grid systems
- Processes for data mirrors, tool sharing, and grid process management
- Tools ported to selected HPC* sites
- HPC* integration with facility and GTL data-analysis and modeling needs

Integrated, Interoperable, Transparent Environment

Provide Integrated Infrastructure Environments

- Community biogrid infrastructure with comprehensive modeling and simulation codes, database mirrors, and appropriate architectures
- HPC* integration with large-scale data production, analysis, and modeling needs
- Continued evaluation of emerging MPP† computing architectures

Objective
Provide computing hardware, storage, network, and grid infrastructure

*HPC: high-performance computing
†MPP: massively parallel processing

4.2.6.1. Capabilities Needed

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

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

4.2.6.2. Some R&D Challenges

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

Creating an Integrated Computational Biology Environment

Community Access to Data and Resources Roadmap

Research and Design

Establish Research and Pilots for Data Representation, Modeling, and Ontologies

- Working groups to study data types
- Databases sited
- Data modeling, representations, and design for pathways, expression, imaging
- Data access and user-environment pilots
- Support of existing early-phase databases (e.g., microbial genomes)
- Ontology pilots
- Preliminary database design, pilots

Modular Tools and Data Structure Development

Deploy Component Databases

- Mature design for databases
- Individual databases for key data types: Expression, pathways, networks, machines, imaging
- Intermediate user environments for database access
- Plan for database integration

Integrated, Interoperable, Transparent Environment

Integrate Databases and User Environments

- Comprehensive integration of GTL databases
- Integration with problem-solving and knowledge-discovery environments for user applications
- Integration with community resources
- Integration with high-performance computing and modeling and simulation tools

Objective
Provide community data resources and user access to data, models, and simulations

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

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

4.2.7. Development Requirements

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

5. GTL Facilities

5.0. Facilities Overview 101

5.0.1. Science and Technology Rationale 102
5.0.2. A New Trajectory for Biology 104
5.0.3. Capsule Facility Descriptions 104
5.0.3.1. Facility for Production and Characterization of Proteins and Molecular Tags 104
5.0.3.2. Facility for Characterization and Imaging of Molecular Machines 105
5.0.3.3. Facility for Whole Proteome Analysis 105
5.0.3.4. Facility for Modeling and Analysis of Cellular Systems 106
5.0.4. Relationships and Interdependencies of Facilities 106
5.0.5. Research Scenarios 107
5.0.6. Facility Development 107

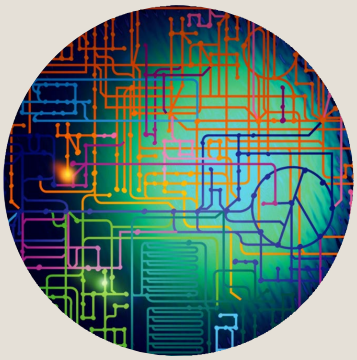
5.1. Facility for Production and Characterization of Proteins and Molecular Tags 111

5.2. Facility for Characterization and Imaging of Molecular Machines 139

5.3. Facility for Whole Proteome Analysis 155

5.4. Facility for Analysis and Modeling of Cellular Systems 173

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, the Bioenergy Research Center, will concentrate on overcoming biological barriers to the industrial production of biofuels from biomass and on other potential energy sources such as biophotolytic decomposition of water to hydrogen. 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://doegenomestolife.org/biofuels/).



To address the analytical and computational capabilities needed to put the GTL research program on track for creating a science foundation for DOE missions, workshops were held between June 2002 and June 2004. Much of the material in this chapter was drawn from the outputs of those workshops, in which nearly 800 different individuals participated. For a list of GTL workshops, meetings, and links to workshop reports, see Appendix D. GTL Meetings, Workshops, and Participating Institutions, p. 239.

Facilities Overview

The proposed GTL user facilities for 21st Century biology and biotechnology will be a major strategic asset in achieving DOE mission goals in industrial biotechnology—a critical arena of national economic competitiveness. The facilities will enable a new era in biology, building on the national investment in genomics.

The research community increasingly is recognizing the need for global analysis of myriad simultaneous cellular activities and is calling for a new research infrastructure. "Progress in microbiology always has been enabled by the technology available, a fact that is still true today. However, many researchers are stymied by lack of access to the expensive instruments that would enable them to make the greatest strides." (Schaechter, Kolter, and Buckley 2004, p. 13; see also Aebersold and Watts 2002; Buckley 2004a; Stahl and Tiedje 2002).

5.0.1. Science and Technology Rationale

In the simplest interpretation, an organism's genome contains encoded information to produce the proteins that are the cell's workhorses. Central to the strategy of GTL and the GTL facilities suite, however, is the fact that information encoded in the genomes of microbes and the metagenomes of microbial communities goes well beyond encoding proteins. Genomes control microbial structure and function at many spatial and temporal scales—molecular, machine, cellular, community, and ecosystem—through an intricate set of interrelated and communicating regulatory and control processes (see Fig. 1. DOE Genomics:GTL High-Throughput User Facilities, p. 103). Microbes display such strong interactions that capabilities are needed to explore these systems in a comprehensive and integrated way at all levels. Proteins and even microbes are thought to function rarely in isolation.

In the Missions Overview, our example mission descriptions demonstrate that the technical challenges of these analyses and the scale of the systems that must be understood exceed any existing capabilities (see Missions Overview, Tables 1–9, beginning on p. 26; 3.2.2. Science and Technology Milestones, p. 44; and sidebar, High-Throughput Model Guides Future Facilities, p. 6). Facilities must be established to dramatically improve research performance, throughput, quality, and cost.

Examples of performance challenges include producing and characterizing complex proteins (e.g., membrane and multidomain); isolating,

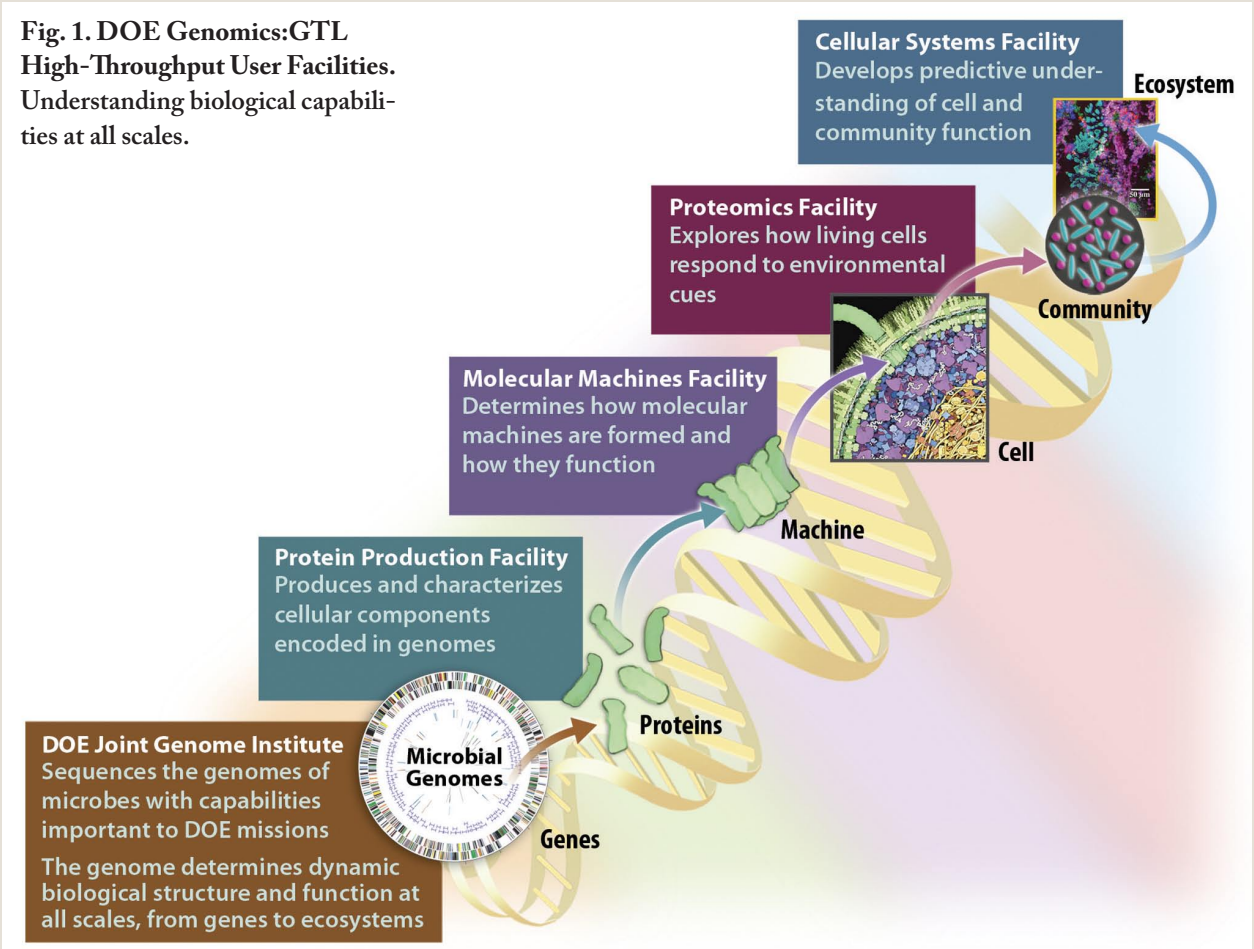


Fig. 1. DOE Genomics:GTL High-Throughput User Facilities. Understanding biological capabilities at all scales.

characterizing, and modeling large or tenuous molecular machines; measuring the full molecular profile of microbial systems; and imaging molecules as they carry out their critical functions in cells in structured communities. Examples of throughput challenges include providing insight into the functions of hundreds of thousands of unknown genes and their modifications; processing thousands of molecular machines; analyzing molecular profiles of thousands of microbial samples under different conditions; and spanning the full range of conditions and processes governing microbial-community behaviors. Quality control includes developing and implementing strict protocols and providing the most sophisticated diagnostics. High-throughput methods and resource sharing among community members will lower the unit cost for production and analyses.

Figure 1, this page, depicts facilities focused on building an integrated body of knowledge about behavior, from genomic interactions through ecosystem changes. Simultaneously studying multiple microbial systems related to various mission problems is powerfully synergistic because enduring biological themes are shared and general principles governing response, structure, and function apply throughout. The biology underlying the challenges of one mission will inform those of the others. Accumulating the data as it is produced, the GTL Knowledgebase and the computational environment that GTL will create will act as the central nervous system of the facilities and program, allowing this information to be integrated into a predictive understanding.

The Office of Science has a tradition of strategic basic research in a multidisciplinary team environment for national missions. These facilities will bring together the biological, physical, computational, and engineering sciences to create a new infrastructure for biology and the industrial biotechnology needed for the 21st Century. DOE's technology programs can work with industry to apply such capabilities and knowledge to a new generation of processes, products, and industries.

5.0.2. A New Trajectory for Biology

As we have learned from the genome projects, consolidating capabilities and focusing on aggressive goals will drive dramatic improvements in performance and cost (Fig. 2. Putting Biology on a New Trajectory, this page). As depicted in Fig. 2, GTL facilities will accelerate discovery and reduce the time for useful applications. With this higher level of performance, microbial systems biology is tractable and affordable to support the next generation of industrial biotechnology for the coming decade and beyond.

5.0.3. Capsule Facility Descriptions

The GTL facilities provide a complementary set of technologies and products. Two facilities are focused on analysis of properties and functions of cellular components, proteins, and molecular machines:

- Facility for Production and Characterization of Proteins and Molecular Tags
- Facility for Characterization and Imaging of Molecular Machines

Two are focused on analysis of microbial-system responses and functions at the molecular, cellular, and community levels:

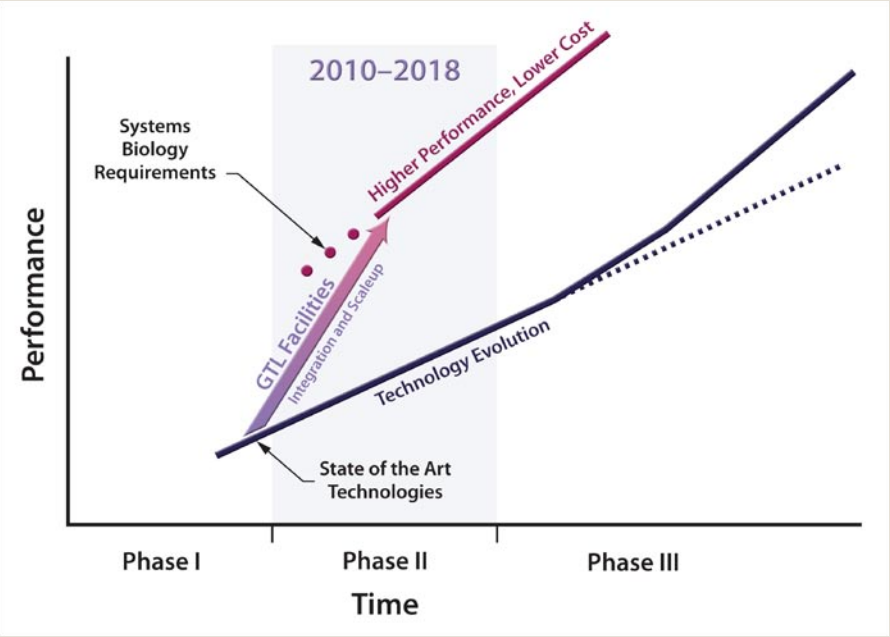
- Facility for Whole Proteome Analysis
- Facility for Modeling and Analysis of Cells and Communities

The ensuing chapters of this roadmap discuss each facility in further detail. Capsule facility descriptions follow.

5.0.3.1. Facility for Production and Characterization of Proteins and Molecular Tags

The Protein Production and Characterization Facility will use DNA sequence to make proteins and reagents for interrogating cell function. Specifically, this facility will have the capability to produce all proteins encoded in any genome on demand; create molecular tags that allow each protein to be identified, located, and manipulated in living cells; and, to gain insights into function, perform biophysical and biochemical characterizations of proteins produced. Using high-throughput in vitro and in vivo techniques will lower the

Fig. 2. Putting Biology on a New Trajectory.



cost of producing proteins to levels that will allow comprehensive analysis of all proteins within the cell. The facility's products and analysis capabilities will be made available to all scientists.

In parallel with protein production will be the generation of "affinity reagents." These small proteins or nucleic acids will permit the detection and tracking of individual proteins in living systems, including complex molecular assemblages; the intracellular position of all proteins and their spatial dynamics; if exported, the extracellular localization and interaction with other community members; and techniques for manipulating protein activity in the environment.

Core facility instrumentation:

- Gene synthesis and manipulation techniques
- High-throughput microtechnologies for protein-production screening
- Robotic systems for protein and affinity-reagent production and characterization
- Computing for data capture and management, genomic comparative analyses, control of high-throughput system and robotics, and production-strategy determination

5.0.3.2. Facility for Characterization and Imaging of Molecular Machines

The Molecular Machines Facility will identify and characterize molecular assemblies and interaction networks. It will have capabilities to isolate and analyze molecular machines from microbial cells; image and localize molecular machines in cells; and generate dynamic models and simulations of the structure, function, assembly, and disassembly of these complexes. The facility will identify molecular machine components, characterize their interactions, validate their occurrence and determine their locations within the cell, and allow researchers to analyze the thousands of molecular machines that perform essential functions inside a cell. It will provide a key step in determining how the network of cellular molecular processes works on a whole-systems basis by completely understanding individual molecular machines, how each machine is assembled in 3D, and how it is positioned in the cell with respect to other components of cellular architecture.

Core facility instrumentation:

- Robotic culturing technologies to induce target molecular machines in microbial systems and supporting robotic techniques for molecular complex isolation
- Numerous sophisticated mass-spectroscopy and other techniques specially configured to analyze samples of purified molecular machines for identification and characterization of complexes
- Various advanced microscopies for intracomplex imaging and structure determination
- Imaging techniques for intracellular and intercellular localization of molecular complexes
- Computing and information systems for modeling and simulation of molecular interactions that lead to complex structure and function

5.0.3.3. Facility for Whole Proteome Analysis

The Proteomics Facility will be capable of gaining insight into microbial functions by examining samples to identify (1) all proteins and other molecules that a microbe (or microbial community) creates under controlled conditions and (2) key pathways and other processes. An organism selectively produces portions of its proteome in response to specific environmental or intracellular cues. Studying its constantly changing protein expression thus leads to a better understanding of how and why an organism turns portions of its genome "on" and "off." Facility users will achieve a comprehensive understanding of microbial responses to environmental cues by identifying, quantifying, and measuring changes in the global collections of proteins, RNA, metabolites, and other biologically significant molecules. These molecules, including lipids, carbohydrates, and enzyme cofactors, are important in understanding biological processes mediated by proteins. Integrating diverse global

data sets, the facility will develop computational models to predict microbial functions and responses, inferring the nature and makeup of metabolic and regulatory processes and structures.

Core facility instrumentation:

- Large farms of chemostats to prepare samples from highly monitored and controlled microbial systems under a wide variety of conditions
- Numerous specialized mass and NMR spectrometers and other instrumentation capable of analyzing the molecular makeup of ensemble samples with thousands of diverse molecular species
- High-performance computing and information capabilities for modeling and simulation experiments of microbial-system functionalities under different scenarios to inform the design of experimental campaigns focused on systems-level goals and to infer microbial-system molecular processes from ensuing data

5.0.3.4. Facility for Modeling and Analysis of Cellular Systems

The Cellular Systems Facility will be the capstone for the ultimate analytical capabilities and knowledge synthesis to enable a predictive understanding of cell and community function critical for systems biology. The facility will concentrate on the systems-level study of living cells in complex and dynamic structured communities. Imaging methods will monitor proteins, machines, and other molecules spatially and temporally as they perform their critical functions in living cells and communities. Microbial communities contain numerous microniches within their structures that elicit unique phenotypic and physiological responses from individual species of microbes. We need to be able to analyze these niches and the microbial inhabitants within. This grand challenge for biology must be addressed before scientists can predict the behavior of microbes and take advantage of their functional capabilities. Modeling in the facility will describe essential features of these biological interactions with the physicochemical environment and predict how the system will evolve in structure and function.

Core facility instrumentation:

- Highly instrumented cultivation technologies to prepare structured microbial communities to simulate natural conditions under highly controlled conditions
- Instruments integrating numerous analytical imaging techniques that can spatially and temporally determine, in a nondestructive way, the relevant molecular makeup and dynamics of the community environment, community, and microbes that comprise it
- Computing and information capabilities to model and simulate complex microbial systems, design experiments, and incorporate data

5.0.4. Relationships and Interdependencies of Facilities

Each of the facilities is technically distinct in the nature of its instrumentation, methods, and overall goals. All will be centered around either production lines designed to maximize quality and throughput and reduce unit costs, the development and operation of frontier instrumentation or unique suites of instrumentation to reach new levels of performance, or combinations of both. While each can serve a user community for a wide range of independent studies, the suite of facilities has complementary strengths and core technologies that together can help provide complete systems knowledge. Figure 3. GTL Facilities: Core Functions and Key Interactions, p. 108, displays how each facility's core functions are complementary to those of the other facilities. The key interactions shown demonstrate their interdependencies and necessary exchange of all information through the GTL Knowledgebase and the program's communication and computing infrastructure.

5.0.5. Research Scenarios

As described in the Missions Overview and related appendices, each mission example has a unique endpoint and research strategy for developing the needed understanding, predictive models, and research capabilities. Table 1. Research Scenarios on Microbial Processes, p. 109, and Table 2. Science Roadmaps for Natural Systems, p. 110, present conceptual research-scenario roadmaps for six cases as illustrations related to Science Milestones and GTL Facilities. Although these systems and problems cover a breadth of microbial phenomenology and system behaviors, they can be studied using the same foundational capabilities. Each of the GTL milestones, as denoted in the left column of Tables 1 and 2, drives the technical core of the facilities, where capabilities resulting from milestone R&D can be scaled up and integrated.

5.0.6. Facility Development

The facility acquisition process will employ project-management practices similar to DOE Order 413.3 Facilities Project Management. The facilities budget will include all costs for the conceptualization, design, R&D and testing, and acquisition of the necessary conventional facilities, instrumentation, computers and software, and supporting technologies, training, and installation of fully operational production lines and analytical facilities upon completion of the project. The process will involve participants from national laboratories, academia, and industry in the necessary workshops and working groups to determine the technical scope and scale of the facilities, technical priorities, and technology development. Many of the long-lead and crosscutting development needs are outlined in the GTL Development Summary chapter. This roadmap is meant to be a starting point for the intensive conceptualization and planning that must occur for successful design, acquisition, and operation of these facilities.

Fig. 3. GTL Facilities: Core Functions and Key Interactions.

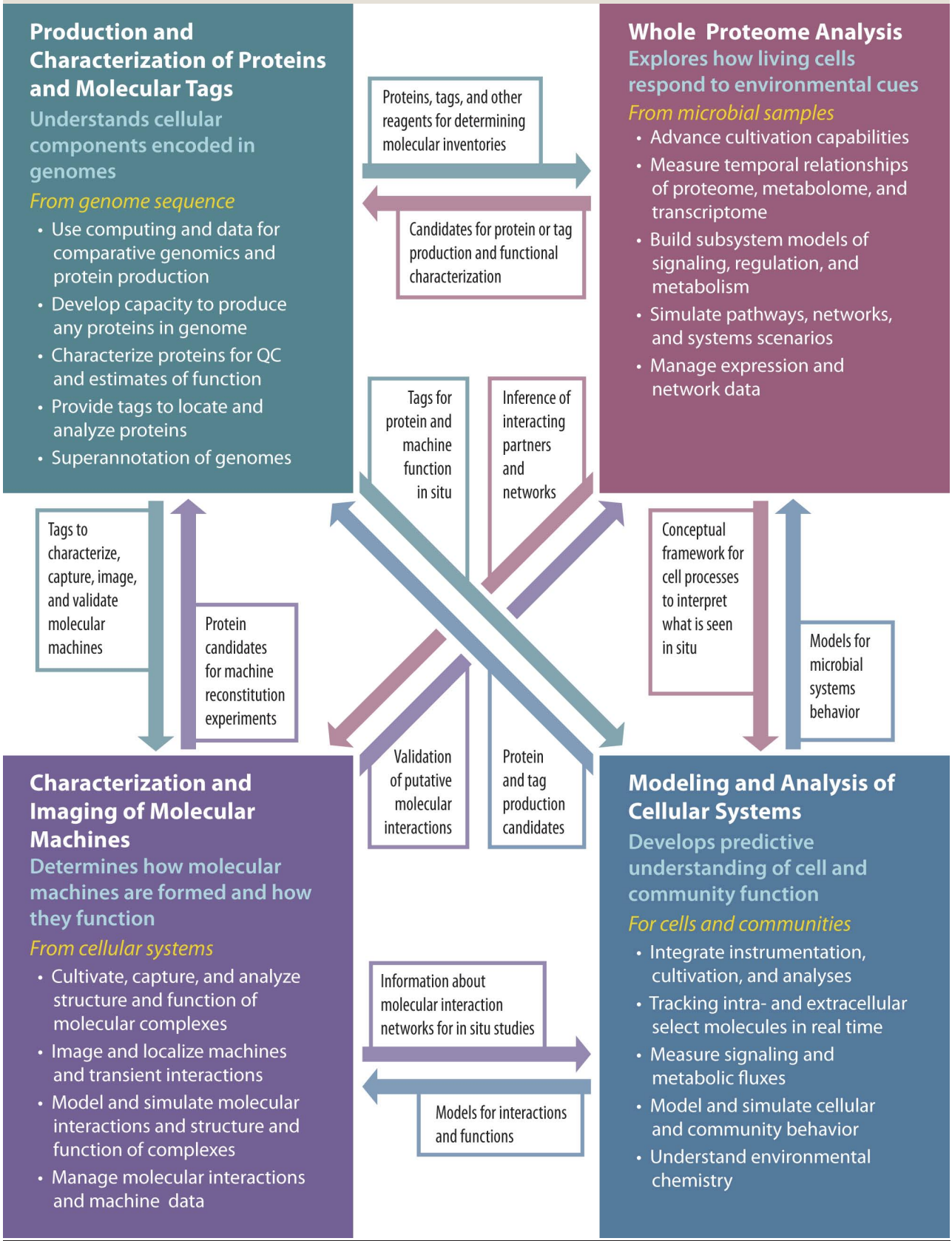


Table 1. Research Scenarios on Microbial Processes: Relationship to Science Milestones and Facilities

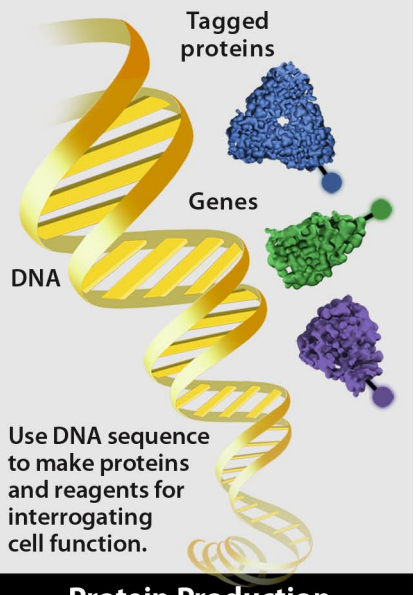
Progression of GTL Science Milestones and Facilities	*Conceptual Science Roadmaps for Microbial Energy and Environmental Processes*		
	Convert Sunlight to Hydrogen and High-Hydrogen Fuels	**Convert Cellulose to Fuels**	**Reduce Toxic Metals in Subsurface Environments**
Milestone 1: Determine the Genome Structure and Potential of Microbes and Microbial Communities Facility for the Production and Characterization of Proteins and Molecular Tags Facility for Characterization and Imaging of Molecular Machines	Analysis of hydrogenase families across microbial species: Screen nature for new variants Range of hydrogenase properties Suite of heterologous expression hosts Characterization of partners, energetics, structures, post-translational modifications Wide range of mutations, variations created and screened Functional and structural analysis of machines	Wide range of microbes surveyed for cellulases, ligninases, and other glucosyl hydrolases Partners and structural information established Structure and imaging of interactions important to efficient function	Survey of subsurface species and genomic potential Comparative genomics and superannotation Generation of knockouts, mutations, transmembrane structures to understand function
Milestone 2: Develop a Systems-Level Understanding of Microbial and Community Function and Regulation Facility for Whole Proteome Analysis Facility for Analysis and Modeling of Cellular Systems	Oxygen sensitivity of hydrogenases Electron-transfer reactions and limitations Reverse-reaction mitigation Partitioning of electrons between hydrogenases and competing pathways Light capture Biophotoelectric antenna	Proteome analysis of expression and regulation Fundamental mechanisms of cellulose deconstruction Transport of sugars Measurement of electron transport chains' redox state, control of electron fluxes Carbon partitioning in cells: Carbon, NAD, NADPH, ATP, ADP	Cellular response to environmental stimuli Proteomics, transcriptomics, and metabolomics to elucidate regulation and responses Intra- and intercellular communications Cells in structures such as biofilms Growth processes, toxicity responses, energy transfer, metabolic responses Microbe-mineral interactions
Milestone 3: Develop the Knowledgebase, Computational Methods, and Capabilities to Advance Understanding of Complex Biological Systems and Predict Their Behavior GTL Integrated Computational Environment for Biology	Pathway models—energetics, electropotential, docking, proton fluxes, cofactors Computational tools for rational design Suites of hosts, pathway cassettes Modeling and measurement of pathways, fluxes, regulation	Design of organisms capable of utilizing all sugars Optimization of sugar transport, regulation Redesign of cellulose structure pH- and temperature-tolerant microbes Principles for enzyme redesign	Modeling capable of visualizing realistic biochemical pathways in cells Interactions of membrane proteins with contaminants and solid-phase electron acceptors Design of experiments in cultured and natural systems
Missions Outputs Systems Engineering	In vivo systems Processes captured in nanostructures, biomimetic systems System design: Light harvesting, conversion to hydrogen or fuel, robustness to oxygen, regulation Transgenic approaches	Improved cellulases and production methods to reduce costs, improve stability Modularized processing to reduce transportation of feedstock Sensors for biomarkers and chemical intermediates	Assessment of long-term cellular and system behavior Remediation strategies Sensors for coupled biochemical and geochemical measurements in situ

Table 2. Science Roadmaps for Natural Systems: Relationship to Science Milestones and Facilities

Progression of GTL Science Milestones and Facilities	*Conceptual Science Roadmaps for Natural Systems*		
	Oceans: Photosynthetically Driven Biological Pumps for Carbon and Energy in Aquatic Systems	**Terrestrial: Microbes in Ecological Communities, Carbon and Nutrient Cycles**	**Deep Subsurface: Microbial Community Processes for Mitigation of Toxic Chemicals and Metals**
Milestone 1: Determine the Genome Structure and Potential of Microbes and Microbial Communities Facility for the Production and Characterization of Proteins and Molecular Tags Facility for Characterization and Imaging of Molecular Machines	Single-cell and environmental community sequence Heterotrophs, autotrophs, viruses, and "twilight zone" organisms Comparative analyses of rhodopsin, hydrogenase genomes Gene synthesis and manipulation	Single-cell and community sequence in situ and in vitro Organisms related to processes in soils Genome annotation	Single-cell and community sequence in situ and in vitro to identify members, functions Superannotation, genome plasticity effects Metagenomics, gene transfer Tags to ID microbes, proteins, metabolites
Milestone 2: Develop a Systems-Level Understanding of Microbial and Community Function and Regulation Facility for Whole Proteome Analysis Facility for Analysis and Modeling of Cellular Systems	Photosynthesis, transporters, biomineralization Proteins, machines, metabolites, and functional assays Systems responses Imaging	All GHGs: CO_2, methane, nitrous oxide, dimethyl sulfide Molecular inventories vs cues Systems interactions with soil, rhizosphere, plants: Inputs and outputs (e.g., stable isotope probes) Proteome and metabolome imaging at cellular and community levels	Community structure and relationship to function Pathways and networks: Mechanisms of intercellular communication and function Stoichiometry and kinetics of intercellular fluxes
Milestone 3: Develop the Knowledgebase, Computational Methods, and Capabilities to Advance Understanding of Complex Biological Systems and Predict Their Behavior GTL integrated computational environment for biology	Modeling of climate-based and mitigational perturbations Individual and multiple life-scale models (cellular, community, ecosystem): Metabolic budgets Multiple photosynthetic processes	Modeling of microbial responses to manipulation of plant inputs into carbon cycle Human inputs directed to soils Response to environmental change understood	Four-dimensional reactive transport models based on genomic, geochemical, and hydrological data Scaling of processes through molecular, cellular, community, and environmental levels; and molecular to long time scales
Missions Outputs Measure environmental responses via sensors	Ecogenomics of sentinel organisms Cellular, community, and ecosystem biochemical assays Accompanying environmental assays	Biomarkers: RNAs, proteins, metabolites, signaling Ecogenomics, functional assays, environmental conditions Carbon and nutrient inventories	Biology and geochemistry: DNA, RNA, proteins, metabolites, geochemical from single-cell to field scales Mesoscale simulation of field conditions Regulatory levels of contaminants
Robust Science Base for Policy and Engineering	Natural behaviors of ocean ecosystems, impact on and of climate change scenarios incorporated into IA models Assessment of efficacy and impacts of intervention strategies	Biological processes for carbon and nitrogen cycling, impact on and of climate change scenarios incorporated into IA models Assessment of potential and strategy for terrestrial carbon sequestration	Predictions of transport and fate Assessment of need for remediation Remediation strategies, designs, and tests

5.1. Facility for Production and Characterization of Proteins and Molecular Tags

5.1.1. Scientific and Technological Rationale 112
5.1.1.1. Value of Proteins for Research 114
5.1.1.2. Value of Protein Characterization for Research 115
5.1.1.3. Value of Molecular Tags for Research 115
5.1.2. Facility Description 116
5.1.2.1. Facility Outputs 117
5.1.2.2. Laboratories, Instrumentation, and Support 117
5.1.3. Development of Methods for Protein Production 118
5.1.3.1. Production Targets 118
5.1.3.2. Specifications for Proteins and Comparisons of Their Production Methods 119
5.1.3.2.1. Comparison of Cell-Based Expression Systems 120
5.1.3.2.2. Cell-Free Systems 122
5.1.3.2.3. Chemical Synthesis 122
5.1.3.2.4. Protein Purification 122
5.1.4. Development of Methods for Protein Characterization 123
5.1.4.1. Requirements, Specifications for Functional Characterization Techniques, Data 125
5.1.5. Development of Approaches for Affinity-Reagent Production 126
5.1.5.1. Specifications for Affinity Reagents and Their Production 127
5.1.5.2. Technologies for Affinity-Reagent Production 130
5.1.6. Development of Data Management and Computation Capabilities 131
5.1.7. Facility Workflow Process 131



Protein Production and Characterization

- Produce proteins encoded in the genome.
- Create affinity reagents that allow each protein to be identified, located, and manipulated in living cells.
- Perform biophysical and biochemical characterizations of proteins produced to gain insights into function.

Facility for Production and Characterization of Proteins and Molecular Tags

The Facility for Production and Characterization of Proteins and Molecular Tags will be a user facility providing scientists with an understanding of the components encoded in the genome by using DNA sequence to make and characterize proteins and reagents for interrogating their functions in cells.

5.1.1. Scientific and Technological Rationale

Systems biology requires that we understand the proteins that make up a cell and the mechanisms of their function. Individual proteins encoded in the genome are the basic building blocks for biological functions potentially useful in DOE missions. Virtually every cellular chemical reaction and physical function necessary for sustaining life is controlled and mediated by proteins generally organized into macromolecular complexes or "molecular machines," which might contain proteins, RNAs, or other biomolecules. A typical microbial genome has 2000 to 5000 genes that encode thousands of proteins and regulatory regions that control their expression. The challenge of understanding these workhorse molecules is technically complex and necessitates that very large numbers of them be produced and analyzed. Experimental analysis has determined the functions of only a few thousand of the millions of proteins encoded by the collective genomes on this planet—and even that understanding is incomplete.

Example of Mission Problem

Proteins Provide Insight into Energy Production

Understanding the functions of bacteria, fungi, and algae is important for determining new ways to produce hydrogen or ethanol economically as a fuel. The genome sequences of these organisms provide a first step, but proteins carry out the useful functions encoded by the genes. To study proteins, they must be produced in quantities sufficient for analysis. In addition, studying these molecules functioning in their natural state (i.e., in the cell) requires the generation of affinity reagents or other molecular tags able to recognize specific proteins. Understanding how hydrogen-generating proteins function inside and outside cells will guide optimization of enzymatic hydrogen production for cell and cell-free applications.

Protein Production and Characterization

We currently have insufficient data and conceptual insights to assign at least one function to about half the proteins found in even the most intensively studied microorganisms. Functional assignments for proteins in unculturable or less-studied organisms often occur by inference from a homologous protein's putative role in an intensively studied organism. A comprehensive understanding of cellular behavior will require experimental data for a significant portion of an organism's proteins (Roberts et al. 2004). We must have the ability to produce and characterize, as needed, essentially all the thousands of proteins encoded in many single genomes and in metagenomes to support functional gene annotation and, ultimately, mechanistic understanding. We also need to be able to produce and screen numerous variants of individual proteins or molecular machines so they can be used for DOE applications.

Facility Objectives

- **Perform comparative genomics against GTL Knowledgebase to determine gene function and to inform needed protein production and characterization**
- **Produce any protein on demand**
- **Characterize all proteins for quality assurance and quality control, for function, and for determining structure-function relationships as needed**
- **Produce affinity reagents and other molecular tags to enable location, tracking, and manipulation of proteins and machines in living systems**
- **Provide clones, proteins, affinity reagents, protocols, and data to scientists**

Having full-length and active forms of proteins in hand for biochemical and biophysical analysis will serve many purposes critical to the next generation of biology. These proteins provide an opportunity for discovery and a starting point for optimizing complex cellular processes from their components and molecular mechanisms. Providing rigorous and comprehensive characterizations for these proteins is invaluable to researchers and frees them to confidently pursue creative experimentation. "Molecular tags" or "affinity reagents" can be produced only by working from the proteins or via protein modification. These tags are critical for detection and potential quantitation of individual proteins and molecular machines in living systems.

The study of microbes, and especially those of DOE relevance, presents a special challenge. Microbial-community systems that we must understand possess millions of genes as opposed to the tens of thousands of even the most-complex higher organisms. The readily available genome sequences and even metagenome sequences of microbial communities have provided our first look into microbes' many functions. Most of the recently sequenced microbial genomes and metagenomes, however, show that roughly 40% of the genes are of unknown function, and, further, the microbes themselves either are not available or are "unculturable." Roughly 200 microbes have been sequenced to date, resulting in a catalogue of unknown genes that now contains 200,000 to 400,000 candidates for investigation. The ability to create and gain insight into proteins from genomic information alone is a crucial first step to understanding these microbial systems. Eventual culture-dependent experimentation on an important subset of microbes will be facilitated greatly by the availability of basic information on proteins and their respective affinity reagents.

Protein production currently is limited by economic and technological constraints and is a widely dispersed and inefficient "cottage industry." While substantial technology exists for generating the easy-to-produce (i.e., small, soluble) proteins, the ability to readily produce large multidomain proteins, membrane proteins, proteins with cofactors, and many other critical proteins is only emerging. For comprehensively understanding microbial systems, access to all proteins in metabolic, signaling, and regulatory pathways and networks is important. The most difficult proteins often are the very ones most vital to cellular function (e.g., those associated with essential transmembrane molecular machines, such as the photosystems in a photosynthetic microbe). In its mature state, the Protein Production and Characterization Facility will spend the greatest part of its effort on hard-to-produce, but critically important, proteins and will enlist the research community to help develop needed methods.

A unique benefit of this facility is that, for the first time, a substantial suite of high-throughput, automated, and increasingly sophisticated characterization assays will be performed on proteins. Thus, protein production and characterization both will benefit as the transition is made from widely dispersed efforts focused on easy proteins to the economy of scale made possible by developing technologies capable of producing any desired protein with an accompanying database of reliable characterizations. The situation is somewhat analogous to genomic sequencing as it transitioned from dispersed, somewhat unreliable sequence data to higher-quality, lower-cost data at high-throughput, automated sequencing centers.

Automated high-throughput protein and affinity-reagent production will have several important impacts, including the following, that will enable the expeditious systemic study of chemical and physical interactions of proteins that underlie biology:

- A production environment will establish the necessary standards, diagnostics, control, and quality to develop and execute the demanding protocols for readily and repeatedly producing difficult proteins.
- A production facility will support a comprehensive and sophisticated array of characterization methods, most unavailable to the individual researcher, that can be applied to both production diagnostics and to protein characterization.
- Large-scale robotics, miniaturization, and automation will greatly enhance throughput and reduce costs.
- Making material and data products available to all scientists will leverage the investment to reach a larger community, whose work will facilitate further production, characterization, and understanding.
- Unlike the current situation, in which only selected portions of labor-intensive data are accessible, the facility's strong computational infrastructure will facilitate data mining of both successful and unsuccessful metadata associated with each protein.

5.1.1.1. Value of Proteins for Research

Ready and economic availability of proteins and affinity reagents will provide the foundation for the next generation of biological research, building on the national investment in genome sequencing. Having widespread access to cutting-edge technology in protein production will level the playing field, increasing the availability of proteins and protocols and creating a broader biotech industry (see sidebar, Protein Microarrays have Multiple Uses, p. 115). Proteins form the starting point for biochemical and biophysical functional studies, for eventual protein engineering, and for creating chimeric or new (optimized) biochemical pathways or even reactions or pathways that work in reverse directions (e.g., carbon dioxide to formate to methane). They offer the ability to study low-abundance proteins such as important regulatory proteins. Many variants (mutations) can be produced and studied for functional analysis. For nonculturable organisms, proteins can be produced from sequence alone to provide a shortcut to functional genome annotation and allow determination of quantitative biochemical binding or reaction constants. Comparative analyses of the structure and functions of protein families can be used to determine design principles. Proteins are reagents for studying metabolomics, post-translational modifications (substrate identifications), biosynthesis of metabolites and intermediates, binding-partner identification, and affinity-reagent generation. Functional proteins are the starting material for reconstituting molecular complexes, making quantitative and qualitative three-dimensional spectral and structural analyses, and mapping molecular interactions (with DNA, metabolites, and other proteins). They also can serve as mass and spectral standards for enhancement of mass spectrometry (MS) data analysis. Proteins, affinity reagents and other molecular tags, and data produced in the Protein Production and Characterization Facility are needed by users of other facilities to capture molecular machines for MS and other analyses and to identify the machines' components. They also are needed for cellular-imaging studies and verification of models (Roberts 2004; Roberts et al. 2004; see Table 1. Analysis of Technology Options for Protein Production, p. 120, and Table 2. Roadmap for Development of Technologies to Produce Proteins, p. 121).

5.1.1.2. Value of Protein Characterization for Research

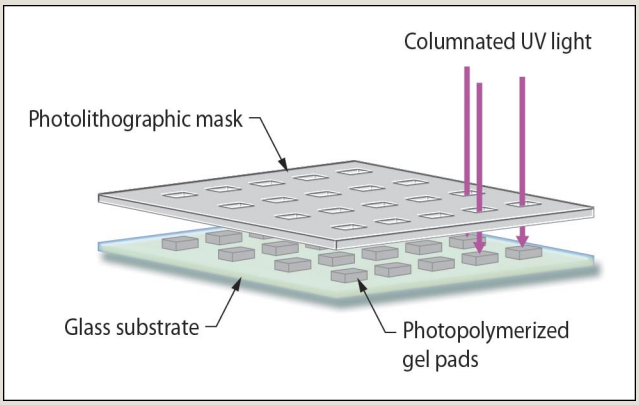
Automated high-throughput and high-quality biophysical and biochemical characterizations of proteins will provide a more rigorous assignment of gene function, resulting in first insights to a mechanistic understanding of microbial capabilities. For the first time, comprehensive reliable data on thousands of proteins will be available to analysts. The production facility can use high-throughput screening to characterize many proteins simultaneously under widely varied controlled conditions. Early analyses will focus on characterizations to determine basic biochemical function and biophysical information (e.g., solubility and insolubility in multiple solutions, multimeric state, presence of metals, and ordered and disordered domains). As the facility matures, the nature and sophistication of these characterizations will expand to determine more complex functions of individual proteins and molecular complexes (see 5.1.4. Development of Methods for Protein Characterization, p. 123, and Table 3. Summary of Characterization Needs and Methods, p. 124).

5.1.1.3. Value of Molecular Tags for Research

Two types of molecular tags are discussed here: Affinity reagents and fusion tags. Affinity reagents comprise proteins, peptides, nucleic acids, and small chemical molecules that bind targets of interest with high specificity and affinity. They commonly are used to detect where particular proteins are localized in cells, recover the protein and its associated molecules from cell lysates, and quantitate protein amounts in complex mixtures. Antibodies, popularly used as affinity reagents, can be generated by immunizing rodents or rabbits with the protein target and harvesting immunoglobulins (e.g., IgM and IgG) from the serum several months later. With the advent of various in vitro methods termed "display technologies," antibody fragments (i.e., scFv, Fab, VH, VL,

Protein Microarrays Have Multiple Uses

Proteins mass produced in the Protein Production and Characterization Facility or by the commercial sector from facility protocols may be delivered as microarrays to investigators in their labs. These devices provide a platform for directly studying global protein interactions and networks. Protein arrays also can serve as global "pulldown" and "affinity" purification platforms for spatially isolating molecular machines and complexes in the Molecular Machines Facility. Protein chips might serve as prepurification steps or as assays in the Proteomics Facility.



An example of a protein microarray is the "biochip" pictured above. This array supports 3D gel pads covalently attached to proteins, nucleic acids, antibodies, aptamers, and functional enzymes. The gel pads allow a solution-phase test environment that avoids subjecting biomolecules to the potentially harsh effects of a surface (e.g., glass slides). These arrays have been used for DNA and RNA analyses (including actual environmental samples), on-chip polymerase chain reaction and various ligation or amplification reactions, antibody arrays, functional protein assays, and protein-protein interaction studies. Combining these arrays (containing, for example, proteins, enzymes, aptamers, or antibodies) with time-of-flight mass spectrometry and automated spectral analysis allows characterization of the mass and identification of agents interacting with the elements embedded on the biochip. [Source: Argonne National Laboratory]

References

1. A. Pemov et al., "DNA Analysis with Multiplex Microarray-Enhanced PCR," *Nucl. Acids Res. Online* 33(2): e11 (2005). Retrieved from http://nar.oxfordjournals.org/cgi/content/full/33/2/e11.

2. I. M. Gavin et al., "Analysis of Protein Interaction and Function with a 3-Dimensional MALDI-MS Protein Array," *BioTechniques*, 39(1), 99–107 (2005).

and VHH) can be isolated from naïve libraries in several weeks' time without the use of animals. In addition to modifying antibody-based molecules, scientists are altering other proteins (e.g., lipocalin, ankyrin, fibronectin domain, and thioredoxin) to bind to specific targets of interest. This is accomplished by modifying the open reading fragments through mutagenesis and selecting among the resulting library of randomized proteins for those that bind targets specifically. Finally, affinity reagents can be selected from libraries of combinatorial peptides, nucleic acids (i.e., aptamers), or small organic molecules (see 5.1.5. Development of Approaches for Affinity-Reagent Production, p. 126; Table 4. Analysis of Technology Options for Affinity Reagent Production, p. 127; Table 5. Roadmap for Development of Technologies to Produce Affinity Reagents, p. 128; and Table 6. Examples of Affinity Reagents and Their Applications, p. 128).

The following items focus on affinity reagents:

- Production of affinity reagents must be designed around their many applications. When proteins are in structured environments, some surfaces are exposed while others are hidden because they are in contact with other proteins or molecules. To deal with this contingency, multiple affinity reagents for each protein will ensure that any exposed surface or epitope can be accessed.
- Affinity reagents are needed that either disrupt or preserve protein activity. They can be used to manipulate proteins, including fabrication of biosensors; map post-translational modifications; determine spatial distributions; array targets in a unique spatial configuration; disrupt protein-protein interactions; promote crystallization of proteins; and stabilize membrane proteins.
- Affinity reagents can be used to assess biodiversity and in diagnostic tools for energy-production processes. They are critical for affinity purification of proteins and complexes, for identifying binding surfaces and mapping interactions in protein complexes, and for characterizing functional states (by targeting epitopes unique to active or inactive forms of the proteins). Finally, they are valuable in flow cytometry to sort cells from mixtures and for use in nanotechnology to anchor proteins during fabrication of novel biohybrid materials.

Another type of molecular tags—fusion tags—are short peptides, protein domains, or entire proteins that can be fused at the genetic level to proteins of interest. The target protein then is imparted with the fusion tag's biochemical properties. In general, the type of fusion tag used is dictated by its application. Short peptide tags (e.g., six-histidine, epitopes, StrepTag, calmodulin-binding peptide) regularly serve to permit facile purification of the recombinant protein, allow detection of the fusion protein, or direct the recombinant protein's interaction with other proteins or inert surfaces. Larger fusion partners such as protein domains (e.g., chitin-binding domain) or proteins (e.g., cutinase, GFP, GST, MBP, and intein) usually are employed to promote folding, solubility, purification, labeling, chemical ligation, or immobilization of the recombinant protein. If desired, the fusion tag can be detached from the protein of interest by cleaving a linker region with a site-specific protease that does not affect the protein (see Table 7. Examples of Fusion Tags and Their Applications, p. 129).

5.1.2. Facility Description

The facility will bring together comprehensive technologies for high-quality mass production and characterization of microbial proteins produced directly from sequence data or other genetic sources such as gene variants or clones. It also will be capable of generating specific capture and labeling affinity reagents for each protein. To derive insights into gene function and assess the best and most cost-effective protein-production strategies, a key capability will be computational comparison of genomic sequences of unknown organisms against the comprehensive GTL Knowledgebase. This user facility will integrate the basic research and technology development necessary to enable its continued scientific focus and usefulness in working with investigators and technologists in academia, national laboratories, and industry (see 5.1.7. Facility Workflow Process, p. 131, and accompanying sidebar with conceptual diagrams and narrative, p. 133).

5.1.2.1. Facility Outputs

Facility products will be distributed to research teams and accessible to the broader community of biologists. In general, proteins will have limited distribution because the facility will establish successful protocols and expression constructs that will allow researchers or commercial concerns to then produce proteins as needed for wider applications. Data and computational analyses will be available freely through the GTL computational environment. Products provided as needed to the user community include:

- Expression vectors (clones) for targeted genes
- Milligram quantities of purified, full-length, functional proteins
- Multiple affinity reagents for each protein, as well as chips with arrayed affinity reagents
- Proteins with a variety of fusion tags
- Initial biophysical and biochemical characterizations of each protein
- Production protocols so researchers and commercial concerns can readily produce proteins for research and biotechnology applications
- Comprehensive production and characterization databases and computational analyses referenced to the subject genome or classes of proteins

5.1.2.2. Laboratories, Instrumentation, and Support

The high-throughput facility's 125,000- to 175,000-sq.-ft. building will house core resources for protein production and characterization and the support necessary to ensure its mission. It will have extensive robotics for efficient sample production and processing and suites of highly integrated instruments for sample analysis and characterization of proteins and affinity reagents.

In the facility will be laboratories and instrumentation for production of large numbers of different DNA molecules, including cloning and insertion into expression vectors and, eventually, gene synthesis capabilities; production of proteins from any biological source; purification; quality assessment; and production of protein variants [e.g., isotopically labeled proteins, post-translationally modified proteins, proteins with novel cofactors, proteins incorporating nonstandard amino acids, and site-specific mutant arrays (high-throughput mutagenesis)]. The facility also will involve production of multiple affinity reagents for each protein; production of membrane proteins and multiprotein complexes; multimodal protein biophysical and biochemical characterization; and combinatorial capabilities to screen for complexes under multiple defined conditions. Methods will comprise cellular or cell-free expression and chemical synthesis. Onsite DNA sequencing will be required for several steps in the process. Informatics capabilities will track each gene or clone, protein, affinity reagent, and the associated data. Quality control will be assessed by onsite MS and a range of other biophysical and biochemical analyses.

Automation and computationally based insights are key to achieving high throughput at steadily declining costs, just as they were in DNA sequencing. Over time, as the GTL Knowledgebase matures (see 3.2.2.3.2. GTL Knowledgebase, p. 52), the GTL computational infrastructure will enable use of DNA sequence to predict the following for each protein: Efficient and successful production methods, likely binding partners, appropriate assay conditions, and, ultimately, information about the functions of each gene. Achieving this goal will require experience and the data created from production and characterization of tens of thousands of proteins.

Offices for staff, students, visitors, and administrative support will be included, as well as conference rooms and other common space. The facility will house all equipment necessary to support its mission. The DOE facility-acquisition process will include all R&D, design, and testing activities necessary to ensure a fully functional facility at the start of operations.

5.1.3. Development of Methods for Protein Production

Proteins have wide variability in their structure and stability—no single production method and characterization scheme will be applicable to every protein. Thus, several methods will be developed simultaneously, including all appropriate variations on cell-based, cell-free, and chemical synthesis.

Whichever method is selected, nearly all protein production is based on transcription from DNA obtained via cloning or possibly direct chemical synthesis of the gene encoding the desired protein. In cases where only gene sequence is available, chemical synthesis alone will be required. The Protein Production and Characterization Facility, as part of its function as a national resource, will develop a sequence-verified library of publicly available protein-coding microbial genes. This library would be available for translation into protein or for use in transformational studies by the other facilities or the larger scientific community.

Technologies should be scalable, economic, and sufficiently robust to work in a production environment. At least 50% of all proteins are anticipated to pose significant problems for any current method, so development work will be required. Some genes have evolved to generate only very small amounts of protein products. Most proteins are idiosyncratic with respect to conditions; for example, some proteins are not readily soluble or they are relatively unstable and require discovery of special conditions for storage, handling, and use. Others will function only in a properly reconstituted assembly and may need to be produced with their partners under specialized conditions. Consequently, a significant component of the facility will be research into new methods of protein production. In addition, many DOE-relevant systems may require techniques compatible with anaerobic or other extreme conditions. The strategy for success includes high-throughput parallel processing to allow exploration of a very large number of conditions and protocols specific to each protein.

Improved techniques are needed to predict from genome sequence the production and purification approaches most likely to succeed with each protein. We also need methods to identify all DNA sequences in a genome that should encode proteins. Thus, computation and informatics is an integral facility component. Algorithms based on data from successful and failed protein expressions are expected to improve future protein-production and -characterization efficiencies.

Disorder and the Formation of Molecular Machines. We need to produce these proteins in their functional state. Disorder is emerging as an increasingly important factor in protein function, particularly in the assembly of protein partners into molecular machines. This key process very often is mediated by disorder-to-order transitions at the binding interfaces as the disordered regions of two proteins become ordered by their interaction. Part of the facility's R&D effort will be to develop characterization methods that will, among other things, allow their general structure (whether ordered or disordered) to be defined and mapped. Whereas disordered protein regions are a hindrance in crystallization for classic protein crystallography techniques, our goal is to allow protein disorder to become a useful tool to predict binding partners and aspects of protein function (Dunker et al. 1998; Romero et al. 1998).

LIMS. A laboratory information management system (LIMS) will provide for machine learning from failures and successes of all facility aspects, the larger program, and other facilities. Experience-based decision making will allow selection of optimal expression, purification, storage, and characterization routes based on bioinformatics. Identification of domains that do and do not inhibit activity and strategies for affinity reagent production will be revealed. Inventory tracking and provenance records will be essential. Development will include better integration of instrument data files for generation of provenance records. For more information on LIMS and other computational and information technologies, see 4.0. Creating an Integrated Computational Environment for Biology, p. 81.

5.1.3.1. Production Targets

The initial numbers of proteins required are large by any current standard and certainly will increase over time with ongoing guidance and review from the researcher and user communities. In addition, each protein

probably will require exploration of a wide range of conditions to define successful production and characterization protocols. Several independent factors drive the need (see 2.0. Missions Overview, p. 21):

- Producing encoded proteins and characterizing them in a low-cost and high-throughput facility will make tractable and affordable the exploration of large numbers of unknown genes from sequenced microbes.
- Metagenomics is becoming more important as a methodology for studying natural systems critical to DOE mission environments. These studies are revealing millions of genes with the recurring 40% unknown ratio. Although more-sophisticated computational analyses can reduce the numbers that must be produced for analysis and for uncovering culturing techniques for some discovered microbes, potentially millions of proteins could or should be beneficially investigated through production.
- Understanding and eventually optimizing such critical microbial functions as redox processes, cellulose degradation, hydrogen production, and all the ancillary metabolic and regulatory pathways will entail screening potentially thousands of naturally occurring variants of hundreds of protein families. Exploring intentional modifications to understand function and to optimize properties could involve very large multiplicative factors on identified targets—gene shuffling can involve thousands of modifications.
- Exploring microbial function and incorporating nonnatural or isotopically labeled amino acids will be beneficial with or without various fusion tags (e.g., six-His, FlAsH tag, and biotin).
- Engineering microbial systems or biobased cell-free systems for energy or environmental applications will require significant exploration of rationally engineered primary and ancillary proteins, machines, and pathways in a concerted and comprehensive way.
- Providing a source of proteins and their characterizations from gene sequence alone would produce a rapid and cost-effective alternative to historical culturing techniques and an important knowledgebase for possible culturing experiments.

Production targets will be determined by research needs and the level of maturity of the particular protein class. Production probably will proceed at multiple scales; the first exploratory pass to determine optimum successful production protocols should be at the smallest and most rapidly executable scale, followed by scaleup of interesting ones accordingly (see sidebar, Workflow Process, p. 133). Three examples follow.

- Screening mode: Microgram quantities, semipure, $>10^4$ to 10^5proteins/year
- Macroscale: Milligram quantities, >90% pure, $>10^4$/year
- Large scale: Hundreds of milligram quantities, >95% pure, $>10^2$/year

Material and data products must be accompanied by protocols that define optimal parameters for production, activity, storage, and use of proteins. The challenge in developing the Protein Production and Characterization Facility is to use various technologies in appropriate ways to cover production needs for all proteins, including small soluble proteins, membrane proteins, multiple domain proteins, and multiprotein complexes. Detailed comparisons of these available options will be a key part of the facility R&D and design process. Table 1, p. 120, provides a summary of technology options for protein production. Table 2, p. 121, is a simplified technology development roadmap covering the necessary research, pilot, and production phases of the R&D process. Each technology application has its own set of challenges. For the easy, soluble proteins, the challenge is scaleup, while the more difficult proteins and complexes require exploration of methods to produce and stabilize them. During facility operations, continued exploration of new techniques for protein production will be needed.

5.1.3.2. Specifications for Proteins and Comparisons of Their Production Methods

Methods eventually must be capable of cost-effectively producing on demand all the proteins coded in any microbial genome for which we have sequence, including the ability to coexpress proteins and purify or reconstitute protein complexes, difficult proteins such as membrane and multidomain proteins, metalloproteins, and proteins that cannot be overexpressed in host cells. Proteins must be properly folded and

active, incorporate correct cofactors and metals, and have correct post-translational modifications. Eventually, optimized versions of proteins should be available on demand, requiring screening of only dozens rather than hundreds or thousands of candidates. Three key methods for protein production and purification are described in sections 5.1.3.2.1–5.1.3.2.4 and in Tables 1 and 2 below.

5.1.3.2.1. Comparison of Cell-Based Expression Systems

Large-scale cell-based expression systems have been used worldwide in structural genomics centers and elsewhere, with *Escherichia coli* as the mainstay system. Yeast and other eukaryotic expression systems have

Table 1. Analysis of Technology Options for Protein Production

Comparative Analyses	Technology Options					Purification
	Cell-Based			Cell-Free	Chemical Synthesis	
	E. coli	Alternative Hosts	Homologous Hosts			
Strengths	Established methods, vectors Renewable Very cost-effective for industrial-scale quantities	Some higher success rates for certain proteins	Codon bias or missing cofactor issues eliminated	Scalable Readily automated Simplified cloning HT screening under readily manipulated conditions Cofactors Labels Production of toxic proteins	Scalable Potential for automation Labels and unusual amino acids incorporated during synthesis	Some tags demonstrated as high throughput, scalable Numerous chromatography reagents available
Weaknesses	Scalability and high-throughput automation	Less developed methods, vectors Cost Not high throughput	Large efforts to develop methods, vectors, strains Scalability and high-throughput automation	Currently only spontaneous disulfide bond formation	Ligations possible at only a small number of amino acid residues Refolding required	Tag removal Tag interference
Development Targets and Needs	More strains, vectors, procedures for difficult proteins	Improved vectors, strains, procedures for difficult proteins	Procedures generalized to engineer uncharacterized microbes	Automation demonstrated Directed disulfide bond formation Difficult proteins	Protein folding problem solved Automated for high throughput	Capability to predict effects of tags Microfluidics Integration with characterization Predictive capability for best purification and storage

June 14–16, 2004, GTL Technology Deep Dive Workshop, Working Group on Genome-Based Reagents

The table above compares and contrasts strengths, weaknesses, and development needs of technologies for use in a high-throughput production environment.

been developed for proteins that fail in *E. coli*–based systems. Their use is not as readily automated as with cell-free systems. Various alternatives are contrasted and compared in the three paragraphs below.

E. coli. Use of *E. coli* for protein production is a robust technology (numerous vectors, strains, extant instrumentation infrastructure) that is relatively inexpensive. Bacterial cultures are a renewable resource (from small- to fermenter-sized cultures), and transformants can be stored indefinitely as DNA or frozen cells. Bacterial hosts can be engineered to coexpress certain proteins or chaperones. Shortcomings include scalability (the number of cultures and culture volume required); difficulty in predicting yields and solubility; product subjectability to proteolysis; costly labeling with certain isotopes; possible absence of necessary cofactors or chaperones; and necessarily large freezer storage capacity (and tracking) of transformants. Development needs include miniaturization of cultures for screening and production; improvements in methodologies and strains; and improvements for generating membrane and other difficult-to-produce proteins.

Alternative Hosts. Use of alternative hosts (yeast, *Pichia*, *Aspergillus*, insect cell lines) may permit better expression of particular proteins, but they have less-developed vector systems and strains and are more costly than bacterial and cell-free methods. In addition, they have slower growth rates compared to *E. coli*, codon-usage

Table 2. Roadmap for Development of Technologies to Produce Proteins

Objectives and Subtopics	Research	Pilots	Production	Products
Protein Production Small soluble proteins	Protocol refinement Optimization for cost-effectiveness	Scale up to 2 k/yr Protocol standards QA standards	Scale up to 25k/yr	Multiple forms of proteins Protein chips Protocols
Protein Production Membrane proteins	Detergents Refolding Novel expression systems Cell-free expression Chemical synthesis Domain identification Domain expression	Evaluate/validate expression systems Protocol standards QA standards	Automate Scale up	Multiple forms of proteins Protocols
Protein Production Multiple domain proteins Proteins with fusion tags	Refolding Novel expression systems Cell-free expression Chemical synthesis Domain identification Domain expression	Evaluate/validate expression systems Protocol standards QA standards	Automate Scale up	Multiple forms of proteins Individual protein domains Protein chips Protocols
Protein Production Multiprotein complexes (when needed for co-expression or stabilization and storage)	Binding-partner identification Refolding Novel expression systems Cell-free expression	Evaluate/validate coexpression systems Protocol standards QA standards	Automate Scale up	Multiple forms of proteins Protocols ID binding partners

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

differences, and possibly missing cofactors or chaperones. These methods require investment in heterologous host systems and improvements for producing membrane and other difficult proteins.

Homologous Hosts. Use of homologous hosts has the advantage that cofactors, accessory proteins, modifying enzymes, and chaperones are present, and codons are optimized for open reading frames. These systems are less developed, however, with uncertain scalability, slow growth rates, low yields, nonexistent or difficult genetics and transformation, and the absence of selectable markers. Furthermore, they are not feasible for proteins from currently unculturable microbes. Development needs include defining optimal growth conditions, development of vectors and transformation protocols, and improvements for producing membrane and other difficult-to-produce proteins.

5.1.3.2.2. Cell-Free Systems

Cell-free expression systems, such as those based on wheat germ or *E. coli* extracts, hold the greatest potential for full automation and hence lower costs and higher throughput. Successful efforts in Japan using these extracts have yielded hundreds to thousands of proteins per year (Kigawa et al. 1999; Sawasaki et al. 2002; Kawasaki et al. 2003; Endo and Sawasaki 2003). Having the ability to automate these systems and the potential to incorporate labeled or nonstandard amino acids adds to their value. However, these methods have not yet seen widespread use or application. A broader experience base needs to be established.

Cell-Free Methods. Amenable to robotics (and microtiter plates), cell-free methods can have either small sample-reaction volumes (30-μL reaction volumes, 30-μg yields) or large. Cell-free proteins can be produced from PCR-amplified DNA templates, eliminating extensive cloning steps and simplifying rapid testing of many construct variations, thereby making this an attractive method for high-throughput screening. Produced protein molecules exist in simpler mixtures, sometimes permitting functional assessment without purification. Multiple proteins can be coexpressed to assemble complexes. Cofactors and detergents can be added, and certain isotopes can be cost-effectively incorporated. Shortcomings include relatively expensive application, although this is expected to decrease substantially as the method becomes more widely used. Disulfide bonds must form spontaneously when reducing agents are removed. Development needs include advances in directed disulfide bond formation, replacement of cell lysates with recombinant proteins and ribosomes, and improvements in generating membrane and difficult-to-produce proteins.

5.1.3.2.3. Chemical Synthesis

Solid-state chemical synthesis is a possible approach for important proteins that fail in all DNA-based expression systems. Currently, this method can produce peptides up to 50 amino acids in length, but longer peptides are made at ever-diminishing efficiencies. Full-length proteins might be synthesized through chemical ligation of multiple peptides. This currently is a costly procedure, and refolding into active protein remains a major problem. This technique has the advantage of producing milligrams of proteins labeled by incorporation of isotopes, chemical modifications, unnatural amino acids, or other chemical groups.

Chemical Synthesis Methods. Requiring no DNA, chemical synthesis can have large yields (>50 mg) for small proteins. There is no contamination by cellular proteins, and incorporating unnatural amino acids, labels, and post-translational modifications is easy. Chemical synthesis currently is not high throughput, and it is labor intensive. It is limited to proteins shorter that 200 amino acids, and the product typically requires refolding. Development needs include cheaper production of thousands of peptides, expansion of peptide ligation sites, reliable refolding, and improvements for generating membrane and difficult-to-produce proteins.

5.1.3.2.4. Protein Purification

Protein purification after expression presents a number of challenges, particularly in a high-throughput environment. In the Protein Production and Characterization Facility, substantial reliance will be placed on experience-based informatics methods to guide the purification strategy for each protein, with the expectation of achieving significant improvement as the database expands. Automated protocols aimed at

eliminating centrifugation will be developed since this step accounts for the major bottleneck in current protein-production protocols.

Purification Methods. Methods based on affinity-purification tags permit generic protocols for purification, but tags can interfere with structure or function and tag removal may be required. Current methods are not high throughput, contaminants may be hard to eliminate, and activity may be lost during purification (i.e., loss of cofactors, denaturation). Development needs include improved instrumentation for high throughput, and the special problems of purifying and storing native membrane proteins should be addressed.

5.1.4. Development of Methods for Protein Characterization

Key and largely unique goals of the Protein Production and Characterization Facility are stabilization and extensive characterization of each produced protein under well-defined conditions, with the resulting data made easily accessible to internal and external users. Given the investment in each expressed protein and its scientific value, investigators plan to subject each to a substantial suite of assays. Measurements for thousands of proteins will be generated robotically under standardized conditions, producing voluminous data. Assays must be rapid and inexpensive, requiring miniscule protein quantities to allow data collection from a broad range of conditions. Technologies such as microfluidics and other lab-on-a-chip methods eventually will provide the required versatility and sensitivity, with attendant sample economies and speed (see sidebar, Micro- and Nanoscale Methods, this page). Some of these protocols should reveal additional functional, structural, biological, chemical, and physical insights.

Serving several purposes, characterization first supports production by validating that the right protein has been produced (without sequence or translation errors), that the protein is stable and nominally folded, and that conditions necessary for long-term stabilization and storage have been met. Subsets of these measurements will be made on all protein attempts, including those to generate only screening levels of unpurified proteins. Since no single measurement provides all the answers, suites of techniques will be employed as they are feasible and required (see Table 3, p. 124).

Once we are assured that validated and stable proteins are produced, a more complete set of biophysical and biochemical characterizations will be made as required by the particular research problem and system. According to program and facility governance, user groups and the review process will adjudicate resource allocation with cost and benefit analyses of each characterization. The more complete characterizations likely will be on a down-selected group—10 to 20% of total protein inventory. These measurements will delve more deeply into structure and function. Not all measurements necessarily will be made in this facility but possibly at other facilities or in researchers' laboratories. Various parameters that might be measured are listed below.

Micro- and Nanoscale Methods Reduce Costs and Improve Performance of High-Speed and High-Throughput Production and Analysis

Recent advances in microanalytical systems support the downscaling of many standard methods, resulting in improved performance and facilitating easier integration of multiple techniques, automation, and parallel material processing. Microfluidic technologies have been used to miniaturize such conventional technologies as chromatographic separations, protein and DNA electrophoresis, cell sorting, and affinity assays (e.g., immunoassays). These methods typically are 10 to 100 times faster (allowing analysis of unstable biological molecules), use 1/100th to 1/1000th the amount of sample and reagents (drastically lowering costs), and offer 2 to 10 times better separation resolution and efficiency than their conventional counterparts. Moreover, the ability to analyze minute amounts of sample reduces sample loss and dilution and allows characterization of low-abundance molecules or screening for exploratory protein-production methods. Microscale miniaturization also enables integration and parallelization of different biochemical processes and components and will be important for all production and analytical processes in the GTL facilities.

Table 3. Summary of Characterization Needs and Methods

Properties of Proteins and Affinity Reagents	Analytical Technologies (Computationally Informed)
Product Validation, QA/QC	
Protein production, identification Post-translational modifications Sequence of polymorphisms, isoforms Cloning artifacts Required cofactors, ligands, binding partners (combinatorial approaches) Stability (cofactors, ligands, binding partners) Folding (cofactors, ligands, binding partners) Storage and handling conditions	Mass spectrometry, affinity tag reaction (e.g., arrays, microfluidics, gels), light scattering, spectral matching (IR, UV), 1D/2D gels, liquid chromatography (e.g., affinity, ion exchange) Centrifugation, light scattering, spectroscopy methods (UV, CD) Screening level (UV-CD, dye binding, partial proteolysis/MS, isotope exchange/MS, FT-IR, SAXS/SANS, WAXS, EM) Robotic HT combinatorial methods (e.g., pH, temperature, salts, buffers, solvents), test with stability diagnostics
Biophysical and Biochemical Characterization	
Prepurification (See items below under postpurification) **Postpurification** Binding partners; identification of reconstitution conditions, intermolecular interactions (dissociation constants) Identification of monomeric or multimeric state Probe of folding landscape, identification of motifs, folding stability, thermodynamics, ordered and disordered regions Discovering substrates (orphan enzymes) Identification of cofactors (e.g., metals, NADH, ATP, ligands) Biological effect of post-translational modifications Identification of DNA and RNA binding, sequence motifs Assignment of function to proteins	HT screening: Dye binding, internal fluorescent labels, metabolite and molecular cocktails/MS (i.e., agonists and antagonists), affinity arrays, MS, biochemical and binding assays, ATP binding, kinase activity, affinity reagent effect on protein activity (neutral or inhibitory) HT, high fidelity: Dye binding, internal fluorescent labels, metabolite and molecular cocktails/MS (i.e., agonists and antagonists), affinity arrays, MS, biochemical and binding assays, ATP binding, kinase activity HT, high fidelity: UV-CD, dye binding, partial proteolysis/MS, isotope exchange/MS, FT-IR, fluorescence emission/lifetime (FlE/L), FRET, SAXS/SANS, WAXS, EM, calorimetry, size-exclusion chromatography coupled with laser light scattering (SEC-LLS) Affinity reagent on protein activity (neutral or inhibitory)
Ultimate Characterization	
Protein primary, secondary, tertiary, and quaternary structures Structural-activity relations Assignment of functions	Computational modeling and simulation Analyses from GTL facilities HT structural measurements: X-ray crystallography, NMR, cryoEM, scanning probe microscopy, FRET, single-molecule spectroscopies
Ultimate Manipulation	
Design of affinity reagents Protein and molecular machine redesign or refinement Pathway redesign Engineering into nanomaterials and devices	Computational modeling and simulation Analyses from GTL facilities Functionalization of nanomaterials, synthetic biology, directed evolution Microbial and cell-free systems design and engineering

- Screen, identify, and measure enzymatic or binding activity, cofactor state and requirements, effect of affinity reagents on proteins (e.g., epitopes, inhibitory or noninhibitory for selected activities)
- Identify agonists and antagonists
- Identify binding partners and determine affinities (dissociation constants) under a suite of conditions, including salts, buffers, pH, temperature, and aerobic or anaerobic
- Identify monomeric or multimeric state
- Identify reconstitution conditions, intermolecular interactions
- Probe the folding landscape, establish structure
- Identify motifs, folding stability, thermodynamics, ordered and disordered regions
- Discover substrates (orphan enzymes)
- Identify cofactors (metals, NADH, ATP, ligands)
- Elucidate biological effect of post-translational modifications
- Identify DNA/RNA binding and sequence motifs

Specific biochemical functions and sensitivities pertinent to DOE applications (e.g., metal reduction, proton or electron transfer, carbon reduction) will be critical. Many of these measurements can be made before the proteins have been purified and thus done in screening mode during the production process. Some measurements could be done with proteins produced to contain sensitive fluorescent probes designed to facilitate inexpensive, high-throughput characterizations with miniscule quantities of protein.

For a set of proteins selected for their unique and mission-relevant properties (e.g., hydrogen and biofuel production, carbon cycling, contaminant immobilization, sensors), the ultimate characterization suite will determine structure at the highest-possible resolution (primary, secondary, tertiary, and quaternary). This approach will use state-of-the-art national synchrotron, neutron, NMR, and electron microscopy facilities and lab-based molecular techniques. These measurements will allow the establishment of structural-activity relations and the understanding of design principles. Computation will be a key part of such analyses.

One of the facility's ultimate roles is to support the refinement and redesign of proteins and affinity reagents for a diverse suite of energy and environmental applications. It will produce and characterize the effects of a wide range of modifications to understand design principles and optimize performance. This includes design of affinity reagents spanning several approaches, not all of which may be proteins or even cellular; protein and molecular-machine redesign or refinement; pathway redesign; and the engineering of biofunctional materials into nanomaterials and devices for energy and environmental applications and research.

As the facility matures, characterizations will shift emphasis from supporting production methods to more advanced characterizations that provide finer detail on structure and function and elucidate design principles.

5.1.4.1. Requirements, Specifications for Functional Characterization Techniques, Data

Methods should be sensitive enough to work with screening-mode levels of proteins where possible and should include cost-effective and high-throughput biochemical and biophysical measurements. Individual measurements should be very inexpensive so they can be repeated under a variety of conditions to reflect salt, pH, buffer concentration, cofactors, ligands, and temperature. They also should have a low coefficient of variation to permit statistical analysis. They should be highly parallelized and scalable and provide QA/QC with feedback to the production process. Computational support will include algorithms for cherry-picking samples for retesting and optimizing activity conditions.

Much of the needed instrumentation is laboratory based (i.e., it can be located within the Protein Production and Characterization Facility). Some measurements could benefit from remote instruments like a high-brightness synchrotron or neutron source. For example, at such a synchrotron facility, high-throughput

systems (flow or robotic enabled) could be developed and evaluated as a means to provide a cost-effective platform for making certain types of valuable measurements on protein samples [e.g., small-angle X-ray scattering (SAXS) or extended range circular dichroism (or UV-CD)]. Results of such developments could be evaluated for their usefulness in the context of this facility's production goals. To take advantage of such an approach, methods would need to be developed for transporting and automating sample handling, data logging and processing, and comparison of results obtained by these methods. Results would need to be integrated with other laboratory-based measurements.

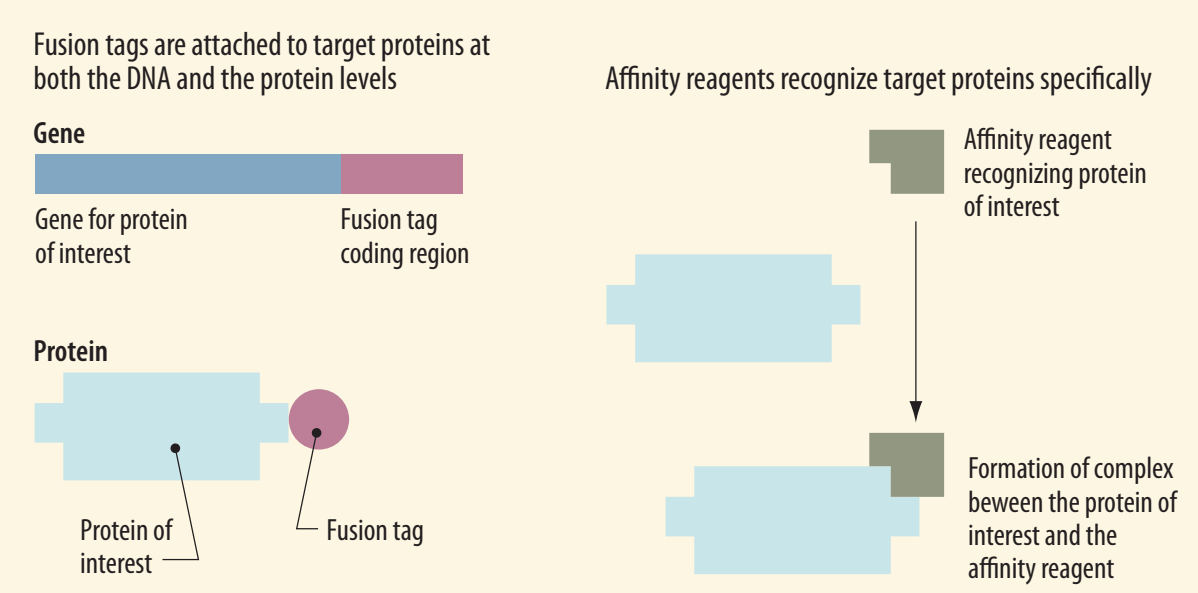
5.1.5. Development of Approaches for Affinity-Reagent Production

Production of multiple high-affinity, high-specificity affinity reagents and suitable fusion tags for each protein presents enormous challenges (see sidebar, Molecular Tags: Fusion Tags and Affinity Reagents, this page). Several promising approaches are under development worldwide, although none has yet emerged as an economical and reliable solution to GTL's high-throughput needs. Overcoming this obstacle is therefore a major target for GTL pilot studies and for this facility in particular (see Table 4, p. 127; Table 5, p. 128; Table 6, p. 128; and Table 7, p. 129).

High-throughput systems must be capable of producing numerous affinity reagents that recognize different domains of each protein. This will require multiple new libraries of affinity reagents from which members with desired affinity and specificity to each target protein can be selected. Different and complementary approaches are under development, including phage and yeast display systems and aptamers. When full proteins cannot be produced, these tags might be created for appropriate epitopes that can be determined by computational analyses. In addition, computational insights eventually might recommend the best affinity-reagent approach for particular proteins. These techniques will require substantial development.

Molecular Tags: Fusion Tags and Affinity Reagents

Fusion tags (orchid) are short peptides, protein domains, and entire proteins that are fused at the genetic level so the cell's endogenously produced proteins of interest (light blue) will have the imparted fusion tag's biochemical properties. Affinity reagents (green) are proteins, peptides, nucleic acids, and small chemical molecules that bind targets of interest with high specificity and affinity. There are many possible affinity reagents for each protein.



Further developmental areas include improved reagent stability and specificity; improved multiplex screening protocols; and rapid, high-throughput affinity-maturation techniques. Reagents also will be evaluated to determine where they bind to their protein targets and whether they disrupt the target's function, thereby dictating how different affinity reagents can be used. Development of modular affinity reagents also would be extremely useful; selected binding domains could be generated rapidly for such different purposes as protein isolation or live-cell imaging.

In many cases, the most useful affinity reagents may be proteins themselves. They can be produced and characterized using technologies already developed for bacterial proteins. They will be standardized reagents, however, so processes can be developed to allow for their rapid and large-scale production, enabling their distribution to scientists worldwide and greatly enhancing the scientific impact of reagents generated in the facility.

5.1.5.1. Specifications for Affinity Reagents and Their Production

Affinity reagent production technologies must be rapid, cost-effective, and amenable to high-throughput automation; they should be capable of being based on antibody fragments, engineered protein scaffolds, combinatorial peptides, and aptamers as the need dictates. They should work with targets that have reduced cysteines or are cell toxic. A computationally based decision process is needed for selecting proteins or epitopes of proteins to serve as targets for affinity-reagent generation. Affinity reagents should bind either individual proteins or complexes, and the collection should recognize three to five different epitopes on a protein and be amenable to epitope subtraction and existing target-detection strategies. The process should identify reagents best suited for particular applications (i.e., Western blot, pulldown, coimmunoprecipitation, staining, complex disruption, inhibited catalytic activity, and inhibited protein-protein interactions).

Table 4. Analysis of Technology Options for Affinity Reagent Production

	Phage Display	Yeast Display	Ribosome and Puromycin Display	DNA or RNA Aptamers	Animals
Strengths	Good diversity Fusion proteins	Liquid and fluorescence-based screening Affinity maturation Fusion proteins	Good diversity Fusion proteins	Good diversity	Many secondary antibodies available
Weaknesses	Slower screening Plate based	Fluorescent tags required that may complicate recognition Reduced cys on targets problematic	Slower screening	Fewer secondary affinity labels Not protein based, so no fusion proteins	Expensive Not high throughput Nonrenewable unless use mAb Slow
Development Targets and Needs	High throughput demonstrated Improved screening	High throughput Improved screening Secondary antibodies that must be developed	Optimization of scaffolds, screening methods, and automation	Optimization of screening methods	Optimization of screen methodologies DNA immunization and improvements in hybridoma production

June 14–16, 2004, GTL Technology Deep Dive Workshop, Working Group on Genome-Based Reagents

The table above compares and contrasts strengths, weaknesses, and development needs of technologies for use in a high-throughput production environment.

Table 5. Roadmap for Development of Technologies to Produce Affinity Reagents

Objectives Subtopics	Research	Pilots	Production	Products
Affinity-Reagent Library Development	Useful molecular scaffolds developed Useful libraries constructed and evaluated Design validated Expression tested System compatibility tested	Automate library Protocol standards QA, standards	Scale up	Affinity reagents Reagent chips Protocols QA, standards
Affinity-Screen Automation	Develop protocols	Scale up to 2k/year Protocol standards QA, standards	Scale up to 25k/year	Affinity reagents Reagent chips Protocols QA, standards
Affinity-Reagent Target Design	Novel vectors Validate designs	Integrate into protein production system Protocol standards QA, standards	Scale up	Immobilized targets Protein chips Protocols QA, standards

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

Table 6. Examples of Affinity Reagents and Their Applications

Examples	Applications
Obtained by Animal Immunization	
IgG and IgM	Detection, purification
Obtained by in Vitro Methods (Affinity reagents based on antibody-like proteins)	
Fab	Detection, purification, therapeutics
FV	Detection, purfication, crystallization
scFV	Detection, purfication, in vivo perturbation, therapeutics
Domain antibodies (VH, VL)	Detection, purfication, therapeutics
VHH (shark and camel heavy-chain antibody VH domains)	Detection, purfication
Fibronectin type 3 domain	Detection, purfication, in vivo perturbation
Affinity Reagents Based on Other Proteins (Scaffolds)	
Affibody (protein A)	Detection, purfication
Anticalin (lipocalin)	Detection, purfication
Ankyrin repeats	Detection, purfication, in vivo perturbation
Thioredoxin	In vivo perturbation
Affinity Reagents Based on Other Molecules	
Conbinatorial peptides	Detection, crystallization, in vivo perturbation
RNA or DNA aptamers	Detect, purification, in vivo perturbation
Small chemical molecules	In vivo perturbation

Table 7. Examples of Fusion Tags and Their Applications

Examples	Applications
Peptide Tags	
Six histidine	Purification by immobilized metal affinity chromatography (IMAC)
Epitope (e.g., myc, V5, FLAG, soft-epitope)	Detection with antibodies, purification, immunoprecipitation
StrepTag	Purification with streptavidin
S tag	Purification, detection
AviTag, Pinpoint	In vitro or in vivo biotinylation
Tandem affinity (TAP)	Purification
Tetracysteine	In vivo labeling, purification
Lanthanide-binding peptide	Labeling
Coiled-coil	Heterodimerization with partner peptide (e.g., E coil with K coil)
Metal, semiconductor, or plastic binding peptides	Immobilization on surfaces, nucleation or growth of nanocrystals, detection of semiconductor materials
Calmodulin-binding peptide	Purification (Ca^{2+} dependent)
Elastin-like peptides	Purification (temperature-dependent aggregation)
Protein Tags	
Fusion partners (glutathione-S-transferase, maltose binding protein, cellulose-binding domain, thioredoxin, NusA, mistin)	Promotion of folding, solubility, expression, or purification of fused protein
Chitin-binding domain	Promotion of folding, solubility, expression, purification, immobilization
Green fluorescent protein or alkaline phosphatase	Monitoring of expression, purification, or binding of fusion partner
Cutinase, O^6-alkylguanine alkyltransferase (AGT), or halo tag	Covalent modification for immobilization, purification, or detection
Intein	Chemical ligation in vitro or in vivo

Affinity reagents should bind their target with modest to high affinity, have lowest-possible failure rate (cross-reactivity, low affinity), be obtainable in reasonable amounts (5 mg, >90% pure) in a cost-effective manner, and be stable and storable. They should be formattable on chips with excellent shelf life and available in fluorescent, biotinylated, or enzyme-linked forms; and formattable for affinity chromatographic methods to purify individual proteins or protein complexes from cells. Ideally, they should be expressible inside cells where they can bind their target and be made conditional or regulatable.

Just as for proteins, no single method will work equally well for producing all affinity reagents, so several methods will be needed. Operationally, methods must be capable of generating reagents from small target amounts (tens of micrograms). They must readily screen diverse libraries with targets and select out the best binders applicable under a variety of conditions; have the capability to screen libraries of more than 10^9 members in a rapid manner for hundreds of targets per day; validate binding to specific target protein; and be amenable to affinity maturation.

Material and data products must be accompanied by protocols that define optimal parameters for production, activity, storage, and use. The challenge is to use various technologies in appropriate ways, including phage display, yeast display, ribosome and puromycin display, DNA or RNA aptamers, and immunization of animals. Table 4, p. 127, provides a summary of technology options for production of affinity reagents.

Table 5, p. 128, is a simplified technology development roadmap covering the necessary research, pilot, and production phases of the R&D process. Each technology application has its own set of challenges. During facility operations, continued exploration of new techniques will be needed.

5.1.5.2. Technologies for Affinity-Reagent Production

Phage Display. This technology can use libraries of combinatorial peptides, antibody fragments, and engineered protein scaffolds. Phage display is amenable to high-throughput screening with robotics; it is protein based, so functionality is added easily by creating fusion proteins with different functional domains; and it has been used for in vivo and subtractive selections. The resulting output, however, may have to go through a second round of evolution as it tends to isolate weak and strong binders at the same time. In addition, candidates should be sorted according to differences in affinity, specificity, epitope overlap, stability, storage, and application, and the output may be misleading about the strength of binding due to multivalent display. The technology may require different scaffolds, depending on the application. Development needs include the optimization of scaffolds and screening methodologies.

Yeast Display. Capable of using libraries of combinatorial peptides, antibody fragments, and engineered protein scaffolds, the yeast display technology can discriminate affinities by flow cytometry, permitting fast assessment and identifying downstream candidates. Good for directed-evolution experiments (enhanced affinity, specificity, expression, or stability) and for epitope identification, yeast display is protein based, so functionality can be added easily by creating fusion proteins with different functional domains. It may need to go through a second round of evolution, however, and its libraries tend to be less diverse than other display formats. Candidates may require sorting by affinity, specificity, epitope overlap, stability, storage, and application. Yeast grow slower than phage, taking more time and effort and needing larger volumes per screening cycle, so making this technology high throughput is more difficult. Yeast display requires different scaffolds, depending on the application. Development needs include optimization of scaffolds and screening methodologies.

Ribosome and Puromycin Display. These methods can work with very large libraries (i.e., 10^{12} members); monovalent display leads to selection of the best binders. The ribosome- and puromycin-display technologies can incorporate mutagenesis during screening and enhance binding during the general selection process. They are protein based, so functionality can be added easily by creating fusion proteins with different functional domains. They are more expensive than phage- and yeast-display technologies, however, and large libraries require more rounds of screening. Candidates need to be sorted by affinity, specificity, epitope overlap, stability, storage, and application; they require different scaffolds, depending on the application. Development needs include optimization of scaffolds and screening methodologies and automation.

DNA or RNA Aptamers. Use of DNA or RNA aptamers is amenable to very large libraries (i.e., 10^{12} members) and high-throughput screening with robotics. Synthesizing large amounts of individual aptamers is relatively expensive, however, and large libraries require more rounds of screening than phage or yeast libraries. Aptamer candidates should be sorted by affinity, specificity, epitope overlap, and application, and they are limited to DNA/RNA. Development needs include optimization of screening methodologies.

Immunization of Animals. This traditional, well-established approach requires animals and large amounts of antigen. Repeated injections are necessary, so it is slow. This is a nonrenewable resource unless hybridomas are generated, so the method is expensive; it is limited by the immune response because common epitopes cannot be subtracted. Development needs include DNA immunization and improvements in hybridoma production (see Table 4, p. 127, for strengths and weaknesses and development roadmap).

5.1.6. Development of Data Management and Computation Capabilities

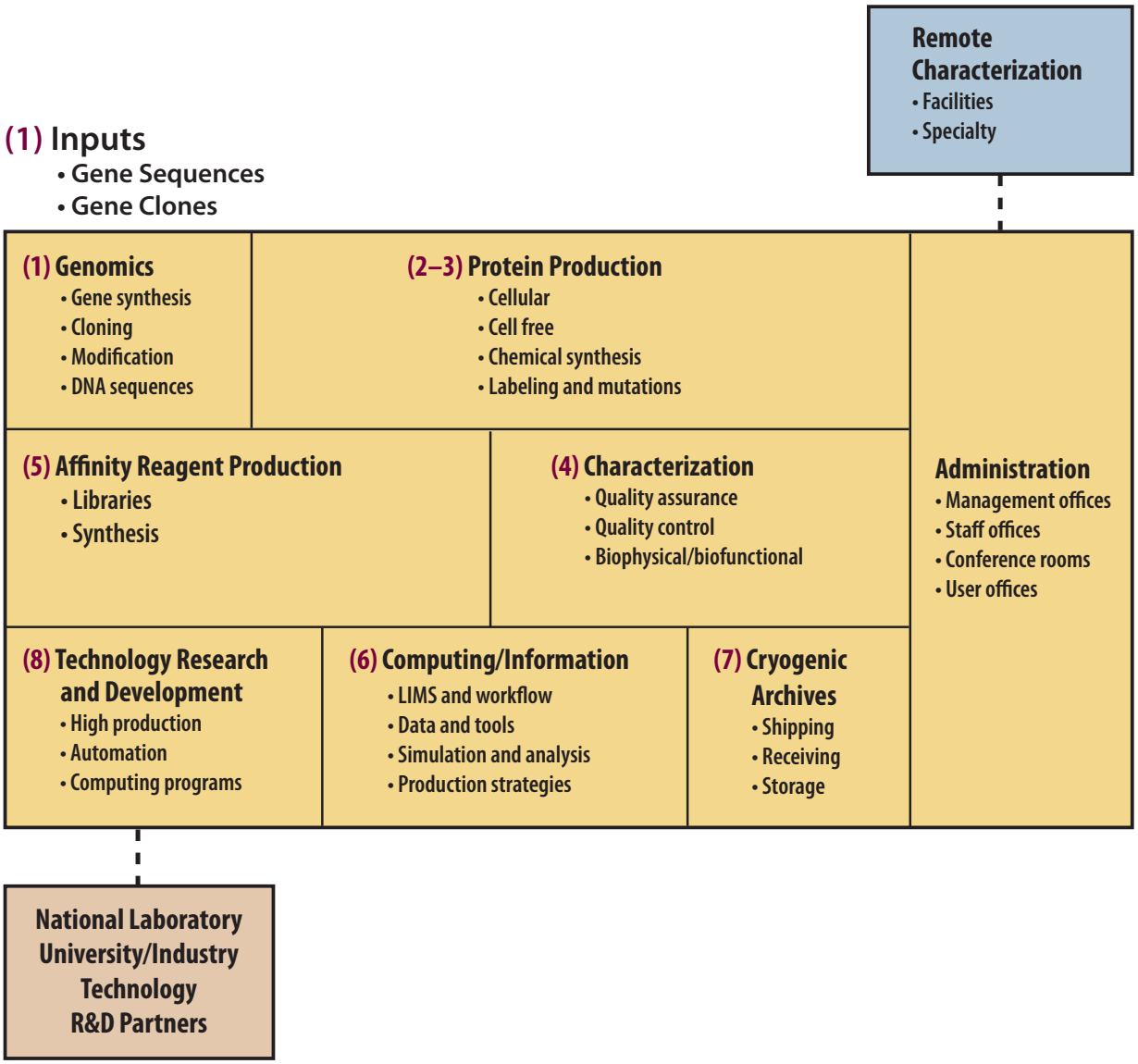
Each step and process in the Protein Production and Characterization Facility will involve very large numbers of biological samples that need to be tracked appropriately through the automated systems. Sophisticated bioinformatics analysis will be greatly needed at all steps so insights can be gained from both successes and failures. Processes will generate vast amounts of valuable data on clones and proteins and their characterization. These and other data will be captured properly and disseminated to the scientific user community. Implementation of appropriate LIMS and data-mining capabilities will be absolutely crucial to achieving high-throughput, cost-effective clone and protein production as well as to enable the use of these materials in contributing to the goals of GTL and the Department of Energy. These criteria will require large computing resources and development of the best scientific tools to properly mine the invaluable data being produced. For more details, see Table 8. Computing Roadmap, p. 132.

5.1.7. Facility Workflow Process

Conceptual diagrams, shown in the insert starting on page 133, depict prospective major facility equipment layout, process flow, and production targets. The process begins with genomics, which includes comparative genomic analyses against the GTL Knowledgebase to (1) gain insight into an unknown genome and identify its protein production targets and (2) produce clones or synthesized genes. Protein production first is pursued in a high-throughput, low-volume screening mode using appropriate microtechnologies, followed by full-scale production with successful protocols and robotics. Characterization is carried out for QA/QC, for initial biophysical and biochemical analyses, and for in-depth studies as needed. With applicable technologies, affinity reagents to selected proteins are produced using pipelines very similar to those for protein production. Computing and information technologies will support and inform all phases of facility processes and provide protocols, supporting data, and characterizations to the scientific community. The facility will have data and sample archives and distribution capabilities.

Table 8. Computing Roadmap: Facility for Production and Characterization of Proteins and Molecular Tags

Topic	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment
LIMS and Workflow Management Participate in GTL cross-facility LIMS working group	Available LIMS technologies Process description for LIMS system Crosscutting research into global workflow management systems Expert system approaches to guiding production protocols	Prototype production design strategy system Prototype protein production LIMS system Prototype biochemical characterization LIMS system Workflow management system for production and characterization Process simulation for facility workflow	Production design Protein production LIMS system Biochemical characterization LIMS system
Data Capture and Archiving Participate in GTL cross-facility working group for data representation and standards	Data models for process metadata and biophysical characterization data Technologies for large-scale storage and retrieval Preliminary designs for databases	Prototype storage archives Prototype user-access environments	Archives for key large-scale data types (e.g., biophysical characterization data) Archives linked to this facility's community databases and other GTL data resources Archives for microbial genome annotation with partners
Data Analysis and Reduction Participate in GTL cross-facility working group for data analysis and reduction	Algorithmic methods for biophysical characterization modalities Grid and high-performance algorithm codes Design for biophysical characterization tools library	Prototype biophysical characterization tools library Prototype analysis grid for biophysical characterization, with partners Analysis tools linked to data archives	Large-scale annotation systems with partners Production-analysis pipeline for biophysical characterization on grid and high-performance platforms Library with production-analysis codes Analysis tools pipeline linked to end-user problem-solving environments
Modeling and Simulation Participate in GTL cross-facility working group for modeling and simulation	Existing technologies explored for protein-fold prediction Technologies explored for low-resolution modeling from scattering data	Genome-scale protein-fold prediction, with partners Prototype code for protein modeling from scatting data	Production pipeline and end-user interfaces for genome-scale fold prediction Production codes for scattering-data modeling
Community Data Resource Participate in GTL cross-facility working group for serving community data	Data-modeling representations and design for databases: protein and reagent catalog, protein biophysical characterization, protein-production methods, and protocols	Prototype database End-user query and visualization environments Databases integrated with other GTL resources and databases	Production databases and mature end-user environments Integration with other GTL data resources Integration with other community protein-data resources
Computing Infrastructure Participate in GTL cross-cutting working group for computing infrastructure	Analysis, storage, and networking requirements for protein production facility Grid and high-performance approaches for large-scale data analysis for biophysical characterizations and establish requirements	Hardware solutions for large-scale archival storage Networking requirements for large-scale grid-based biophysical data analysis	Production-scale computational analysis systems Web server network for data archives and workflow systems Servers for community data archive databases



Workflow Process of the Protein Production and Characterization Facility

Note: Numbers and italicized words in parentheses below refer to terms used on charts beginning on next page.

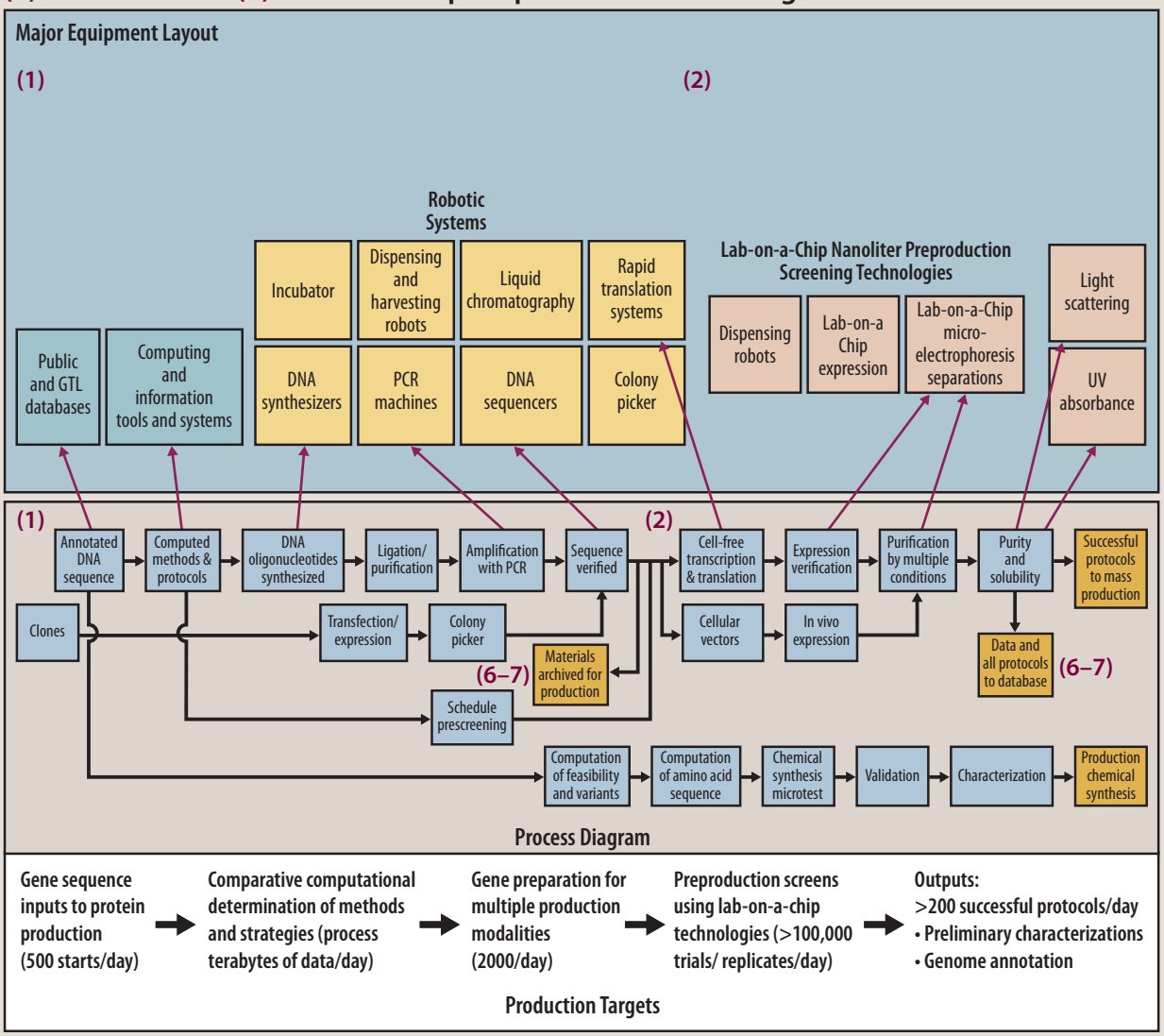
Inputs (1)

In its DNA sequence, every gene contains information needed by a cell to produce a specific protein. Scientists can use this information to make the same protein in the laboratory. The Protein Production and Characterization Facility will make proteins beginning with one of two inputs: Actual pieces of DNA that serve as molecular templates for producing given proteins (*Gene Clones*) or gene sequence information stored in databases—virtual pieces of DNA (*Gene Sequences*).

Genomics (1)

With standard techniques, the gene sequence information can be used to construct a gene clone (*Gene Synthesis*). Cloning is accomplished by inserting the synthesized DNA segment into a cloning vector, usually a specific microbe or bacterial virus designed to overexpress the protein of interest (*Cloning*). Choice of vector will vary, since all DNA sequences cannot be cloned in the same vector, nor can all proteins be produced in the same vector. In some cases, specific DNA sequence

(1) Genomics and (2) Lab-on-a-Chip Preproduction Screening Lines



modifications will be needed before insertion [e.g., to increase the resultant protein's solubility or to change the way it interacts with other proteins (*Modification*)].

Cloning and modification can introduce errors into a given DNA sequence. A critical quality-control step, one of several in the protein-production process, is verification that the gene clone's DNA sequence is correct. This process uses the high-throughput DNA sequencing technology developed as part of the Human Genome Project (*DNA Sequencing*).

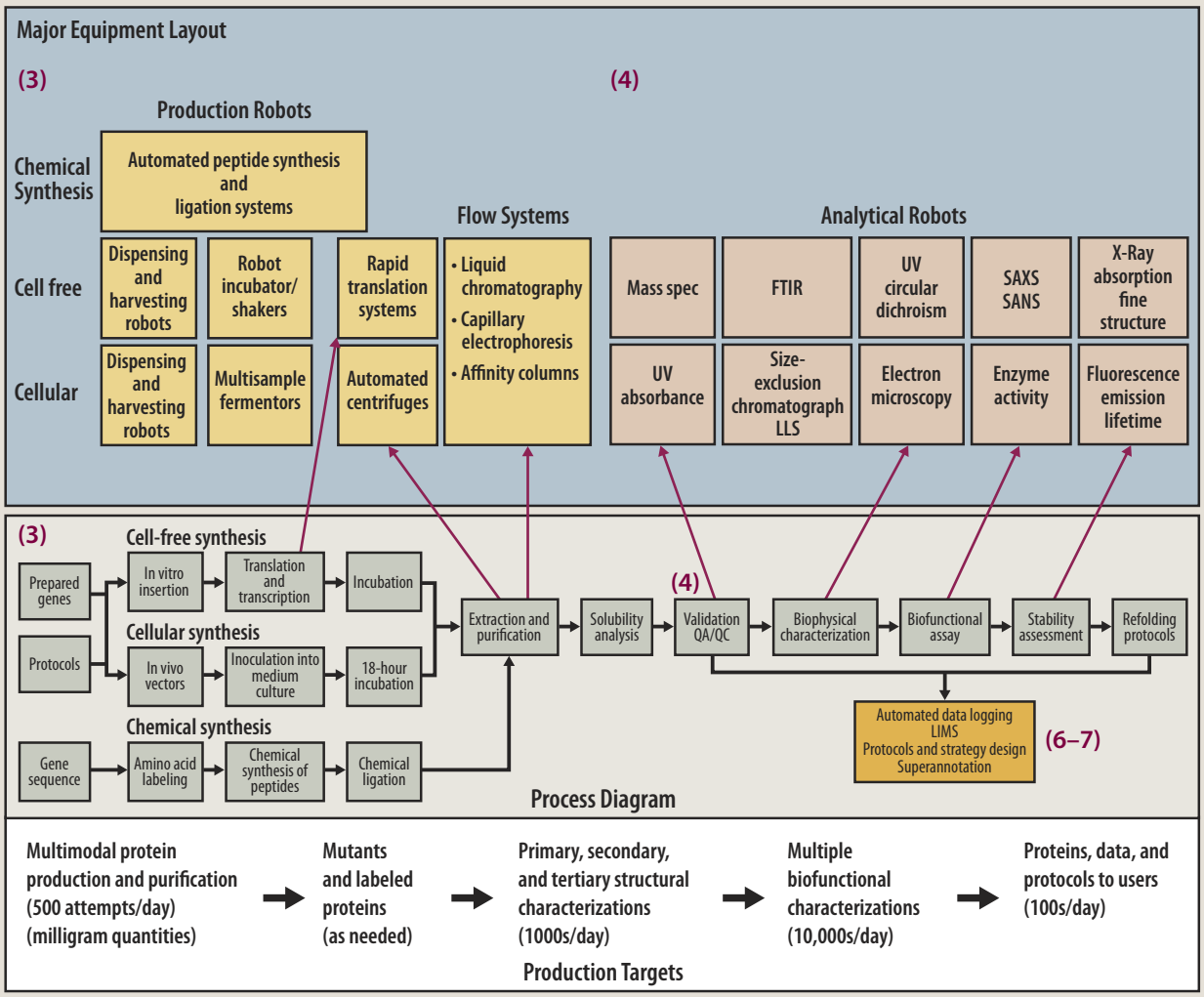
Virtually all steps in this process can be automated. A technician can obtain gene sequence information from a database and use genomics software to automatically direct a series of robots to produce a gene clone, verify the sequence, insert the clone into the appropriate vector, and produce DNA samples ready for making proteins. A laboratory can run this process simultaneously on hundreds of different target gene samples.

Protein Production (2–3)

No single method will work equally well for all proteins, so several methods will be needed to produce difference proteins from gene clones or gene sequence information.

Preproduction screening will optimize production and purification methods for each protein of interest. Various production conditions will be tested using nanoliter volumes of reagents and a "lab on a chip" on which large numbers of synthesis and analysis steps can be carried out in parallel. Robotics and microfluidic

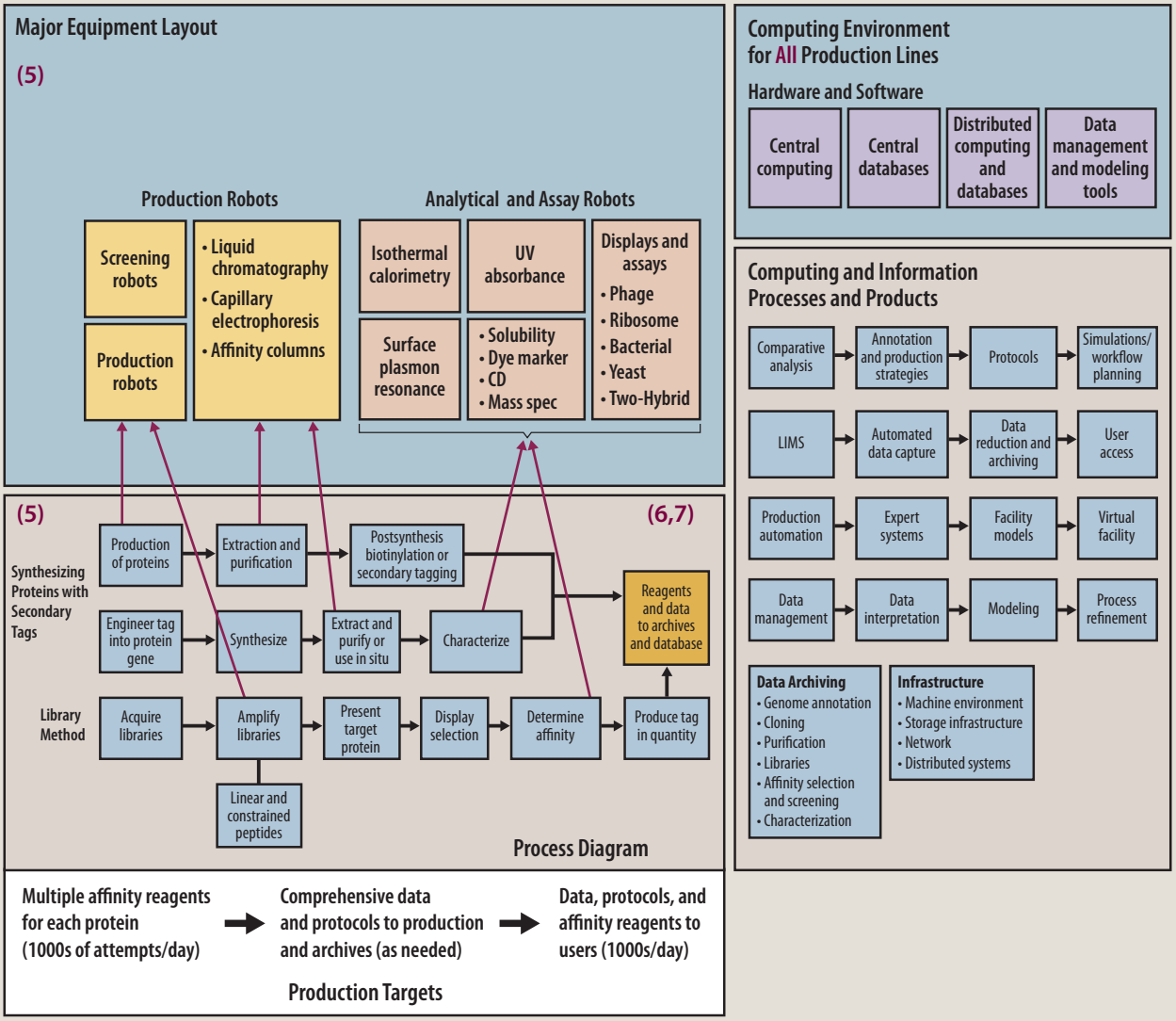
(3) Protein Production and (4) Characterization Lines



processes will be used to test various combinations of cloning vectors, reagents, and reaction conditions. The presence, level, and purity of protein expression will be checked using microchannel separations of reaction products combined with molecular-weight markers and various detection techniques (e.g., mass spectrometry, ultraviolet absorbance, and light scattering). Data will be entered into a computer and analyzed, and the best conditions and methods for large-scale protein production will be identified automatically.

During cellular protein production, vectors carrying the gene clone of interest are inserted into a bacterial host whose cellular machinery is used to produce the specific protein of interest (*Cellular Production*). The protein is extracted from the host cells and purified. Alternatively, proteins can be produced by mixing a DNA template with a set of purified enzymes and chemicals normally used by the cell for protein production; only the protein of interest is produced without the need for a living cell (*Cell-Free Production*). Finally, well-established chemical-synthesis methods can be used to make short strings of amino acids that must then be hooked together to make complete proteins (*Chemical Synthesis*). These methods, especially chemical synthesis, can be used to introduce specific changes in a protein sequence such as modification of protein subunits or incorporation of radioactive isotopes needed in downstream analysis. All proteins will undergo purification using a variety of separation technologies (e.g., liquid chromatography, capillary electrophoresis, or affinity columns). Proteins also will need to be collected and maintained under specific conditions that enable them to fold into their natural, functionally active configurations.

(5) Affinity Reagent Production Line



The protein-production process can be automated and run simultaneously on hundreds of samples to generate a vast array of normal or modified proteins ready for characterization.

Characterization (4)

In addition to verifying the sequences of gene clones, we also need to characterize the proteins produced (and the processes used to produce them) to ensure their purity and biological behavior (*Quality Control* and *Quality Assurance*).

All proteins produced will be run through a battery of tests and screening procedures (*Biophysical Characterization*) to assess their quality and to provide initial insights into their structures. For each protein, molecular weight, stability, and proper folding must be determined. No single test will be sufficient to characterize every protein adequately and accurately. Instead, a combination of various spectroscopic, separation, and imaging techniques will be used. Some proteins of particular interest to DOE, such as those involved in hydrogen production or cleanup of environmental contaminants, will be characterized further for biological function by assaying for specific enzymatic activity or binding properties.

Automated systems will simultaneously characterize hundreds of proteins for purity and, in some cases, function.

Affinity Reagent Production (5)

A very useful product of this facility will be affinity reagents that can serve as molecular markers needed to "see" the proteins in cells as parts of multiprotein complexes or as they interact with other proteins or molecules in their normal functions. Multiple affinity reagents, produced by a variety of methods, will be needed for each protein, since each reagent will recognize and bind to a particular feature (e.g., a specific physical conformation or shape as well as specific sites responsible for protein function or activity).

Affinity reagents can be produced from "libraries" of potential binders (*Libraries*). Each contains, for example, millions of different antibody-like molecules. These libraries can be screened rapidly to identify sets of affinity reagents for each protein. Proteins also can be produced or synthesized (see Protein Production above) with molecular markers or tags built into each (*Synthesis*).

Almost all steps in this process can be automated and run in parallel so millions of potential affinity reagents can be made simultaneously and hundreds of proteins can be screened against these large libraries to identify binding markers.

Computing and Information (6)

Both the production and research components of this facility need robust tools for tracking the many processes and products and associated R&D operations. A laboratory information management system (*LIMS*) is needed to track every sample and product that goes into or out of the facility and every process carried out as part of the facility (*Workflow*). LIMS will enable tracking of process efficiencies, product locations, status and availability of all facility research tools, and status of ongoing user projects. LIMS will allow facility managers and researchers to monitor production strategies (*Production Strategies*) for both proteins and molecular tags, keep track of all data generated by the facility including successes and failures, and use all that information to predict, for example, which specific strategy would be most likely to work for a given protein (*Data and Tools*). Developing these data-analysis and process-simulation capabilities will increase facility operational efficiency and reduce costs (*Simulation and Analysis)*. Moreover, the publicly available protocols of "lessons learned" will be a valuable resource that speeds progress in laboratories of scientists not physically using this facility.

Cryogenic Archives (7)

Samples (DNA, proteins, affinity reagents) used and produced by this facility will be stored for future use, shipped to current users, and received from new users (*Shipping, Receiving, Storage*). Part of the centralized LIMS, all storage, shipping, and receiving data are key components in operating this high-throughput user facility. Many aspects of sample storage and shipping are automatable.

Technology Research and Development (8)

Item 8 is illustrated on first chart only, p. 133. While technologies currently exist to carry out all production and analysis steps described above, additional research and development are needed to make each individual step more efficient, cost-effective, and part of an automated, high-production assembly line (*High Production, Automation*). Development and use of computational tools for all aspects of facility operations will be extremely important (*Computing Tools*).

5.2. Facility for Characterization and Imaging of Molecular Machines

5.2.1. Scientific and Technological Rationale 140

5.2.2. Facility Description 141

5.2.2.1. Laboratories, Instrumentation, Quality Control, Computing, and Support 141

5.2.2.2. Production Targets 143

5.2.3. Technology Development for Expression, Isolation, and Purification of Molecular Machines 143

5.2.4. Technology Development for Identification and Characterization of Molecular Machines 144

5.2.4.1. Identification of Macromolecular Complexes 145

5.2.4.1.1. Mass Spectrometry 147

5.2.4.1.2. Separation-Based Techniques 147

5.2.4.1.3. Yeast 2-Hybrid 147

5.2.4.1.4. In Vivo Imaging Technologies 148

5.2.5. Technology Development for Biophysical Characterization 149

5.2.5.1. Structural Techniques 151

5.2.5.1.1. Crystallography 151

5.2.5.1.2. CryoEM Imaging of Isolated Complexes 151

5.2.5.1.3. Nuclear Magnetic Resonance 151

5.2.5.1.4. X-Ray Scattering 152

5.2.5.1.5. Neutron Scattering 152

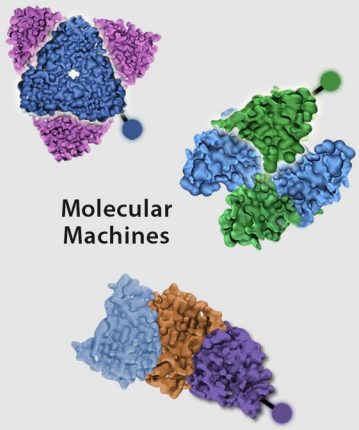
5.2.5.2. Other Biophysical Techniques 152

5.2.5.2.1. Calorimetry 152

5.2.5.2.2. Force Measurements 152

5.2.5.2.3. Mass Spectrometry for Structural Characterization 152

5.2.6. Development of Computational and Bioinformatics Tools 153



Identify and characterize molecular complexes and other interactions.

Molecular Machines

- Isolate and analyze molecular machines from microbial cells.
- Image structure and cellular location of molecular machines.
- Generate dynamic models and simulations of molecular machines.

Facility for Characterization and Imaging of Molecular Machines

The Facility for the Characterization and Imaging of Molecular Machines will be a user facility providing scientists with the basis for understanding biochemical processes in microbes by determining how molecular complexes are formed and how they function.

5.2.1. Scientific and Technological Rationale

Microbes are biological "factories" that perform and integrate thousands of discrete and highly specialized processes through coordinated molecular interactions involving assemblies of proteins and other macromolecules often referred to as "complexes" or "molecular machines." These biologically important protein-protein interactions (as well as protein-RNA, protein-DNA, and other biomolecular complexes) modify and dictate molecular states, which, in turn, integrate to define cellular physiology in response to genetic and environmental cues.

Understanding molecular machines, key players in various biochemical pathways, is central to systems biology. Many machines are short-lived or unstable and changing in composition, modification state, and subcellular location as they carry out vital functions that dictate how a cell or organism interacts with its environment. Many types of protein complexes exist in cells; complexes are associations that may be precursors to machines, associations that may not form contiguous machines, or associations that include a machine and appended molecules. A large number are assembly intermediates, while others are fully functional molecular machinery.

Key cellular multienzyme complexes can result in increased reaction rates, reduced side reactions, and direct transfer of metabolites, while many truly are machines that have moving parts or move other cellular entities (e.g., folding mechanisms and motors). So-called array machines such as light-harvesting systems, ribosomes, and others carry out intricate conversions in many organisms. Complexes also can be classified in an operational perspective from subcellular fractionation as stable and soluble, transient and soluble, and membrane associated.

As important as these machines are in cellular function, our current knowledge of them is quite limited. This is partly because proteins and other components of the complexes most often have been studied individually and in isolation and partly because they are highly

dynamic and inherently difficult to study. A cell's collection of molecular machines has intricate interrelationships that must be understood to determine how various environmental conditions influence pathways and how they differ from one organism to another. For example, specific pathways that will enhance hydrogen generation might be turned on or off by altering another pathway in an organism. We must determine the location and interactions of the molecular machines as they perform their critical functions in cells. This will require the most sophisticated and modern imaging technologies capable of resolving these details at multiple scales, from hundreds of nanometers to angstroms. Imaging technologies for identifying and locating (and collocating) machines in living cells will be incorporated into the Molecular Machines Facility. More extensive dynamic measurements that might track these machines through the life cycle of a cell will be incorporated into the Cellular Systems Facility, where the internal workings of cells will be monitored within well-defined communities and environments.

Facility Objectives

- **Discover and define the complete inventory of protein complexes in a microbe.**
- **Isolate complexes from cells using high-throughput techniques.**
- **Identify molecular components of complexes.**
- **Analyze the structure and predict the function of molecular machines. Determine basic biophysical and biochemical properties of these complexes.**
- **Validate the occurrence of complexes within cells and determine their location.**
- **Develop principles, theory, and predictive models for the structure, function, assembly, and disassembly of multiprotein complexes. Verify models with experimental data.**
- **Provide high-fidelity data and tools to the greater biological community.**

The goal of the Molecular Machines Facility is to provide researchers with the ability to isolate, identify, and characterize these functional microbial components and to validate their presence in cells using imaging and other analytical tools. The facility also will generate dynamic models and simulations of the structure, function, assembly, and disassembly of these complexes. Such efforts will provide the first step in determining how the large, dynamic network of cellular molecular processes works on a whole-system basis, how each machine is assembled in three dimensions, and how it is positioned in the cell with respect to other components of cellular architecture. Centralizing these analyses within a specialized and integrated facility will allow them to be conducted with higher performance, efficiency, fidelity, and cost-effectiveness. Many of the technologies discussed in this chapter are part of a long-lead and global development plan described in 6.0. GTL Development Summary, p. 191.

5.2.2. Facility Description

5.2.2.1. Laboratories, Instrumentation, Quality Control, Computing, and Support

The Molecular Machines Facility will have several key capabilities to provide detailed insight into the form and function of protein complexes in a cell (see Fig. 1. Core Capabilities for Molecular Machines Facility, p. 142). The high-throughput facility will consist of a 125,000- to 175,000-sq.-ft. building housing core resources for cultivation, isolation, stabilization, identification, and analysis of molecular machines as well as necessary support systems. It will have extensive robotics for efficient sample production and processing and suites of highly integrated analytical instruments for sample analysis and molecular-machine characterization.

Instrumentation in the Molecular Machines Facility will include mass spectrometry (MS) for complex identification; electron, optical, and force microscopes for in vivo and in vitro imaging, localization, and characterization of complexes; and other analytical tools. The facility will make optimum use of state-of-the-art capabilities at such national user resources as synchrotrons, neutron sources, and electron microscopes as needed. Laboratories will be required for microbial cell growth, molecular biology, automated high-throughput sample

preparation, gene expression, protein-complex analysis based on MS, imaging of protein complexes, biophysical characterization, and quality assurance. Integrated with these laboratories will be computing resources for sample tracking; data acquisition, storage, and dissemination; algorithm development; and modeling and simulation. For multiprotein machines with structurally characterized components, high-performance computing will play a very significant role in building structural models of the machines and performing molecular dynamics simulations of their intermolecular interactions. The next generation of massively parallel processors in the 40- to 100-teraflop range will allow simulations of sufficient size and fidelity to make important contributions in explaining the mechanisms of machine construction and function.

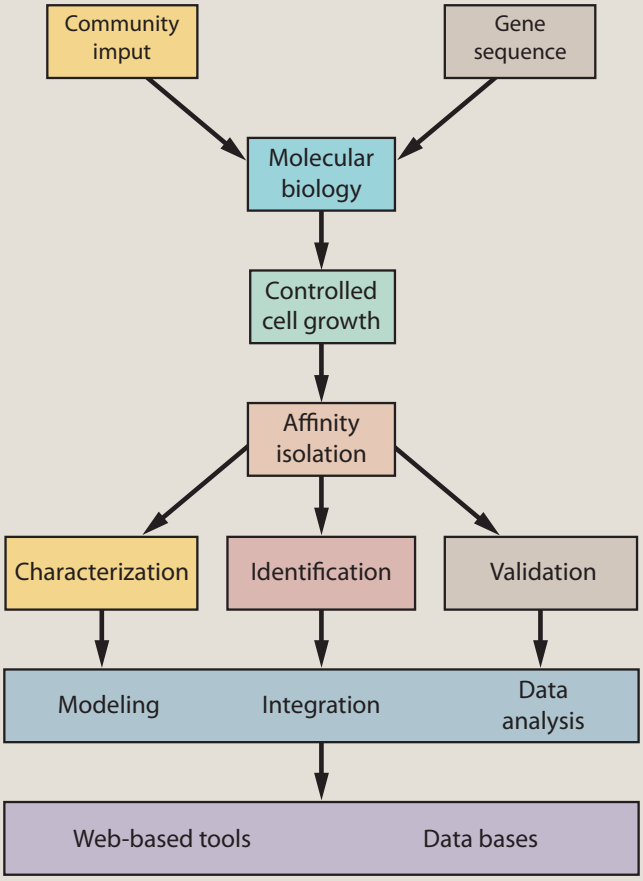


Fig. 1. Facility for Molecular Machines Analysis Core Capabilities.

Stringent quality-control protocols will be applied at each step. To get a complete picture of the complex network of molecular interactions, investigators will culture cells under a number of different conditions. They will work from insights provided by the Proteomics Facility, which will make temporal analyses to determine when and under what conditions specific proteins and machines occur. These protocols will result in potentially thousands of samples to be run through the analysis pipeline for each microorganism. Because of the diverse nature of protein complexes—stable, transient, membrane-associated, and others—multiple isolation approaches must be included. Additional technologies, especially imaging and other structural and biophysical characterization techniques, will be required to validate the machines' presence in living cells and to provide essential data that will enable insight into molecular-level interactions, kinetics, and thermodynamic properties.

The facility's computational requirements will be vast. Handling large amounts of data from diverse sources will be required, and these data must be integrated to provide a more complete view of the cell's interaction networks and to support sophisticated models of intermolecular interactions, structures, and function. In its analysis of protein machines, the facility will use the protocols and vast wealth of data on individual proteins being produced by structural genomics programs in other agencies, including the National Institutes of Health and National Science Foundation.

Offices for staff, students, visitors, and administrative support will be included, as well as conference rooms and other common space. The facility will house all equipment necessary to support its mission. The DOE facility-acquisition process will include R&D, design, testing, and evaluation activities for ensuring a fully functional facility upon completion.

5.2.2.2. Production Targets

To meet GTL program goals, researchers will need to generate protein complexes potentially involving thousands of different proteins (both natural and modified) from each organism studied. This means that many different species of microorganisms (many now unculturable) will need to be grown under a variety of carefully controlled conditions, producing millions of different protein complexes.

For a single microbe, the comprehensive mapping of the entire interactome (the summation of all protein-protein interactions in a cell) in a reasonable timeframe, with thousands of potential targets per microbe, will require throughput of the protein complex purification and identification pipeline of at least 10,000 pull-down attempts per year (to be statistically significant, these procedures must be run in triplicate and have a control, necessitating 40,000 attempts). The GTL program will need to analyze tens of microbes per year, which will require the ability to run about 100,000 pull-down attempts annually. All associated isolation, identification, and characterization procedures must be completed. The exact number of procedures will be determined by the governance processes that adjudicate the allocation of facility resources and set research and production priorities.

5.2.3. Technology Development for Expression, Isolation, and Purification of Molecular Machines

Technology must be developed to express intact protein complexes in wild-type and recombinant cultures under well-characterized conditions so molecular machines can be isolated and analyzed in initial studies as well as in those where machine functions are being optimized for specific characteristics. Maintaining high-quality, reproducible growth conditions will be essential for ensuring that high-quality data are generated. Conditions to be controlled must include environment (temperature, pH, media, substrate, light, oxygen); growth state (exponential, steady state, balanced, stationary); operation (batch, continuous); and harvest (age, lag, concentration, handling conditions). Due to the complexity of each process involved in producing the machines and the need for replicates, other quality-assurance and -control (QA-QC) techniques will be paramount to the facility's success (see Table 1. Technology Development Roadmap for Cell Growth and Processing, p. 144).

The isolation of molecular machines from cells is a challenging task. Molecular machines often are held together by weak interactions, making them fragile and difficult to isolate for analysis. Many such complexes are present only briefly or in very low amounts—sometimes just a few per cell (e.g., regulatory complexes, which are singularly important). Current techniques are inadequate for the robust, high-throughput isolation of protein complexes. The development and automation of such improved techniques is therefore an essential early goal of GTL pilot projects for this facility (see 3.3. Highlights of Research in Progress to Accomplish Milestones, p. 55). Data and reagents to be produced by the Protein Production and Characterization Facility will be central to isolating multiprotein complexes; in particular, affinity reagents would be used to isolate or "pull down" complexes (see Table 2. Technology Development Roadmap for Complex Isolation, p. 145). As a long-term goal, novel techniques for analyzing protein complexes in single microbes will be developed. The typical simplified "pipeline" for molecular-machine analysis would involve growing native or wild-type microorganisms under a reference state; using reagents to isolate the complexes from harvested cells; and then analyzing the complexes by MS, imaging, and other analytical tools. This process would be repeated under different growth states established by the Proteomics Facility, p. 155, to enable the comprehensive identification of machines chosen to be studied for the target organism.

Comprehensive identification of multiprotein complexes will require automating current methods for final sample preparation (i.e., desalting, buffer exchange, sample concentration, stabilization, and proteolytic digestion of samples). An important component of this facility is a highly integrated laboratory information management system (LIMS) that will track samples and manage data from cell cultivation through data archiving (see 5.2.6. Development of Computation and Bioinformatics Tools, p. 153).

5.2.4. Technology Development for Identification and Characterization of Molecular Machines

This facility is intended to provide detailed information on machine functions and the contributions of each to overall cell function. This analysis is a prerequisite for predicting a microbe's behavior under a range of natural and artificial conditions relevant to DOE missions. Due to the complexity and diversity of functions performed by molecular machines, multiple combinations of techniques and instrumentation must be used to identify and fully characterize all possible machines that a microbial cell is capable of producing.

Integration of multiple analytical and computational technologies will play a key role. Knowledge of a machine's static composition and structures is obtained by a variety of techniques. This information provides a starting point for following the machine's behavior in a living cell, for example, by scientists in their own laboratories and by users of the Cellular Systems Facility, p. 173. Imaging techniques can be used to follow the labeled components of a machine to trace its formation, movement, and dissociation in vivo by nondestructive techniques such as various types of fluorescence microscopy. Similarly, the high spatial resolving power of electron microscopy (EM) and X-ray microscopy can be used to localize machines in cells frozen at key functional time points. Further, X-ray and neutron diffraction and small-angle scattering can be used to help identify structural relationships among complex components.

Many analytical techniques can be used to identify and characterize proteins and protein complexes. Advantages, disadvantages, and potential areas of key method development are discussed in the sections below. Although not an exhaustive summary, they describe technology gaps that must be the subject of this facility's R&D.

Table 1. Technology Development Roadmap for Cell Growth and Processing

Technology Objectives	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
Develop technologies for protein machine production: Cell growth and processing Cultivation systems: • Modified for endogenous protein isolations • Wild type for exogenous complex isolations	Define conditions to express and process active molecular machines Develop methods: • Reproducible growth, real-time monitoring, sampling • Novel culture approaches • High-throughput controlled fermentations • Sample archive documentation • Functional assays for unknown isolated molecular machines Evaluate commercial systems	Controlled cell growth, processing: • Modified cultures for endogenous complex isolation • Wild-type microbes for exogenous complex isolation • Large numbers of microbial clones with encoded tags Database development Controlled bioreactors for cellular imaging Automation and standardization Standards, protocols, costs, QA/QC refinements Evaluation, incorporation of new technologies Development of methods for microbes and machines requiring specialized conditions	Establish high-throughput pipeline based on defined requirements, standards, protocols, costs Scale up parallel processes for multiple organisms Evaluate and incorporate new technologies Use parallel processes for scaleup	Production of well-defined microbial samples for extraction and characterization of active molecular machines Database from controlled cell growth with analysis of protein complexes and associated biocompounds Well-managed biosample archive Protocols

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

5.2.4.1. Identification of Macromolecular Complexes

The four types of macromolecular machines (each containing proteins, nucleic acids, and small biomolecules) are water-soluble stable protein-protein complexes, water-soluble transient protein-protein complexes, membrane-associated complexes, and protein–nucleic acid complexes. Water-soluble complexes typically reside inside the cell, and stable complexes can be tagged readily for isolation and characterization. Technologies for this type of system are the most developed for high-throughput analysis but are by no means sufficiently mature to be applicable to the wide range of macromolecular complexes that conduct life's processes in microbial cells. Transient complexes typically cannot be isolated from cells and therefore must either be identified while in the cell or stabilized before isolation and analysis. Complexes that last for only fractions of a second may best be hypothesized first using computational approaches but can be detected experimentally with emerging techniques. Membrane-associated complexes contain fewer polar (hydrophilic) regions, making them poorly soluble in aqueous solutions. Protein–nucleic acid complexes can fall into any of these categories. Technologies for identifying these various types of macromolecular

Table 2. Technology Development Roadmap for Complex Isolation

Technology Objectives	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
Develop high-throughput technologies for molecular machine isolation for full population of biomolecular complexes Soluble stable complexes Membrane-associated complexes Transient complexes	Define needs: • Evaluation of commercial laboratory and LIMS resources, if available • Protocol refinement, automation • QA/QC • Multistage isolation schemes using affinity reagents to minimize background interferences • Microfluidic-based affinity reagent isolations to minimize sample size requirements • Stabilization and cross-linking of less-stable and transient complexes • In vivo validation approaches Develop: • Continuous, automated processing • Multiplexed pulldowns • Novel affinity reagents and isolation schemes Develop: • Solubilization of membrane-associated complexes • Stabilization of complexes Develop stabilization and cross-linking	Pilot-scale isolation method: • Scaleup from 100 assays per week to thousands per week • Automated, continuous processing • Assessment of bottlenecks, costs • QA/QC • Evaluation and incorporation of new technologies • Methods for rapid elucidation of protein complex network linkage maps	Establishment of multiple parallel pipelines Evaluation and incorporation of new technologies	Complexes isolated to permit identification, imaging, and biophysical characterization Protocols Methods Databases and query tools

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

complexes are summarized in the following pages and in Table 3. Technology Development Roadmap for Complex Identification and Characterization, this page.

Analytical techniques for the identification and characterization of nucleic acid complexes are far less developed, in general, than those for protein-protein interactions. Many of the techniques discussed below also can be applied to this type of complex, but more development will be required, as shown in Table 3, this page.

Table 3. Technology Development Roadmap for Complex Identification and Characterization

Technology Objectives	Research, Design and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
Develop technologies for complex identification and characterization Analysis by mass spectrometry (MS): • Identification and quantification of both digested peptides and intact proteins Data processing: • Data interpretation • Data archiving	Develop for mass spectrometry: • High-throughput complex analysis by MS • Improved data analysis • Improved methods for quantitation and determination of complex stoichiometry • Improved MS detection limits and dynamic range • Identification of protein complex modifications via top-down MS • Combined isolation and identification approaches • Improved online separations • Integrated, high-sensitivity analytical tools, eventually for single cells • Improved cleavage and digestion approaches • Improved ionization for broad classes of proteins • Microfluidic-based assays Evaluate commercial hardware, software, and instrumentation	Pilot scale: • Optimization of protocols with regard to throughput, reproducibility, costs • Improved MS data-analysis tools • Database development and query tools Assays: • Integrated "lab on a chip" • Probe-based affinity • Binding affinity • Automated neutron, cryoEM, and X-ray small-angle scattering • New technologies evaluated and incorporated • MS labeling for identification of contact interfaces	Establish high-throughput, automated pipelines: • Scale up via multiple parallel production lines Refine QA/QA protocols Automate data acquisition and data analyses Evaluate and incorporate new technologies	Capability for high-throughput protein complex analysis by MS Highly validated data of identified protein complexes Confirmatory analyses of protein complexes via biophysical techniques New tools for complex-identification analysis Databases for complex identification and characterization Examples: • Binding affinity • Interaction interfaces
Biophysical Characterization Structural characterization Binding affinities Others	Establish structural and functional assays: • EM, SANS, SAXS, NMR • Approaches to identify contact faces • High-throughput binding affinity assay			

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

5.2.4.1.1. Mass Spectrometry

This technique, the workhorse analytical tool for all aspects of protein identification, can be adapted readily to the analysis of protein complexes. MS is particularly useful in identifying modified components (e.g., post-translational modifications, mutations, and others) that are important in effecting biological function. In addition, it has high sensitivity and is amenable to high-throughput analyses. Thus, MS currently is recognized as the most broadly applicable tool for large-scale identification of macromolecular complexes. Complexes must be isolated from cells before analysis, which requires the development and use of affinity reagents to isolate the target complex. Further, MS does have limitations for application to membrane-associated complexes because of the requirement to solubilize, separate, and ionize complexes before mass analysis. Although membrane-associated complexes can be solubilized in detergents and other solvents, these modified solutions are not readily adaptable to today's separation and ionization techniques typically employed with MS. Isolated membrane-associated complexes can be digested enzymatically before MS analysis of the resulting peptides, however.

Improved technologies for MS ionization, mass analysis, and detection are needed to handle the full range of complexes in cells with high sensitivity and wide dynamic range. Development in these areas should enhance the ability to analyze membrane-associated complexes. In addition, better sample-handling techniques before mass analysis, including microsample preparation and separation techniques, are necessary to improve detection limits and decrease the amount of sample. Improved methods for isolating complexes from cells are desired, especially affinity reagents and other isolation approaches that are more robust and universal. MS has great potential for quantitative determination of amounts of complexes in a cell and also for establishing complex stoichiometries; however, additional development of quantitative techniques is essential. Application of MS to transient complexes has been reported using cross-linking reagents and other approaches for stabilization, but future work should validate these approaches and make them more robust for routine use. Finally, improved computational tools are needed to provide automated MS data interpretation. Table 4. Performance Factors for Different Mass Analyzers, p. 148, compares available mass-analyzer technologies, their most common ionization modes, resolving powers, mass accuracies, and mass-to-charge ranges. Each of these techniques has some range of applicability in the Molecular Machines Facility.

5.2.4.1.2. Separation-Based Techniques

These technologies include a number of methods for characterizing and fractionating a wide range of complexes based on hydrodynamic radius. Separations are achieved, for example, via sedimentation velocity, size-exclusion chromatography, 2D electrophoretic gels, field-flow fractionation, and equilibrium dialysis. Many of these techniques are amenable to microtechnologies. Separation generally is accomplished with a range of such detection techniques as staining, fluorescence, and MS, many of which have wide-capacity capabilities and are fairly low cost. Protein components from complexes, however, can be identified only if standards are available to compare retention characteristics. The exception occurs when MS is used as the detector and the separated peaks can be identified from the resulting mass spectra. Recent developments have shown that microfluidic devices have very high peak resolving powers and very fast analysis times (seconds vs many minutes). Although additional development is required, they have the potential for analyzing components from single cells. In addition, they can be integrated with multiple sample-preparation steps, greatly decreasing the amounts of both sample and reagent needed for analysis. Simple versions of these "labs on a chip" have become commercially available and could be of immediate use for screening samples before full MS analysis is available (see Fig 2. Capturing Protein Complexes Using Fusion Tags, p. 149).

5.2.4.1.3. Yeast 2-Hybrid

These assays are applicable to any complex for which the cloned DNA encoding the machine components exists. A readily automatable technique, it provides good coverage of the various types of binary (pair-wise) interactions. It is a very good screening tool but has a number of problems with both false positives and false

negatives. The incidence of false-positive results increases as complexes become less stable; thus, the assays have limited use with transient complexes. Moreover, capabilities are needed to enhance applications to domain mapping and obtain low-order structure information. In general, this technique can be very useful as an initial screening tool before analysis by MS and other techniques.

5.2.4.1.4. In Vivo Imaging Technologies

Imaging tools can be used to provide high spatial resolution images of complexes in individual living cells. An important application of imaging tools will be to verify the formation of complexes identified by MS and map their locations in the cell as they perform their functions. Affinity reagents modified with fluorescent or other labels (depending upon detection modality) will be produced by the Protein Production and Characterization Facility. These reagents will be used to "tag" specific complex components to identify the locations of complexes within the cell and produce information on the dynamics of their assembly and disassembly. This information will provide additional insights into understanding the function of protein machines and will furnish valuable data for system-wide studies to be conducted in the Cellular Systems Facility.

Many types of imaging technologies can be employed to identify macromolecular complexes, including those based on optical, vibrational, X-ray, electron, and force microscopies. Within these general categories, some specialized techniques have specific applications to the analysis of macromolecular complexes in situ in live, fixed, or frozen cells. The strengths of imaging techniques typically include excellent detection sensitivity (in some cases, single-molecule detection) and the ability to characterize complexes in their natural environments in cells. Imaging techniques are applicable to all classes of complexes, providing that the identity of one or more components of the complex is known and that appropriate labeled molecules can be synthesized. For in situ measurements, the labeled molecules must be introduced successfully into cells in a manner that approximates natural conditions (i.e., does not interfere with protein associations and folding).

Currently, most imaging techniques are labor intensive and slow; robotics and automation, however, have the potential to provide faster sample throughput, and improved computational tools will enhance data

Table 4. Performance Factors for Different Mass Analyzers

Mass Analyzer	Most Common Ionization Modes	Resolving Power (FWHM)*	Mass Accuracy	Mass/Charge Range
Quadrupole	ESI	1000 to 2000	0.1 Da	200 to 3000
Time-of-flight (reflection or Q-TOF)	MALDI ESI	2000 to 10,000	0.001 Da	10 to 1,000,000 (200 to 4000 for Q-TOF)
Sector	ESI	5000 to 100,000	0.0001 Da	1000 to 15,000
Quadrupole ion trap	ESI	1000 to 2000	0.1 Da	200 to 4000
Linear trapping quad	ESI	1000 to 2000 (5000 to 10,000 in zoom scan mode)	0.1 Da	200 to 4000
Fourier transform ICR-MS	ESI MALDI	5000 to 5,000,000	0.0001 Da	200 to 20,000

*Full width at half maximum (FWHM) defines how close two peaks can be and still be resolved (resolving power). The mass divided by the FWHM is the resolving power.

Table 4 compares performance factors for the different mass analyzer technologies envisioned for use in the Molecular Machines Facility. Ionization modes, resolving power, mass accuracy, and mass-to-charge range are important factors qualifying these techniques for various applications.

visualization and manipulation. Issues specific to some of the techniques are summarized here, and some additional information on other imaging tools is given in Table 1, p. 144; Table 2, p. 145; and 5.4. Facility for Analysis and Modeling of Cellular Systems, p. 173.

Tagged Localization. This technique can be used with optical, X-ray, or electron microscopies to identify sets of biomolecules labeled with appropriate tags. This in situ method is applicable to live (visible-light), fixed, or frozen cells (X-ray and electron); to tagged transient complexes; and to membrane-associated complexes. Development in optics would improve instrumentation and more versatile excitation sources, and continued probe enhancement is needed. Examples of recently reported tags used with various imaging modalities are lanthanide dyes, quantum dots, nanoparticles, and tetracystein-based ligands.

Fluorescence Resonance Energy Transfer (FRET). FRET can identify pairs of biomolecules labeled with tags and provide information on biomolecular interrelationships. This in situ method is applicable to live cells, tagged transient species, and membrane-associated complexes. It is particularly good for structure and binding of extracellular ligands.

Scanning Probe Microscopy (SPM). Capable of very high spatial resolution, SPM can identify protein associations by attaching a tagged probe molecule to the scanning tip. Depending on the length of analysis time, the probe can detect single molecules and thus capture information on transient complexes. Labor intensive and slow, this technique is best suited for the study of membrane-associated complexes with whole cells or for the study of isolated complexes. The probe, for example, can be used to identify sites on a cell surface for interactions. Identification is a one-at-a-time process unless multiprobe devices, each with individual probe molecules, can be employed. Now under development, such devices hold promise for allowing this technique to be applied in a highly parallel fashion.

5.2.5. Technology Development for Biophysical Characterization

Generating isolated molecular complexes offers a unique but extremely challenging opportunity to characterize a complex with a host of biophysical techniques toward the ultimate goal of fully understanding a specific machine's structure, activity, and underlying interactions and mechanisms. Initially, a suite of well-established techniques will be employed to characterize the basic biophysical properties of an isolated

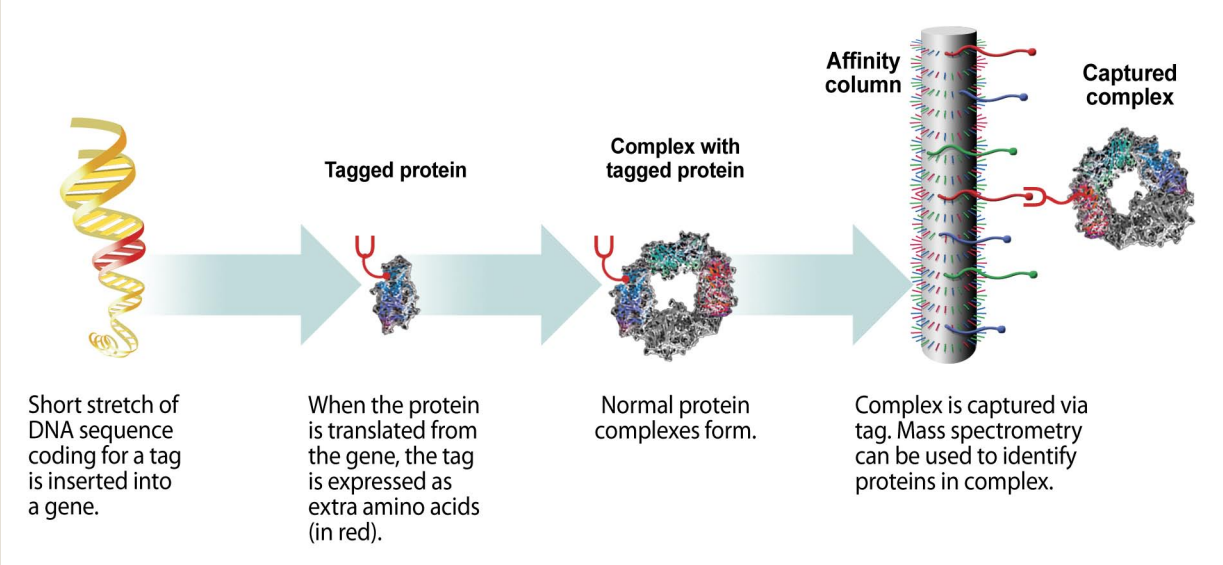


Fig. 2. Capturing Protein Complexes Using Fusion Tags.

complex. Obtaining systematic experimental information about dynamic complex behavior (including assembly and disassembly), combined with ongoing improvements in computational tools and modeling methods, will allow accurate simulations of molecular-machine activity at the heart of cellular function.

For stable protein-protein and protein–nucleic acid complexes, mechanistic understanding comes most readily with the highest levels of structural detail (general shape). Thus, atomic resolution generally is the ultimate goal in analysis of any biological structure. Crystallography and some imaging techniques offer this potential but have very specialized sample requirements and limitations, are not high throughput, and provide only a static picture of the complex.

Solution-based techniques such as cryoEM, NMR, and X-ray and neutron diffraction offer information that is lower resolution but can be related more directly to the molecule's structure in a more natural environment. Multiple tools obviously will be needed to obtain a more complete view of the structure of protein complexes, including shape, relationship of interaction faces, and stoichiometry. Three-dimensional images are obtained readily for proteins and protein complexes or machines that can be expressed, isolated, purified, and then crystallized for X-ray diffraction studies or dissolved to a sufficiently high concentration for NMR studies and scattering experiments. Such structural images have been obtained for quite large protein machines, for example, the bacterial ribosome containing some 55 proteins, additional strands of RNA, and other molecules. Some of these structural techniques are described below (see Table 5. Technology Development Roadmap for Complex Validation and Characterization, this page).

Table 5. Technology Development Roadmap for Complex Validation and Characterization

Technology Objectives	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
Develop technologies for complex validation and characterization Analysis of complexes in vitro and in vivo: • Data processing • Data archiving	Develop: • In vivo imaging for validation and spatial and temporal studies • New labels for optical microscopy • Multimodal imaging approaches • Automated image acquisition • High-throughput image analysis • Improved spatial resolution • Environmental sample-manipulation techniques Evaluate commercial hardware, software, and instrumentation	High-throughput EM High-throughput optical methods Image-analysis software Automated sample acquisition Multimodal imaging	Automate image acquisition Automate data analysis Scale up acquisition and analysis Establish database Evaluate and incorporate new technologies	Data and characterizations: • Existence of complexes • Dynamic spatial relationships of proteins and other macromolecules in complexes • Local chemical and physical environment of complexes in cells

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

5.2.5.1. Structural Techniques

5.2.5.1.1. Crystallography

X-ray crystallography is employed widely for characterizing proteins and machines. Its strengths include high structural resolution, high reliability, and practically no limit on machine size. Extending these techniques to the scale of machines is a challenge in both data collection and analysis, with problems such as phasing requiring innovations. Although synchrotrons and enhanced detectors have improved greatly the speed of analysis once a crystal is available, difficulties in sample (crystal) preparation ultimately limit throughput and applicability to protein complexes. This technique requires moderate quantities of samples with high purity. Neutron crystallography has inherently lower throughput than the X-ray technique but is the method of choice for certain types of information about protein and nucleic acid complexes. Its attributes also include high spatial resolution and practically no size limit. A particular strength of neutron crystallography is the use of hydrogen-deuterium (H/D) contrast techniques to identify locations of key hydrogen atoms. Difficulties in sample preparation and requirements for large sample quantity and purity, however, greatly limit this technique's applicability.

5.2.5.1.2. CryoEM Imaging of Isolated Complexes

Electron cryomicroscopy (cryoEM) is an emerging tool with which the 3D structure of a molecular machine in a single conformation can be determined at subnanometer resolution without requiring a crystal. Studies can be conducted at different chemical or physiological states of the molecular machines so a snapshot of mechanistic processes can be captured. The flexible docking of individual components with the medium-resolution cryoEM map can provide snapshots of the molecular machine as it is being assembled. CryoEM has been applied successfully to several different molecular machines—ribosomes, chaperonins, and ion channels; throughput, however, is slow. New generations of instrumentation allowing higher-throughput data collection will be coupled with more robust and automated image-processing software. The prospect is high that cryoEM can extend molecular-machine imaging to near-atomic resolution in a single conformation. Such advancements would allow a molecular machine's polypeptide backbone to be traced. The challenge of reaching near-atomic resolution lies in software improvement for image reconstruction.

Future excitement in studying molecular machines via structural techniques lies in the interplay among results of multiple methods for refining mechanistic models at the atomic level. For instance, dynamic motion observed via fluorescent microscopy can be used to refine cryoEM structures of a mixture of conformational states. Simulation and modeling will provide feedback to iterative refinement cycles.

Purifying molecular machines in structurally homogeneous states will be difficult because a functional machine may have flexible domains and moving parts. These dynamic characteristics of molecular machines will present a great challenge to obtaining structures of molecular machines that exist only in mixed conformational states. CryoEM can record images of molecular machines with mixed conformations at moderately high resolution. Novel software must be developed for in silico separation of molecular-machine images in different conformations. A team effort of experimental and computational scientists will be needed to tackle this problem at both algorithmic and software levels. These types of investigations will require the fastest available computers for data sorting and structural refinement (see 4.2.1.5. Structure, Interactions, and Function, p. 88).

5.2.5.1.3. Nuclear Magnetic Resonance

NMR is well suited for detailed studies of select targets in simple mixtures of small molecular assemblies. It can probe the structures of biomolecular complexes at low resolution but requires large quantities of pure complexes at relatively high concentrations (100 μM or more) free of nonspecific aggregation. Improvements are needed in data handling and analysis, sensitivity, sample throughput, and mass range (currently <100 kDa). NMR provides information on small biomolecular assemblies at atomic resolution. Of particular

importance is chemical-shift mapping and H/D exchange techniques that can be used to observe dynamics. NMR requires isotopic labeling (e.g., ^{15}N, ^{13}C, ^{2}H) to identify specific moieties within the complex. Today, NMR has limited usefulness for the analysis of large complexes above about 100 kDa, but this limitation is likely to be circumvented in the future.

5.2.5.1.4. X-Ray Scattering

This technique can be applied broadly to a range of macromolecular complexes as long as the complexes can be purified. Small angle X-ray scattering (SAXS) provides moderate-resolution information on complex structure as well as stoichiometry. Quality control and standard database infrastructure are needed. This technique has the potential of being high throughput with the development of specialized robotic sample changers on instruments at synchrotron sources and of improved data-acquisition and -analysis tools.

5.2.5.1.5. Neutron Scattering

Small-angle neutron scattering (SANS) has many of the attributes and limitations of X-ray scattering (above). An additional attribute of SANS is that H/D contrast techniques can give more insight into interaction interfaces of macromolecular complex components. Improvements needed are similar to those listed above for X-ray scattering.

5.2.5.2. Other Biophysical Techniques

A number of other biophysical techniques, both mature and developing, can be employed to obtain information on kinetics, binding affinities, interaction interfaces, and others. Some of these techniques are outlined below.

5.2.5.2.1. Calorimetry

This group of techniques assesses interactions among complex components as well as complex stability. A relatively mature technique that can be used to characterize molecular interactions, calorimetry gives a quantitative measure of thermodynamic parameters associated with the interactions. Data interpretation requires extensive analysis, which would be facilitated by computational-tool development. Calorimetry is limited by its requirement of moderately large quantities (micrograms) of pure materials, although newer techniques may reduce these amounts. Also, the samples must be monodisperse (no aggregation). This technique does have the potential to be high throughput.

5.2.5.2.2. Force Measurements

Related to force microscopy (described above in 5.2.4.1.4 under Scanning Probe Microscopy, p. 149), force measurements assess interactions among complex components using chemically modified or tagged probe tips. This technique is capable of single-molecule detection and can assess a large range of forces. It requires a specific probe for each assay, however, and is labor intensive and slow. It is in the early stages of development but eventually could be made highly parallel using multiple probe tips.

5.2.5.2.3. Mass Spectrometry for Structural Characterization

MS can provide information on biomolecular interactions at low resolution when gas-phase H/D exchange reactions are used. In that case, surfaces inaccessible to exchange do not incorporate deuterium, providing information to identify solvent-accessible surfaces and protein interfaces. MS is applicable to larger biomolecular assemblies and has high sensitivity. It is most useful when 3D structural data are available. Under development, this structural application of MS is data intensive, requiring improved data handling and interpretation techniques (see Table 5, p. 150).

5.2.6. Development of Computational and Bioinformatics Tools

The Molecular Machines Facility has great need of computational tools for sample tracking, data acquisition, data interpretation, quality assurance, modeling and simulation, and many other tasks. A wide variety of these tools are being developed, and some specific application areas are outlined below (see Table 6. Computing Roadmap: Facility for Characterization and Imaging of Molecular Machines, p. 154).

- **Data-Handling and -Integration Techniques.** Not only will huge quantities (gigabytes and more) of MS data be obtained daily, but the data from many other analytical and structural tools must be integrated to understand the (1) complex network of interacting molecules in a microbial cell and (2) temporal and spatial dynamics of these biomolecular complexes. Computational tools for MS, while developed more than for almost any other analytical technique, still need further refinement to allow truly high throughput data acquisition and interpretation. As described above, imaging tools will require improved data acquisition and processing to improve sample throughput. Once all the data are collected, strategies must be designed for archiving and distributing these data to the biological community (see Table 6, p. 154).
- **Probabilistic Sequence or Structure Techniques.** These methods require a priori knowledge of classes of biomolecular interactions, but they can be high throughput and inexpensive. This tool is not CPU limited but needs more algorithm development and continuously updated databases. Also needed is further benchmarking with actual biological applications and improvements in strategies for integrating diverse data and providing reliability estimates.
- **Genome Context Analysis.** Relying on the size and extent of genome databases in its present state, this analysis does not give reliable predictions. The technique, therefore, requires more algorithm development and benchmarking for actual biological use, along with improved strategies for integrating diverse data.
- **Function-Based Inference of Participation in Complexes.** Though inexpensive once the required algorithms and databases are in place, this technique can provide interaction data that may be difficult to access experimentally, especially on short-lived complexes. These methods are not CPU limited but need more algorithm development, continuous database improvements, and benchmarking with actual biological applications.
- **Sequence and Structural-Motif Methods for Predicting Transmembrane Regions.** Limited only by availability of sequence and structural data on these regions, the strengths and development needs of these methods essentially are the same as for function-based techniques discussed above.
- **Computational-Sequence and Structural-Motif Methods for Predicting Regulatory Sites, Nucleic Acid–Binding Domains, and Target Sequence from Protein Structure.** These methods are limited only by availability of sequence and structural data on nucleic acid–binding proteins. Inexpensive to apply once algorithms and databases are in place, the techniques are probabilistic, requiring a priori data and the development of reliability estimates. Although not CPU limited, these methods do need more algorithm development, continuous improvement of databases, and ongoing benchmarking.

Table 6. Computing Roadmap: Facility for Characterization and Imaging of Molecular Machines

Topic	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment
LIMS and Workflow Management Participate in GTL cross-facility LIMS working group	Available LIMS technologies Process description for LIMS system Crosscutting research into global workflow management systems Approaches to guiding experiment-based production protocols to optimize protein production	Prototype molecular machine characterization LIMS system* Characterization design strategy Workflow management for identification and characterization Workflow process simulation	Molecular machines LIMS and workflow system Workflow integrated with other GTL facilities and experimental strategy system
Data Capture and Archiving Participate in GTL cross-facility working group for data representation and standards	Data-type models* Technologies for large-scale storage and retrieval Preliminary designs for databases	Prototype storage archives Prototype user-access environments	Archives for key large-scale data types* Archives linked to community databases and other GTL data resources GTL Knowledgebase feedback
Data Analysis and Reduction Participate in GTL cross-facility working group for data analysis and reduction	Algorithmic methods for various modalities* Grid and high-performance algorithm codes Design for tools library Approaches for automated image interpretation in confocal light microscopy and FRET	Prototype visualization methods and characterization tools library* Prototype grid for data analysis, with partners Prototypes for automated image interpretation in confocal light microscopy/ FRET Analysis tools linked to data archives	Production-analysis pipeline for various modalities* on grid and HP platforms Automated image interpretation in confocal light microscopy, FRET Repository production-analysis codes Analysis tools pipeline linked to end-user problem-solving environments
Modeling and Simulation Participate in GTL cross-facility working group for modeling and simulation	Technologies for: Fixed and flexible docking and constrained molecular dynamics Low-resolution cryoEM data modeling and reconstruction Reconstruction of protein interaction and regulatory networks Multiscale stochastic and differential equation network models	Automated production pipeline (experimentally guided molecular docking and machine dynamics; efficient modeling methods for 3D CryoEM data reconstruction) Mature methods for reconstructing protein-interaction and regulatory networks	Production pipeline and end-user interfaces for genome-scale fold prediction Production codes for scattering-data modeling
Community Data Resource Participate in GTL cross-facility working group for serving community data	Data-modeling representations and design for databases: Protein machine catalog, protein machines models and simulations, interaction network models and simulations, protein machine methods and protocols	Prototype database End-user query and visualization environments Integration of databases with other GTL resources	Production databases and mature end-user environments Integration with other GTL resources and community protein-data resources
Computing Infrastructure Participate in GTL crosscutting working group for computing infrastructure	Analysis, storage, and networking requirements for Molecular Machines Facility Grid and high-performance approaches for large-scale data analysis for MS and image data; requirements established	Hardware solutions for large-scale archival storage Networking requirements for large-scale grid-based MS and image data analysis	Production-scale computational analysis systems Web server network for data archives and workflow systems Servers for community data archive databases

* Data types and modalities include MS, NMR, neutron scattering, X-ray, confocal microscopy, cryoEM, and process metadata. Large-scale experimental data results are linked with genome data, and feedback is provided to GTL Knowledgebase.

5.3. Facility for Whole Proteome Analysis

5.3.1. Scientific and Technological Rationale 156

5.3.2. Facility Description 158

5.3.2.1. Production Targets 158

5.3.3. Technology Development for Controlled Microbial Cultivation and Sample Processing 159

5.3.3.1. Development Needs for Cultivation 161

5.3.3.2. Development Needs for Sample Processing 161

5.3.4. Large-Scale Analytical Molecular Profiling: Crosscutting Development Needs 162

5.3.5. Technology Development for Transcriptome Analysis 162

5.3.5.1. Global mRNA Analysis 162

5.3.5.1.1. Microarray Limitations Requiring R&D 163

5.3.5.2. Small Noncoding RNA Analysis 164

5.3.5.2.1. sRNA-Analysis Development Needs 164

5.3.6. Technology Development for Proteomics 164

5.3.6.1. Methods for Protein Identification 164

5.3.6.2. Methods for Quantitation 165

5.3.6.3. Methods for Detecting Protein Modifications 166

5.3.6.4. Proteomics Development Needs 166

5.3.7. Technology Development for Metabolomics 167

5.3.7.1. Measurement Techniques 167

5.3.7.2. Metabolomics Development Needs 169

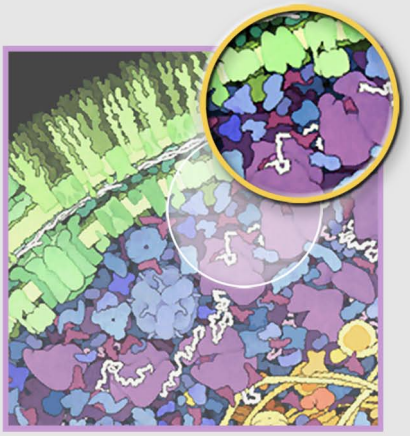
5.3.8. Technology Development for Other Molecular Analyses 169

5.3.8.1. Carbohydrate and Lipid Analyses 169

5.3.8.2. Metal Analyses 169

5.3.9. Development of Computational Resources and Capabilities 169

Facility for Whole Proteome Analysis



Identify proteins and other molecules produced by cells in response to environmental cues.

Proteomics Facility

- Measure molecular profiles and their temporal relationships.
- Identify and model key pathways and other processes to gain insights into functions of cellular systems.

The Facility for Whole Proteome Analysis (Proteomics Facility) will be a user facility enabling scientists to analyze microbial responses to environmental cues by determining the dynamic molecular makeup of target organisms in a range of well-defined conditions.

5.3.1. Scientific and Technological Rationale

The information content of the genome is relatively static, but the processes by which families of proteins are produced and molecular machines are assembled for specific purposes are amazingly dynamic, intricate, and adaptive. All proteins encoded in the genome make up an organism's "proteome." Proteins are molecules that carry out the cell's core work; they catalyze biochemical reactions, recognize and bind other molecules, undergo conformational changes that control cellular processes, and serve as important structural elements within cells. The cell does not generate all these proteins at once but rather the particular set required to produce the functionality dictated at that time by environmental cues and the organism's life strategy—a set of proteins that are produced just in time, regulated precisely both spatially and temporally to carry out a specific process or phase of cellular development.

Understanding a microbe's protein-expression profile under various environmental conditions will serve as a basis for identifying individual protein function and will provide the first step toward understanding the complex network of processes conducted by a microbe. Insight into a microbe's expression profile is derived from global analysis of mRNA, protein, and metabolite and other molecular abundance. Characterizing a microbe's expressed protein collection is important in deciphering the function of proteins and molecular machines and the principles and processes by which the genome regulates machine assembly and function and the resultant cellular function. This is not a trivial feat. A microbe typically expresses hundreds of distinct proteins at a time, and the abundance of individual proteins may differ by a factor of a million. Technologies emerging only recently have the potential to measure successfully all proteins across this broad dynamic range; these technologies and others to be further developed will form the facility's core (see Fig. 1. Proteomics Facility Flowchart, p. 157).

Measuring the time dependence of molecular concentrations—RNAs, proteins, and metabolites—is needed to explore the causal link between genome sequence and cellular function (see Fig. 2.

Gene-Protein-Metabolite Time Relationships, p. 158). Generally, a microbial cell responds to a stimulus by expressing a range of mRNAs translated into a coordinated set of proteins. Measuring RNA expression (transcriptomics) will provide insight into which genes are expressed under a specific set of conditions and thus the full set of processes that are initiated for the coordinated molecular response. An even-greater challenge will be detection of precursor regulatory proteins or signaling molecules that start the forward progression of a metabolic process. An example is master regulator molecules that simultaneously control the transcription of many genes (see sidebar, Genetic Regulation in Bacteria, p. 67). When activated and functioning, proteins expressed by RNA will yield metabolic products. Each organism has a unique biochemical profile, and measuring the cell's collection of metabolites, "metabolomics," is one of the best and most direct methods for determining the cell's biochemical and physiological status. Each of the molecular species' distinct temporal behaviors and their interrelationships must be understood. In this facility, temporal measurements—snapshots in time—will be made by taking a time series of samples from large-scale cultivations (see Table 1. GTL Data: Thousands of Times Greater than Genome Data, p. 159). The Cellular Systems Facility, by contrast, will nondestructively track processes as they happen within the microbial-community structure.

Facility Objectives

- **Identify and quantify all proteins, both normal and modified, expressed as a function of time (proteomics).**
- **Analyze all mRNA and other types of RNA (transcriptomics).**
- **Analyze all metabolites, the small biochemical products of enzyme-catalyzed reactions (metabolomics).**
- **Perform other molecular profiling. Lipids, carbohydrates, and enzyme cofactors are examples of other molecular species that can inform investigations of cellular response.**
- **Carry out modeling and simulation of microbial systems. Test models and inform experimentation, inferring molecular machines, pathways, and regulatory processes.**
- **Provide samples, data, tools, and models to the community.**

High-capacity computation is needed to integrate all the data from transcriptomics, proteomics, and metabolomics with additional information obtained from research programs and other GTL facilities. These data will be combined to understand and predict microbial responses to different intracellular and environmental stimuli. Petabytes of data generated from all these different measurements will require a substantial investment in computational tools for reducing and analyzing massive data sets and integrating diverse data types.

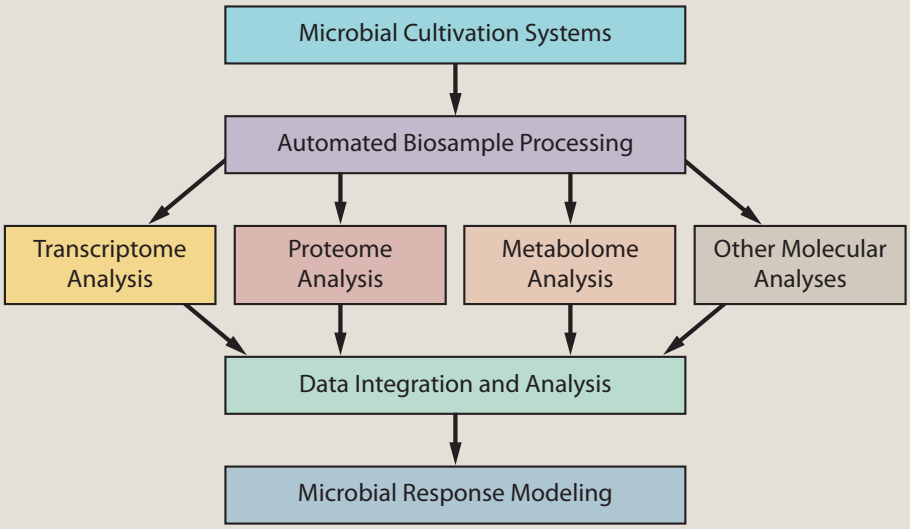


Fig. 1. Facility for Whole Proteome Analysis Core Capabilities.

5.3.2. Facility Description

This user facility will provide capabilities and supporting infrastructure to enable conceptualizing and modeling a cell's molecular response to environmental cues by identifying critical molecular changes resulting from those conditions. The Proteomics Facility, consisting of a 125,000- to 175,000-sq.-ft. building, will house core facilities for controlled growth and analysis of microbial samples. The facility's laboratories will grow microorganisms under controlled conditions; isolate analytes from cells in both cultured and environmental samples; measure changes in genome expression; temporally identify and quantify proteins, metabolites, and other cellular constituents; and integrate and interpret diverse sets of molecular data (see Fig. 1, p. 157). This high-throughput facility will have extensive robotics for efficient sample production and processing with suites of highly integrated analytical instruments for sample analysis.

The facility's computational capabilities will include data-management and -archiving technologies and computing platforms to analyze and track facility experimental data. In addition, computational tools will be established for building and refining models that can predict the behavior of microbial systems. Captured in data, models, and simulation codes, this comprehensive knowledge will be stored in the GTL Knowledgebase to be disseminated to the greater biological community, enabling studies of microbial systems biology.

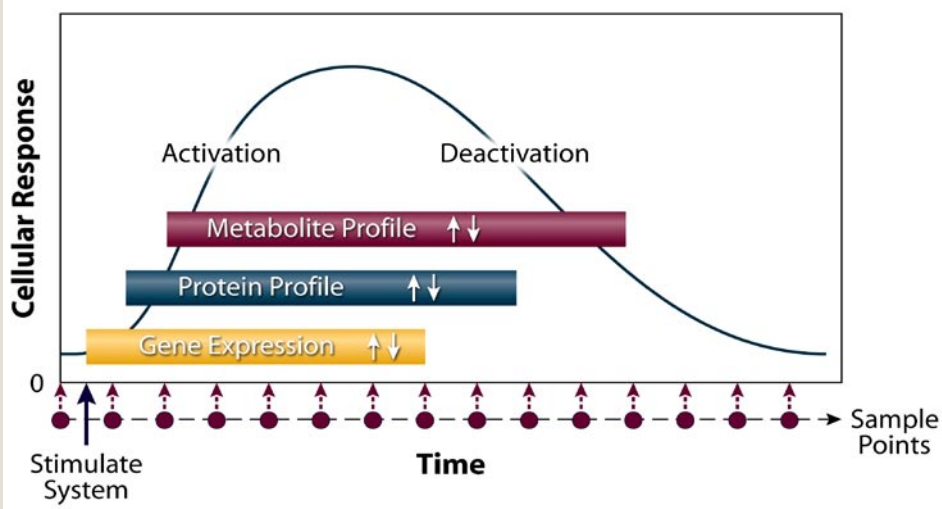


Fig. 2. Gene-Protein-Metabolite Time Relationships. To accurately establish causality between measured gene, protein, and metabolite events, sampling strategies must cover the full characteristic time scales of all three variables. Little is known about the time scale of gene, protein, and metabolite responses to specific biological stimuli or how response durations vary among genes and species. [Figure adapted from J. Nicholson et al. (2002).]

Offices for staff, students, visitors, and administrative support; conference rooms and other common space; and all the equipment necessary to support the proposed facility's mission will be included. The DOE design and acquisition process will include all R&D, design, testing, and evaluation activities necessary to ensure a fully functional facility upon completion.

5.3.2.1. Production Targets

Table 1, p. 159, illustrates the capacity needed for analyzing a single microbial experiment at various levels of comprehensiveness. This facility's goal would be to perform at least tens of such analyses per year using a phased approach, with the initial potential for that number to grow rapidly. Samples will be derived from experiments in mono- and mixed-population cultures and environmental samples.

5.3.3. Technology Development for Controlled Microbial Cultivation and Sample Processing

Automated, highly instrumented, and controlled systems will be developed for producing microbial cultures under a wide range of conditions to permit the high-throughput analysis of proteins, RNA, and metabolites. With the goal of producing and analyzing thousands of samples from single- and multiple-species cultures, technologies must be improved to provide continuous monitoring and control of culture conditions. To ensure the production of valid, reproducible samples, the Proteomics Facility must be able to grow cultures under well-characterized states, measure hundreds of variables accurately, support cultures at a scale sufficient to obtain adequate amounts of sample for analysis, and grow microbial cells in monoculture as well as in nonstandard conditions such as surfaces for biofilms (see Table 1, this page). These cultivation systems will be supported by advanced computational capabilities that allow simulation of cultivation scenarios and identification of critical experimental parameters. This facility will set the standard for cultivation, which other GTL facilities and research programs will use as starting points for their studies.

Table 1. GTL Data: Thousands of Times Greater than Genome Data

Experiment Templates for a Single Microbe

Class of Experiment	Time Points	Treatments	Conditions	Genetic Variants	Biological Replication	Total Biological Samples	Proteomics Data Volume in Terabytes	Metabolite Data in Terabytes	Transcription Data in Terabytes
Simple	10	1	3	1	3	90	18.0	13.5	0.018
Moderate	25	3	5	1	3	1,125	225.0	168.8	0.225
Upper mid	50	3	5	5	3	11,250	2,250.0	1,687.5	2.25
Complex	20	5	5	20	3	30,000	6,000.0	4,500.0	6
Comprehensive	20	5	5	50	3	75,000	15,000.0	11,250.0	15

Profiling Methods

Proteomics: Looking at a possible 6000 proteins per microbe, assuming ~200 gigabytes per sample

Metabolites: Looking a panel of 500 to 1000 different molecules, assuming ~150 gigabytes per sample

Transcription: 6000 genes and 2 arrays per sample ~100 megabytes

Typically, a single significant scientific question takes the multidimensional analysis of at least 1000 biological samples.

This table shows how quickly GTL experiments will generate terabytes (10^{12} bytes) of proteomic, metabolomic, and transcriptomic data. Global proteomics currently generates ~1.0 terabytes (TB) a day with expected 5- to 10-fold increases per year. Not only massive in volume but also very complex, these data span many levels of scale and dimensionality. For example, in a simple study of a microbial system under a single treatment (such as pH or toxin exposure), three different growth states may be studied, with ten samples taken over the growth of the culture. Replicates of each of these samples will be run as part of quality-assurance protocols. This will result in a total of 90 (3 × 10 × 3) analyses and the generation of more than 18 TB of proteomics data, 13.5 TB of metabolomics data, and 0.018 TB of transcriptomics data. If, however, a more complete set of data is taken to achieve greater temporal fidelity and better understand mechanistic response, the amount of data can grow rapidly. This example of growth in data output demonstrates one of the major data-management challenges of GTL. Strategies and technologies for data compression must be developed that avoid "data decimation," which means knowing all the information that must be extracted from raw data before any is discarded. Current proteomics efforts are employing preliminary technologies for near real-time data reduction.

Biological systems inherently are inhomogeneous; measurements of the organism's average molecular expression profile for a collection of cells cannot be related with certainty to the expression profile of any particular cell. For example, molecules found in small amounts in ensemble samples may be expressed either at low levels in most cells or at higher levels in only a small fraction of cells. Consequently, as a refinement, techniques such as flow cytometry will be used to separate various cell states and stratify cell cultures into functional classes.

Standardized, statistically sound sampling methods and quality controls are essential to ensure reproducibility and interpretability of advanced analyses. Robotics and liquid-handling systems will be developed and automated for initial isolation of proteins and other molecules from microbes, final sample preparation (e.g., desalting, buffer exchange, and sample concentration), and treatment of samples as required for analysis. Microtechnologies such as microfluidic devices will be developed wherever applicable to improve performance and speed, reduce sample handling and potential sample losses, and reduce use of materials and costs (see Table 2. Controlled Cultivation and Sample Processing Technology Development Roadmap, this page).

Table 2. Controlled Cultivation and Sample Processing Technology Development Roadmap

Technology Objectives	Research, Design, and Development	Demonstration: Pilots, Modular Deployment	Integration and Production Deployment	Facility Outputs
Controlled Cell Growth, Analysis Flexible, highly instrumented and monitored cultivation systems Online metabolite monitoring Sample preparation, characterization, stabilization Sample archiving, tracking User environments Community outreach, education	Define and determine: • Appropriate parameters, culture variability • Workflow processes • Scale factors • Hardware, software, instrumentation Develop: • Reactor and instrumentation, interfaces, sampling methods • Reactor-based growth models, simulations • Searchable sample archive • Isotope labeling • High-throughput cultivation, isolation of community members	Pilot: • High-throughput controlled cell growth, processing • Methods for large sample collection • Online analytical systems for high-throughput metabolite measurements • Experiment and sample database • Automation, standardization, protocols Develop methods: • Commensal cocultures • Extremophiles • Biofilms and structures • Sample receipt and delivery	Establish high-throughput pipeline based on defined products, standards, protocols, costs Scale up parallel processes for multiple organisms Process automated, reproducible samples Scale up user-access protocols for sample receipt, growing, delivery	Coordinated high-quality analyses of microbial samples for nucleic acids, proteins, metabolites, and others as needed Detailed cultivation and sampling parameters Efficient, high-capacity, annotated biosample archives User environment for access, protocols, process

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

5.3.3.1. Development Needs for Cultivation

- **New Technologies for Online Monitoring.** New sensors are needed to measure environmental variables, volatile and soluble metabolites, and microbial physiology to monitor and adjust conditions continually to ensure the quality of cell growth.
- **Culture Heterogeneity.** Heterogeneity is found in even the most "homogeneous" cultures produced in continuously stirred tank reactors (chemostats). Individual cells in the culture are at various stages in growth and cellular-division cycles, and subpopulations can form on reactor surfaces. Different types of culture heterogeneity also are caused by stochastic effects in microbial populations (Elowitz et al. 2002). We are just starting to develop techniques for assessing this variability and determining its impact on downstream analyses of harvested biosamples.
- **Biofilms and Structured Communities.** Emerging techniques support the growth of microbial structured communities in the form of, for example, biofilms and clusters. Even in clonal populations, the formation of structures can result in a distribution of distinct and unique phenotypes in the microniches of biofilms and other structures (see sidebar, Life in a Biofilm, p. 18).
- **Definition of Media Components and Culture Parameters.** Such culture parameters as dissolved oxygen, pH, density, and growth rate are important for interpreting the culture's metabolic responses and for providing another level of quality assurance from one experiment to another. Components of growth media influence microbial metabolism and physiology and should be defined chemically to ensure reproducibility and to account for chemical mass balance, an indicator of how the culture is processing nutrients.
- **Large Culture Volumes.** Current methods for proteomics based on mass spectrometry (MS) require large-scale cultivation for the very large number of samples required. Improvements in downstream analytical technologies, however, could reduce sample volumes and the need for such large cultures.
- **Growth in Nonstandard Conditions.** Ideal culture conditions in the laboratory should reflect community conditions in natural environments. Several microbes that DOE is studying either require extremes of salt, pH, temperature, aerobic or anaerobic conditions, and light or they exhibit certain unique phenotypes in microniches with unknown and difficult-to-characterize physicochemical states. Cultivation technologies that accommodate such a range of metabolic requirements must be considered, improved, and, in some cases, developed.

5.3.3.2. Development Needs for Sample Processing

- **Biosample Stabilization.** Harvested biosamples must reflect accurately the conditions under which they were produced. This requires the development and use of harvesting procedures that rapidly and effectively stabilize samples. For example, samples of intracellular metabolites should be quenched as quickly as possible (within a few hundred milliseconds) to maintain in vivo concentrations.
- **Sampling Time Scales.** Gene, protein, and metabolic events within cells operate on significantly different time scales. The resulting gene expression, protein synthesis, cell signaling, and metabolic responses to an environmental stimulus are related functionally but can last from milliseconds to hours. Inferred causal correlations among these different kinds of molecular events depend on well-defined temporal relationships in sampling. Having technologies and methods in place is important for accurately measuring the time-dependent patterns of change for a variety of molecular responses (see Fig. 2, p. 158).
- **Environmental Samples.** Analysis of real environmental samples will be a critical capability of this facility. As methods are refined and made more robust, examining environmental samples with their increased complexity and lack of controls will become more feasible, with protocols supporting these analyses.

5.3.4. Large-Scale Analytical Molecular Profiling: Crosscutting Development Needs

Several technological factors impact the kinds of measurements that can be made on the molecular inventories of cells: (1) limit of detection (the lowest number of molecules that can be detected), (2) dynamic range (ability to detect a low abundance of a molecular species in the presence of other more-abundant molecules), (3) sample complexity or heterogeneity, and (4) analysis throughput. All these factors must be improved to develop technologies that can make the high-throughput molecular measurements required for GTL research.

The kinds of measurements that GTL needs for systems biology will require great improvement in throughput—not just for individual instruments within an analysis "pipeline," but for the entire system. MS technologies today vary in dynamic range from about 10^3 to 10^6. Although usually adequate for proteomic measurements, this dynamic range is not sufficient for global analysis of metabolites. To explore the full range of metabolites of an individual organism today, researchers must use a time-consuming combination of technologies that makes data comparisons and analyses difficult. Another limitation of current technologies is poor detection of molecules present in low numbers. A cell may have only a few copies of some molecules with important biological effects, making them impossible to detect without substantial concentration steps before analysis.

A comprehensive understanding of microbial response can be achieved only by linking and integrating results from many different kinds of molecular analyses. Every technology and method multiplies the scale and complexity of data and analysis (see Table 1, p. 159). Computational methods for designing and managing experiments and integrating data must be part of plans for developing experimental procedures from the ground up.

Exceptional quality control, from cultivation to experimental analysis and data generation, must be maintained to ensure the most reliable data output. To draw meaningful conclusions from transcriptomic, proteomic, and metabolomic studies, researchers need data generated from protocols that have been highly validated in a process similar to that currently used in gene sequencing. This will require understanding error rates and variability in measurements and defining how many measurement replicates are needed for confident identification of biologically significant changes. Today, months are required to measure the proteome of even a simple microbial system, making replicates of proteome measurements impractical for most individual laboratories.

In addition to these crosscutting challenges to multiple analytical methods, research and development are needed for methods and technologies specific to each type of molecular analysis conducted at this facility, as described below.

5.3.5. Technology Development for Transcriptome Analysis

Large-scale RNA profiling involves quantifying and characterizing the entire assembly of RNA species present in a sample, including all mRNA transcripts (the transcriptome) and other small RNAs not translated into proteins (see Table 3. Transcriptome Analysis Technology Development Roadmap, p. 163).

5.3.5.1. Global mRNA Analysis

Microarrays have become a standard technology for high-throughput gene-expression analysis because they rapidly and broadly measure relative mRNA abundance levels. The mRNA expression patterns revealed by microarrays provide insights into gene function, identify sets of genes expressed under given conditions, and are useful in inferring gene regulatory networks. The most common types of microarrays are slide based and affixed with hundreds of thousands of DNA probes, with each probe representing a different gene. In addition to glass slides, probes can be attached to such other substrates as membranes, beads, and gels. When

the probes bind fluorescently labeled mRNA target sequences from samples, the relative mRNA abundance for each expressed gene can be determined. The more target mRNA sequence available to hybridize with a specific probe, the greater the fluorescence intensity generated from a particular spot on an array.

Data from global microarray analysis must be validated with lower-throughput, more-conventional methods such as Northern blot hybridization, as well as real-time polymerase chain reaction that can be used to benchmark these facility results for comparison to researchers' lab measurements.

5.3.5.1.1. Microarray Limitations Requiring R&D

- **Global Quantitative Expression.** Relative abundance of mRNA can be measured, but quantitation is poor.
- **Interpretations of Microarray Results.** Unexpected formation of secondary mRNA structure, cross hybridization, or other factors could produce artificially low expression levels for particular genes. In addition, gene function and regulation based entirely on mRNA expression data may miss functionally related genes not expressed together or may incorrectly predict functional relationships between genes that just happen to be coexpressed. Gene expression is a piece of the systems biology puzzle that also requires proteomic and metabolomic analyses to obtain a comprehensive understanding of gene function and genome regulation.
- **Sensitivity.** The lower limit of detection for current microarray technologies is 10^4 copies of a target molecule, which is not sufficient for many applications. Low-abundance cellular mRNA cannot be detected.
- **Time Resolution.** Today's techniques lack sufficient time resolution to measure constantly changing mRNA levels.

Table 3. Transcriptome Analysis Technology Development Roadmap

Technology Objectives	Research, Design, Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Products
High-Throughput Gene-Expression Profiling Sample processing Data processing Quantitation QA/QC	Define: • Workflow processes • Improved detection limits, reproducibility, dynamic range • Hardware, software • Lab automation, robotics • QC instrumentation, processes Develop: • Multipurpose, multiorganism array platform • In vivo testing platforms • Expression database • Commercial array applications	Expression pipeline optimization, scaleup: • Improved standards, protocols, costs Pilot: • Array processing pipeline • Expression database • In vivo testing pipeline	Establish high-throughput pipeline based on defined requirements, standards, protocols, costs, and adopted industry standards: • Array processing pipeline • Expression-experiment database • In vivo expression-testing pipeline	High-quality, comprehensive expression data linked to experiment archive and culture and sampling data

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

- **Sufficient Replicates.** Running statistically sound numbers of replicate microarray experiments can significantly decrease false-positive results and increase the statistical significance of all ensuing and coordinated experimental results.

5.3.5.2. Small Noncoding RNA Analysis

We have only begun to realize the importance of noncoding small RNA molecules (sRNAs, <350 nucleotides) in many different cellular activities. Many sRNAs are known to regulate bacterial response to environmental changes. Regulatory sRNAs can inhibit transcription or translation or even bind an expressed protein and render it inactive. Other types of sRNAs with elaborate 3D structures have catalytic or structural functions within protein-RNA machines (Majdalani, Vanderpool, and Gottesman 2005).

5.3.5.2.1. sRNA-Analysis Development Needs

- **Finding sRNA Genes.** Even with the availability of complete genomes and computational tools for sequence analysis, finding genes that code for functional sRNAs rather than proteins presents a new computational challenge. Because there are so many different types of sRNAs (with many yet to be discovered) and no genetic code to aid the prediction of sRNA transcripts, more-reliable approaches to sRNA gene discovery require further development. For example, traditional methods such as BLAST and FASTA for comparing the sequences of proteins or protein-coding genes are not as useful for sRNA sequence comparisons.
- **Detecting and Quantifying sRNAs.** Still in its infancy, sRNA analysis cannot tell us how many sRNA genes we should expect to find in a microbial genome. Without reliable sRNA sequence information, experimental screening for sRNAs is difficult. Methods must be developed to isolate various sRNAs and distinguish functional RNA molecules from nonfunctional RNA by-products of cellular activities.

5.3.6. Technology Development for Proteomics

Proteome analyses at the facility will focus on identifying and quantifying both normal and modified proteins expressed by a microbe at a particular time. The most widely used proteomics technologies today include a separation technique such as gel electrophoresis and liquid chromatography combined with detection by mass spectrometry. MS will be used to measure molecular masses and quantify both the intact proteins and peptides produced by enzymatic protein digestion (see Molecular Machines Facility, Table 4. Performance Factors for Different Mass Analyzers, p. 148). Identification of expressed proteins will require both moderate-resolution "workhorse" instruments such as quadrupole and linear ion traps as well as high-performance mass spectrometers capable of high mass accuracy, including Fourier transform ion cyclotron resonance (FTICR) and quadrupole time-of-flight (Q-TOF) mass spectrometers. Data output from these instruments will require extensive dedicated computational resources for data collection, storage, interpretation, and analysis.

Currently, few laboratories are capable of carrying out large-scale proteomics experiments. Specialized technologies needed for proteome analysis are still evolving, and no standards exist for representing proteomics data, making comparisons of results among laboratories difficult. The Proteomics Facility will be a venue for the scientific community to validate these techniques and develop cross-referenced standards. It also will be in the forefront of research into completely new techniques that have capabilities going beyond those currently available (see Table 4. Proteomics Technology Development Roadmap, p. 166). Current techniques are described in the following sections.

5.3.6.1. Methods for Protein Identification

One of two general classes of MS-based approaches for measuring the proteome, gel-based methods use two-dimensional electrophoresis (2DE) to separate complex protein mixtures by net charge and molecular mass. Proteins separated on the gel are extracted and enzymatically digested to produce peptides that can be identified with MS, typically by matrix-assisted laser desorption ionization (MALDI) combined with a TOF

instrument. Recent developments in 2DE separations under nondenaturing conditions have shown that this process yields proteins that retain structural conformations, thus preserving enzymatic activity that holds the possibility of detecting other functional characteristics.

- Increasingly, proteomic techniques use liquid-chromatography (LC) separations coupled with electrospray ionization (ESI) MS for the characterization of the separated peptides or proteins. Intact proteins or peptides generated from enzymatic digestion of proteins are analyzed by direct accurate mass measurement or by tandem mass spectrometry (MS/MS), or some combination of these approaches. MS/MS analysis can provide characteristic spectra that can be searched against databases (or theoretical MS/MS spectra) to identify proteins.
- An alternate approach takes advantage of high mass accuracy of FTICR mass spectrometers to identify proteins, substantially eliminating the need for MS/MS analysis. This approach uses accurate mass and time (AMT) tags for peptides or proteins derived from the combined use of LC separation properties and the accurately determined molecular mass of a peptide or protein. Such measurements allow a certain peptide or protein to be identified among all possible predicted peptides or proteins from a genomic sequence. A database of verified AMT tags for an organism is generated using "shotgun" LC-MS/MS methods for peptide identification as described above. Once this initial investment is made (currently less than a week of work for a single microbe), use of AMT tags can achieve much faster, more quantitative, and more sensitive analyses. These methods will be augmented by new data-directed MS approaches that allow species displaying "interesting" changes in abundances (e.g., between culture conditions), but for which no AMT tag initially exists, to be targeted for identification by advanced MS/MS methodologies (as well as generation of an AMT tag for the species). The combined result will be capabilities to broadly and rapidly characterize proteomes (Lipton et al. 2002) (see Molecular Machines Facility, Table 4, p. 148).

5.3.6.2. Methods for Quantitation

The facility will require that all proteome analyses be quantitative and that the data generated have associated levels of uncertainty so that, for example, changes in protein abundances as a result of a cellular perturbation may be determined confidently. Although MS-based techniques are excellent for protein identification, protein-quantification methods are still under development, and the most-effective approaches are not yet clear.

Challenges for quantitation using MS are related to variations in peptide or protein ionization efficiencies, possible ionization-suppression effects, and other experimental factors affecting reproducibility. Recent research has suggested that quantitative results are achievable in conjunction with LC separations by using very low flow rates with ESI. Although significant effort is needed to develop methods for routine automated measurements, the use of spiked (calibrant) peptides or proteins also provides a basis for absolute quantitation in proteome measurements. Combined with appropriate normalization methods, direct-comparison analyses to understand proteome variation after a cellular perturbation appear to be possible in the future.

In addition, highly precise quantitative measurements are feasible by analyzing mixtures of a proteome labeled with a stable isotope and an unlabeled proteome. These approaches, which introduce a stable-isotope label as an amino acid nutrient in the culture, have the advantage that high-efficiency labeling can be obtained without significant impact on the biological system. Capabilities are envisioned for absolute-abundance measurements and stable-isotope labeling for high-precision analyses that will be beneficial and complementary. In many cases, the facility will apply both methods of quantitation simultaneously to provide precise information for comparison of two different proteomes as well as intercomparison of changes across large numbers of experimental studies.

In addition to limitations in ionization, several other issues must be resolved to achieve better MS-based quantitation: Incomplete digestion of proteins into peptides, losses during sample preparation and separations, incomplete incorporation of labels into samples, and difficulties with quantifying extremely small or large proteins.

Table 4. Proteomics Technology Development Roadmap

Technology Objectives	Research, Design, Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Products
High-Throughput Protein Profiling Sample processing MS for global proteomics Other analysis techniques Data processing and analysis QA/QC	Define: • Workflow processes • Lab automation and robotics • QC instrumentation and processes • Improved detection limits, reproducibility and dynamic range • Hardware, software, instrumentation Develop methods: • Peptide identification and quantitation • Identification of protein modifications • Analysis of intact proteins, including membrane associated proteins	Whole-proteomics pipeline: • Optimization and scaleup • Improved standards, protocols, costs • Pilot of global proteomics database development • Determination of global state of modification of cellular proteins Evaluate and implement: • Hardware advances • Software advances • Instrumentation advances	Establish high-throughput pipeline based on defined standards, protocols, costs	High-quality, comprehensive proteome data linked to experiment archive and culture and sample data

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

5.3.6.3. Methods for Detecting Protein Modifications

Covalent protein modifications (e.g., phosphorylation or alkylation) and other modifications (e.g., mutations and truncations) can affect protein activity, stability, localization, and binding. The majority of cellular proteins are, in fact, modified by one or more chemical processes into their functional form. MS techniques can be used to detect and identify modified peptides. For example, when a phosphate group, lipid, carbohydrate, or other modifier is added to a protein, the modified amino acid's molecular mass changes. Any technique based on mass analysis of peptides, however, can miss modifications on peptides that are not detected. This "bottom-up" analysis recently has been complemented by a "top-down" analysis scheme in which intact proteins are analyzed by ESI FTICR MS. This top-down approach has provided greater detail on both the types and sites of these modifications. Improvements in the ability to effectively ionize a wider range of intact proteins are needed, however.

5.3.6.4. Proteomics Development Needs

- **Analyzing Intact Proteins.** Although today's MS techniques are well suited for analyzing peptides produced by enzymatic digestion of proteins, improved capabilities for the MS analysis of intact proteins are needed, especially higher molecular-weight proteins and membrane-associated proteins. In both cases, ionization is a major limitation.
- **Improving Separation Methods.** The proteome's complex, heterogeneous nature requires separation of peptides or proteins before analysis. Improved separation technologies are needed to provide higher-speed, yet higher-performance, separations. A longer-term solution may include improved MS-based approaches

that use selective ionization and ion mass selection (e.g., MS/MS, gas-phase reactions) to minimize the need for high-performance separations.

- **Improving Dynamic Range.** High-throughput MS-based analysis at the Proteomics Facility will require at least a tenfold improvement in dynamic range over today's best performance.
- **Measuring Protein Turnover Rates.** The ability to introduce stable-isotope labels (e.g., in cultures) opens the doors to global measurements of protein-turnover rates, based on the partial incorporation of stable-isotope labels observed in the isotopic distributions for peptides or proteins measured with mass spectrometers in proteome studies. These measurements reflect the rates at which proteins are being produced, destroyed, or modified; they can be expected to be complex (i.e., vary with protein subcellular localization) and provide valuable data not otherwise obtainable on important aspects of the biological systems.
- **Developing New Ionization Methods.** Ionization methods and the mechanisms underlying their variability are not well understood. New or improved methods are needed for greater ionization efficiency to extend current detection limits and more-uniform ionization to improve quantitative capabilities.
- **Developing Computing Tools and Data Standards.** Such tools are needed to handle data-analysis bottlenecks. Although commercial software packages for data interpretation are quite advanced, additional improvements are needed for automatic analysis of large volumes of data and incorporation of data into larger data structures and the GTL Knowledgebase.

5.3.7. Technology Development for Metabolomics

Metabolites are the small molecular products (molecular weight <500 Da) of enzyme-catalyzed reactions. Metabolite levels are determined by protein activities, so a comprehensive understanding of microbial systems is not possible without measuring and modeling these small molecules and integrating the information with data from proteomics and other large-scale molecular analyses.

5.3.7.1. Measurement Techniques

The high chemical heterogeneity of metabolites requires that technologies be combined to fully explore the entire metabolome of even an individual organism. This heterogeneity, however, also means that metabolome components are much more varied in nature than proteome components and therefore potentially much easier to measure (see Table 5. Global Metabolite Analysis Technology Intercomparison, p. 168). A variety of separation and MS techniques and nuclear magnetic resonance (NMR) commonly are used to measure the metabolome.

- **MS and Chromatographic Separations.** Multiple forms of MS analyzers, including TOF, quadrupole and linear ion traps, and FTICR, can be combined with different separation technologies that have a variety of advantages and disadvantages. While thin-layer chromatography and gel electrophoresis have been combined successfully with MS, the two most common approaches include gas chromatography (GC) MS and LCMS.
  - **Gas Chromatography MS.** Gas chromatography can provide high-resolution separations of many chemical compounds, and MS is a very sensitive method for detecting and quantifying most small organic compounds. For quantitative measurements, an isotopically labeled analogue of the target molecule is required for optimum measurement accuracy. A major drawback is that most metabolites are polar and thus not volatile enough to be analyzed by GC methods. These polar compounds therefore must be derivitized into less polar, more volatile forms before GCMS analysis. This approach is used widely, but the chemical-derivativization steps can decrease sample throughput and introduce sample loss.
  - **Liquid Chromatography MS.** Also used in proteomics analyses, LCMS circumvents the need for derivitization required by GCMS. Like GCMS, LCMS is highly sensitive and capable of detecting

attomoles of target compounds. LCMS, however, generally provides lower-resolution separations than GCMS, which can limit its applicability in metabolite analyses involving more than 1000 species. Recent progress in higher chromatographic separations using "ultraperformance" liquid chromatography shows the potential to provide increased chromatographic resolving power (more GC-like peak resolution) that will permit enhanced detection and quantitation capabilities with shorter run times. LC can be interfaced with a variety of mass analyzers, providing detailed information on metabolite identification at very low detection limits. As with GCMS, isotopically labeled standards are required for quantitative measurements with very high accuracy. These assays can be run on such widely available instruments as quadrupole or linear ion traps. In addition, higher-performance MS instrumentation such as FTICR can be used to obtain high mass accuracy as an aid to identify metabolites.

- **Nuclear Magnetic Resonance.** One of NMR's advantages is its noninvasive, nondestructive nature that can be used to generate metabolic profiles. By analyzing samples in a liquid state, NMR can be adapted for automation and robotic liquid handling. An important NMR limitation is sensitivity, but several methods being studied have the potential to overcome this limitation. For example, recent research has shown that angular momentum of hyperpolarizable gases like xenon can increase dramatically the number of detectable spins. This has the potential to improve NMR sensitivity by a factor of 20,000. Interfacing

Table 5. Global Metabolite Analysis Technology Intercomparison

	GC-MS	LC-MS	NMR
Strengths	Highly sensitive detection of small, nonpolar organic compounds Robust Highly reproducible Well-developed databases Well-established techniques for quantitative measurements Use of high-performance mass analyzers, such as FTICR, to provide accurate mass measurement and minimize the need for separations	Highly sensitive detection High throughput Minimized need to derivitize molecules prior to analysis Potential for single-cell analysis Use of high-performance mass analyzers, such as FTICR, to provide accurate mass measurement and minimize the need for separations	Structural information provided Nondestructive Direct analysis of liquids Highly reproducible Automatable Dynamic range similar to MS
Weaknesses	Derivitizing less volatile metabolites lowering throughput and introducing potential for sample loss Difficult to discover new compounds	Poor analytical reproducibility in multivariate setting Ion suppression and matrix effects Lower resolving power than GC, leading to poor separation of molecules in complex matrices	Sensitivity Resolution Limited application to complex mixtures
Development Needs	Robustness Improved chromatographic resolving power Improved dynamic range Metabolite databases Computational tools for predicting metabolites		Robustness Dynamic range Cryogenic probes Microprobes and nanoprobes Robust interfaces with chromatography

The table above compares and contrasts strengths, weaknesses, and development needs of technologies for use in a high-throughput production environment.

NMR with chromatographic methods such as LC can resolve molecular species that usually are overlapped in the spectra, thus improving detection and structural assignments.

- **Metabolic Flux Analysis (MFA).** MFA is used to quantify all the fluxes in a microorganism's central metabolism. To measure metabolic fluxes, a ^{13}C-labeled substrate is taken up by a biological system and distributed throughout its metabolic network. NMR and MS technologies then can measure labeled intracellular metabolite pools. Intracellular fluxes are calculated from extracellular and intracellular metabolic measurements. Currently, MFA can be applied only to a highly controlled, constantly monitored system in a stationary metabolic state. MFA's main benefit is the generation of a flux map to identify targets for genetic modifications and formulate hypotheses about cellular-energy metabolism.

5.3.7.2. Metabolomics Development Needs

- **Defining Metabolic Data Standards.** Currently, methods are not standard for formatting, storing, and representing metabolic data.
- **Developing Standardized, Comprehensive Databases of Metabolites.** Although many of the most common metabolites are catalogued and commercially available, the most biologically interesting molecules are unknowns produced by metabolic reactions unique to specific organisms or organism interactions.
- **Developing Methods for Studying Multimetabolite Transport Processes.** Transporters regulate metabolic concentrations just as much as enzymes in some cases.

Table 5, p. 168, compares and contrasts the strengths, weaknesses, and development needs of technologies discussed above. Table 6. Metabolite Profiling Technology Development Roadmap, p. 170, outlines steps in preparing the appropriate mix of these technologies for a high-throughput production environment.

5.3.8. Technology Development for Other Molecular Analyses

5.3.8.1. Carbohydrate and Lipid Analyses

Macromolecules such as lipids and carbohydrates make up cell surface and structural components, impact the function of proteins through covalent modifications, and, as substrates and products of enzyme activities, serve as key indicators of active metabolic pathways. Organic and metallic cofactors, present in many molecular machines, play essential roles in protein folding, structure stabilization, and function. Some current technologies used to analyze these molecules include LC, MS, and NMR. Methods for lipid analysis are mature, but new technologies for carbohydrate analysis are needed. A major obstacle will be to distinguish among many different chemical entities with similar properties and isomers.

5.3.8.2. Metal Analyses

Metal ions are present in many molecular machines relevant to DOE missions. Technologies are needed for measuring metal abundance, coordination state, levels of metalloproteins, and metal trafficking in cells and communities. Current metal-analysis technologies include optical emission and absorption, inductively coupled plasma (ICP) MS, X-ray spectroscopy, electrochemistry, and others. They are relatively mature compared with other global analyses but may need further development to meet the facility's specific needs.

5.3.9. Development of Computational Resources and Capabilities

Computing will be an integral part of all activity within this facility: Managing workflow, controlling instruments, tracking samples, capturing bulk data and metadata from many different measurements, analyzing and integrating diverse data sets, and building predictive models of microbial response. Databases and tools will be created to give the scientific community free access to all data and models produced by the facility (see Table 7. Computing Roadmap, p. 171).

Table 6. Metabolite Profiling Technology Development Roadmap

Technology Objectives	Research, Design, Development	Demonstration: Pilots, Modular Deployment	Integration, Production Deployment	Products
High-Throughput Metabolite Profiling LC/MS and NMR methods for metabolite discovery Sample processing Data processing Analysis, quantitation, QA/QC	Define requirements: • Workflow processes • Detection limits, reproducibility, dynamic range • Lab automation, robotics, QC • Robust LC-NMR techniques • Hardware, software, instrumentation Develop methods: • Identification and quantitation • QA/QC with metrics • Sample processing	Establish pilot metabolite-profiling pipeline Optimize, scale up Develop improved standards, protocols, costs	Establish high-throughput pipeline based on defined requirements, standards, protocols, costs	High-quality, comprehensive, metabolite-profiling data linked to experiment archive and culture and sampling data

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

- State-of-the-art systems for tracking and maintaining accurate metadata for all experimental samples (e.g., culturing details, sample-processing methods used).
- High-performance computational tools and codes for efficiently collecting, analyzing, and interpreting highly diverse data sets (e.g., MS data for proteins and metabolites, microarrays, and 2DE gel images). Tool capabilities, including data clustering, expression analysis, and genome annotation, would be linked closely to advances in computing infrastructure being proposed by DOE.
- Databases, biochemical libraries, and software for interpreting spectra and identifying peptides and metabolites. Mass spectra for most metabolites are not in standard libraries. Organism-specific metabolic databases are needed.
- Computational tools for abstracting network and pathway information from expression data and genome annotation. These tools will be used for building mathematical models that represent subcellular systems responsible for protein expression and proteome state (including modified proteins) as a function of conditions. Simulation would be employed to evaluate the state of knowledge contained in these models and validate the accuracy of experimental parameters.
- Database development for expression measurements, metabolome measurement, and networks and pathway systems, models, and simulation codes that may exceed petabytes.

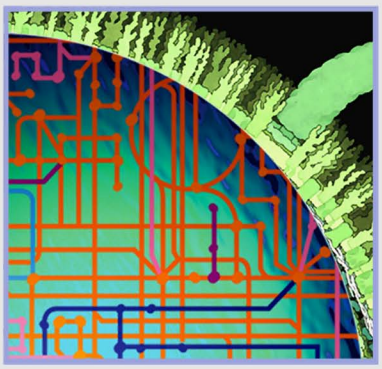
Table 7. Computing Roadmap: Facility for Whole Proteome Analysis

Topic	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment
LIMS and Workflow Management Participate in GTL cross-facility LIMS working group Develop technologies and methods to: • Manage massive dataflow • Process and integrate data • Manage workflow • Conduct QA/QC • Deploy collaborative tools for shared access to data and processes	Archival storage systems Prototype bulk data capture and retrieval systems Prototype inter- and intralab limited LIMS Shared LIMS and workflow technology for each analytical capability	LIMS for each analytical pipeline Data archives for each analytical pipeline Inter- and intralab LIMS	Establish LIMS for each analytical pipeline workflow Products: Output data products Cross-facility access and tracking Information management systems and automation Efficient, analytically rigorous pipelines Components and integration to GTL process
Bioinformatics Participate in GTL cross-facility working group for data representation and standards Provide user environments, community access, database development Integrate data-analysis methods Develop large-scale integrated experiment designs, analysis pipelines	Workflow processes and database needs Evaluation of technical solutions Large-scale storage and retrieval solutions Entire workflow processes and methods for experimentation and analysis Algorithms Quality control and assessment measures	Statistically designed experiments Multidimensional data-analysis and integration tools for large-scale experimentation Multilevel databases for bulk and derived data for each profiling method Analysis pipeline for derived data Community-access systems Cross-facility data-sharing processes and analysis methods Archival, computing, and network capacity to match demand	Bulk data archives for key data sets Process to link archives to production activities Local facility data archive Cross-facility data-sharing processes and analysis methods Mature bulk data archives, analysis piles Scaleup of archival activities, computing, and network capacity to match demand Products: • Whole proteome analysis for each GTL organism • Experiment templates and data sets for modeling and simulation • Defined experiment archive integrated with data and analysis from each analytical pipeline • Molecular profiling context-dependent database
Computing Infrastructure Participate in GTL cross-cutting working group for computing infrastructure Establish scientific computing with massive data reduction, archival storage application development Develop infrastructure: hardware, software, code control, libraries, environments Use ultrahigh-speed internet connection to GTL facilities	Operations process Computational architecture Large-scale data mining Access and security plans and processes Performance and quality metrics of service Capacity planning Backup and recovery strategy Testing plans Workflow Dev-Test-Pro strategy for implementations	Test network Development environment Validation methods Data archive Access methods Storage and retrieval methods Application integration and implementation Production infrastructure Cross-facility data sharing Infrastructure: hardware, software, and network	Production environment and data archive Bulk data archives for key data sets Process to link production activities to local facility data archive Cross-facility data-sharing processes and analysis methods Mature bulk data archives and analysis pipelines

5.4. Facility for Analysis and Modeling of Cellular Systems

5.4.1. Scientific and Technological Rationale 174
5.4.1.1. Probing Mixed Microbial Populations and Communities 175
5.4.1.2. Foundations for Community Analyses 177
5.4.2. Facility Description 178
5.4.2.1. Laboratories, Instrumentation, Quality Control, Computing, and Support 178
5.4.2.2. Performance and Production Targets 178
5.4.3. Technology Development for Cultivation of Microbial Communities 179
5.4.3.1. Requirement Examples 180
5.4.4. Development of Genomic Capabilities 181
5.4.5. Technology Development for Imaging and Spectroscopy 181
5.4.5.1. Analytical Characterization of Cellular Systems 181
5.4.5.1.1. Examples of Analytical Requirements 184
5.4.5.1.2. Monitoring and Interacting with Cellular Systems 185
5.4.5.1.3. Technology Development Progress and Benefits 185
5.4.5.2. Imaging Macromolecular Complexes 186
5.4.5.3. Development Options 187
5.4.6. Development of Computing Capabilities 187

Facility for Analysis and Modeling of Cellular Systems



Achieve an in silico, predictive understanding of microbes in their natural environments.

Cellular Systems

- Integrate knowledge and models to understand the structure and functions of cellular systems, from single cells to complex communities.
- Integrate imaging and other technologies to analyze molecular species from subcellular to ecosystem levels as they perform their functions.

The Facility for the Analysis and Modeling of Cellular Systems will be a user facility to provide scientists with insight into the responses and functionality of microbes and microbial communities in complex environments. Modeling and real-time functional mapping of processes from the molecular through the ecosystem levels will be used.

5.4.1. Scientific and Technological Rationale

The Facility for Analysis and Modeling of Cellular Systems will be the GTL capstone needed to provide the ultimate integration of analytical capabilities and knowledge synthesis critical for systems biology. Users of this facility will investigate how microbial communities and their subsystems of cells function together to sense, respond to, and modify their environment. They will accomplish this by dynamically identifying, localizing, and quantifying molecular machines and all other important biomolecules and their interactions as they carry out their critical roles throughout microbial and community life-cycles. This grand challenge for biology ultimately must be addressed before scientists can develop and test models to predict the behavior of microbes and take advantage of their functional capabilities.

This facility will provide the ultimate testing ground for fully integrated models developed from component models created from ongoing research and from previous facility data, modeling, and experimentation. The experimental capabilities of the facility will drive a new generation of systems models. Essential aspects of the computational challenges and conceptual roadmaps are described in 4.2.1. Theory, Modeling, and Simulation Coupled to Experimentation of Complex Biological Systems, p. 85; and roadmap tables beginning on p. 91 (see Fig. 1. Probing Microbial Communities, p. 176).

The other three GTL facilities will provide new high-throughput production and analysis capabilities to define and understand component parts and processes of microbial systems and analyze physiological and molecular conditions on a global level (i.e., measure the properties of samples comprising large numbers of cells). One of the key insights from recent research, however, is that microbial communities are dynamic and highly structured physically and functionally, suggesting that ensemble measurements that look at properties averaged across many cells can reveal only part of the picture.

To address this challenge, the Cellular Systems Facility will focus on dynamic systems-level studies, ranging from molecular processes within individual living cells to complex, structured microbial communities. Microorganisms in such communities—microbial mats and biofilms, for example—occupy various microniches established as a result of coupled biological, chemical, and physical interactions. Each member of the community carries out unique functions that can vary in space and time but are integral to community stability and overall function. The Cellular Systems Facility will provide the underlying capabilities to allow the spatial and temporal analysis of these complex microbial systems in a concerted and integrated way, from molecular processes to ecosystem functionality. This is a daunting challenge, partly because the complex multicellular drama is playing out at submicron scales. Nonetheless, we will need to dynamically image and functionally analyze the critical substructures and molecular species within microbes and their communities and develop models that describe and predict their behaviors. This capability builds on the Molecular Machines Facility, which will focus on intracomplex imaging to determine molecular makeup and structure and on intracellular imaging to localize machines within the cell. (See box, Cellular Systems Facility Objectives, this page; Fig. 1, p. 176; sidebar, Group Living and Communicating, p. 18; Fig. 1. DOE Genomics:GTL High-Throughput User Facilities, p. 103; and Fig. 3. GTL Facilities: Core Functions and Key Interactions, p. 108, from 5.0. Facilities Overview).

Facility Objectives

- **Relate community composition, structures, and functions to environmental physicochemical conditions measured at the scale of microbial communities—an overlay of community physical and functional maps.**
- **Determine community composition, relative positioning of members, and phenotypes.**
- **Analyze overall community functionality and distributions and fluxes of molecular species.**
- **Dynamically image critical molecules and substructures as they function intra- and intercellularly.**
- **Develop models of microbial function at the molecular, cellular, community, and ecosystems levels.**
- **Provide protocols, data, models, and tools to the community.**

The analytical and conceptual challenges of this ultimate step in systems microbiology will require unprecedented technical and computational resources and infrastructure far beyond the reach of individual investigators.

5.4.1.1. Probing Mixed Microbial Populations and Communities

In microbial communities, the complex and dynamic set of interactions that we seek to analyze is taking place in an area far smaller than the period at the end of this sentence. To understand community function, we first must be able to analyze environmental and community structure and composition at high resolution (Fig. 1, p. 176). On an even-smaller scale (roughly 1/500th the size of a period), we must be able to peer inside an individual microbe in a nondestructive way to locate and continually track essential biomolecules that reveal the cell's inner workings; these biomolecules include DNA, RNAs, proteins, protein complexes, lipids, carbohydrates, and metabolites (Riesenfeld, Schloss, and Handelsman 2004; Johnston et al. 2004; Schaechter, Kolter, and Buckley 2004). Models then must be developed to describe key features of these biological interactions within the physicochemical environment and predict how the system will evolve structurally and functionally. Robust models are required to conceptualize these intricate systems, formulate meaningful hypotheses, manage the ensuing complexity and sheer volume of information that comes from experiments, and, ultimately, allow incorporation of the resulting science into applications.

Key information we seek about microbial-community function includes:

- Community arrangement and its physical environment
- Members and where they reside in relation to each other and their environment

- Phylogenetic, phenotypic, and functional properties of members. Do the genes in one organism regulate gene expression in another organism?
- Microbial interactions among themselves and with their environment
  - Information flow (e.g., genes, signaling molecules) within the community
  - Ecological interactions (e.g., predation, symbioses) among the various members
  - Food webs and communications
  - Excretions, secretions, and consumptions
  - Interactions of secreted components and nutrients with the environment
- Energy and element flow through communities and cells and its regulation
- Intrinsic biological (genetic) and environmental factors that control the structure, stability, and functioning of communities

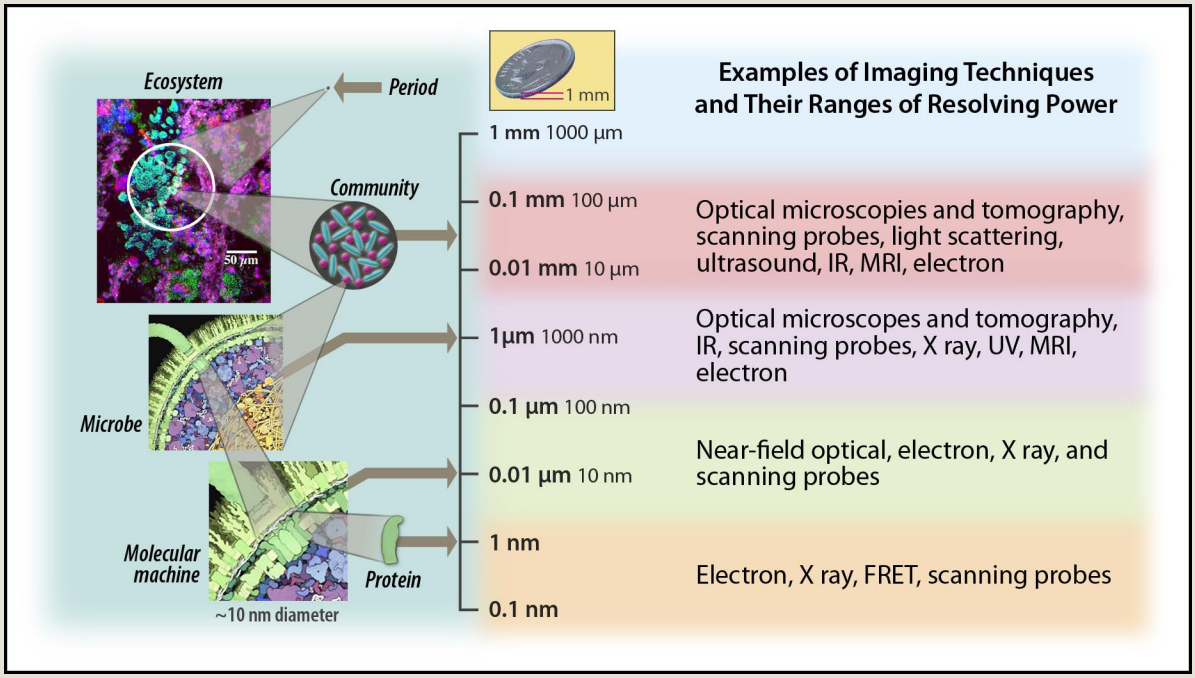


Fig. 1. Probing Microbial Communities. Microbial communities and ecosystems must be probed at the environmental, community, cellular, subcellular, and molecular levels. The environmental structure of a community will be examined to define members and their locations, community dynamics, and structure-function links. Cells will be explored to detect and track both extra- and intercellular states and to determine the dynamics of molecules involved in intercellular communications. Probing must be done at the subcellular level to detect, localize, and track individual molecules. Preferably, measurements will be made in living systems over extended time scales and at the highest resolution. A number of techniques are emerging to address these demanding requirements; a brief listing is on the right side of the figure. These and other techniques are discussed in section 5.4.5, beginning on 181.

- Trajectory of community evolution
  - Life strategies of each population in the community
  - Role of lateral gene-transfer processes in microbial evolution and community metagenome
  - Senescence, death, and turnover rate (consequence of death?)
  - Community resilience (biodiversity, stability)

5.4.1.2. Foundations for Community Analyses

Before undertaking these analyses, we will have a growing body of knowledge (incorporated in the GTL Knowledgebase) and capabilities from work funded by other agencies and within the GTL program and facilities. These resources will include:

- Cooperative analyses of comprehensively annotated genomes of individual microbes and the community metagenome to estimate the genetic potential of individuals and the community.
- Many critical proteins encoded in the community's genome expressed and characterized to produce a substantial body of functional characterization data incorporated into gene-annotation data sets. These data will provide significant insights into "interesting" processes that need to be pinpointed and analyzed in the context of a complete system. Since studies can be performed on the basis of sequence alone, we will have circumvented the fact that these microbes are largely unculturable.
- Ability to produce multiple affinity reagents for any produced proteins and other such biomolecules as RNA and some metabolites that can be used to locate, track, and manipulate these entities within living systems. Fusion tags can be incorporated in a variety of ways.
- Extensive measurements at the global level on bulk and ensemble samples to ascertain the phylogenetic and physiological state of member microbes or the entire community under relevant conditions. Temporal relationships will be revealed through repeated sampling and process interplay via extensive linked measurements of the transcriptome, proteome, metabolome, and other biological and physicochemical variables.
- Analysis of the structure and function of critical molecular complexes in vitro and insights to determine where and in what context they carry out their cellular functions.
- Extensive databases and exploratory tools to begin deriving underlying principles at the molecular, cellular, and community levels and the ability to begin encompassing complexities in detail as processes play out in real, nonlinear, coupled systems.
- Extensive models at molecular, cellular, and community levels to support creating and simulating hypotheses in a systems context. These models and simulations will be used to design and gain insight into experimental campaigns and protocols and provide advanced knowledge of key experimental variables that must be captured in ensuing research.

Even when all this information is at hand, unraveling how all these entities and processes act together in a continuous, concerted way—from molecular to community levels—will remain a grand challenge in accomplishing DOE mission goals. All technologies that created this body of knowledge must be specialized in innovative ways to provide the same information at a microscopic (actually nanoscopic) scale. This facility will be capable of analytical measurements that are nondestructive and done in real time in living cells within a well-defined global and dynamic system. Understanding how these individual cells interact and function as a unit—a microbial tissue in some respects—to carry out complex processes is key to unlocking their vast potential for important applications and achieving our science goals.

5.4.2. Facility Description

The Cellular Systems Facility will combine advanced computational, analytical, and experimental capabilities for integrated analysis of spatial and temporal variations in biological systems—how, when, and why the various system components appear, disappear, function, and remain. The facility will determine the state of cellular systems, from the internal makeup, structure, and dynamics of individual microbial cells to complex communities and their environments. To achieve a systems-level understanding, simulation and modeling must be coupled tightly with experiments to define and analyze the complex regulatory and metabolic processes in microbial cells and communities. This facility will emphasize concurrent and dynamic measurements of proteins, molecular complexes, intracellular metabolites, regulatory molecules, and gene transcripts. The aim is to establish the state of cells within populations and communities as a function of changes in physicochemical and biological conditions, emphasizing measurements at spatial and temporal resolutions appropriate for the entities being measured.

5.4.2.1. Laboratories, Instrumentation, Quality Control, Computing, and Support

The facility will be 100,000 to 150,000 gross square feet, with laboratories containing necessary cultivation, isolation, and analysis instrumentation. Its frontier instruments will incorporate capabilities and techniques for extensive environmental control and monitoring, manipulating communities in real time in various ways, and temporal and spatial resolution. The exact configuration of these instruments awaits necessary technical developments as described below. The facility will have requisite offices, common space, and conference facilities for staff, administration, and users.

The Cellular Systems Facility will provide to the user community:

- Frontier instruments incorporating the capabilities to create, sustain, and monitor structured microbial communities and analyze them at the molecular through ecosystem levels.
- Models and the tools and computing and information infrastructure for developing and evaluating such models, along with the user-facility infrastructure needed to undertake such tasks.
- The GTL Knowledgebase as a source of data on all known aspects of microbial systems that have been studied as a foundation for further experimentation, model development, and the ensuing simulations.

The facility therefore will be highly data intensive, providing extensive linked data sets on the dynamic behavior of microbial cells and communities. It also will be compute intensive, providing unprecedented data analysis, modeling, and simulation. In addition, it will involve new computational approaches for data storage, analysis, and use in complex models. These systems-level data sets will be made available to the entire scientific community. This new knowledge will be invaluable for advancing the annotation of microbial and community genome sequence, identifying regulatory and metabolic networks in microbial systems, and understanding microbial contributions to ecosystems function.

5.4.2.2. Performance and Production Targets

The Cellular Systems Facility will have the experimental and computational capacities to measure and analyze hundreds of microbial systems per year at unprecedented levels of detail. A key distinguishing feature will be the sophistication and performance of its instrumentation and capabilities in computational modeling and simulation. Some key performance features as described in the following sections include:

- Performance of physical (i.e., structure) as well as functional (i.e., molecular profiling) mapping
- Spatial and temporal resolution to map all levels of the functional processes within a microbial community (i.e., nanometers to millimeters and microseconds to hours)
- Performance of nondestructive measurements
- Culturing capabilities for maintenance of realistic environmental conditions and microbial communities

- Integration of modeling capabilities and supporting computing infrastructure, including data-intensive bioinformatics, compute-intensive molecular modeling, and complexity-dominated systems modeling

This GTL facility, more than the others, must overcome major challenges associated with the lack of available technologies and instruments for measuring the dynamic state of living microbial cells. As stated above, it will benefit from technologies, instrumentation, and data developed by current and future GTL R&D and pilot projects by the time the facility comes online. Along with state-of-the-art capabilities when operations begin, the facility also will include extensive ongoing development of essential instruments and technologies to advance the measurement of activities and characteristics of cellular systems at the single-cell level.

DOE has an extensive and successful history of developing and applying new technologies to complex problems in the physical and chemical sciences. As in the genome projects, the agency can draw on multidisciplinary teams of biological, physical, computational, and other scientists and engineers from national laboratories, academia, and industry. GTL brings a tremendous opportunity for using these same talents in the Cellular Systems Facility to devise technological solutions to some of biology's most complex problems.

5.4.3. Technology Development for Cultivation of Microbial Communities

The classic definition of an unculturable microbe is that it cannot be grown in homogeneous suspension. Because of this, there is a dearth of information about the metabolic capacity of microorganisms that resist cultivation under laboratory conditions (Keller and Zengler 2004). This is due primarily to the difficulty or impossibility of simulating the chemistry and interspecies interactions of highly structured communities by suspended cell-culture techniques. Most microbes reside in these structured communities and display unique phenotypic states in response to microenvironments within those communities. Many current technologies require large populations of cells to measure gene expression, proteome, and metabolites, masking the true cell-to-cell heterogeneity. The inability to cultivate most structured communities formed by natural microbial populations limits our discovery of new genes, gene products, and resultant functionalities. New cultivation techniques must support the development of meaningful community structures, and the functions of community members must be measurable in that environment (see sidebar, Laboratory Cultivation Techniques to Simulate Natural Community Structure, p. 180).

The facility's cultivation systems will allow for the precise control and manipulation of environmental conditions and for the monitoring and culturing of microorganisms in meaningful community structures. Many microorganisms of scientific and biotechnological importance, including those sequenced by the Biological and Environmental Research Program, by GTL, and by the Joint Genome Institute's Community Sequencing Program, are relevant to various DOE missions. Some thrive in unusual or extreme environments or those in which gradients or temporal changes occur in physicochemical conditions.

Scientific investigations of these organisms and the communities in which they live thus require flexible, highly controlled, and instrumented systems that can provide a range of environmental conditions and sophisticated measurements. These conditions include monoculture or mixed cultures; nutrient status; extremes of pH, temperature, and salinity; exposure to contaminants and radiation; gas composition and pressure; light intensity; and the presence of solid phases. The ability to control and monitor the environment allows for rigorous investigations and interpretations of gene expression, regulation, and function at the level of individual cells, cell populations, and mixed communities. When required, cells of unusual or difficult-to-culture microorganisms will be cultured in sufficient quantities to provide protein for biophysical, structural, and functional analyses. In other cases, very small numbers of unusual or difficult-to-culture microorganisms might be studied using novel microscale approaches combined with specialized sensitive analyses of gene expression, proteomics, metabolism, and metabolite flux. Ultimately, for meaningful analysis of communities, measurements at the individual cell level will be essential. To select for study any cell in such a structured community, we need to be able to (1) remove it from its environment without inducing significant changes in the properties being measured or (2) conduct analyses of living cells in situ (i.e., without disrupting its environment or harming the cell).

Laboratory Cultivation Techniques to Simulate Natural Community Structure

Microbes associating within a biofilm surface offer the opportunity for members of a discrete population and individual organisms representing different species to establish fixed spatial relationships over extended periods of time. Surfaces enable microorganisms to establish high cell densities in localized areas. For example, products of cell metabolism in a colony of one type of microorganism diffuse to adjacent surface areas, forming strong concentration gradients within the intercellular volume of a biofilm (Beyenal, Davis, and Lewandowski 2004; Beyenal et al. 2004).

To identify the function of genes preferentially expressed by specific populations in the structured community, new cultivation techniques are being developed that incorporate surfaces for microbial colonization and RNA extraction (Finelli et al. 2003). During the past decade, researchers have developed reactors in which biofilms can be imaged using confocal scanning laser microscopy (CSLM) and other light-microscopic techniques (Wolfaardt et al. 1994). When combined with fluorescent in situ hybridization (FISH) to distinguish populations of cells in multipopulation biofilms and fluorescent reporters (green fluorescent protein) of functional gene expression, CSLM has been used to demonstrate how gene expression by one population affects gene expression in another proximally located population (Moller et al. 1998).

Pacific Northwest National Laboratory

The mobile pilot-plant fermentor shown here has a 90-L capacity and currently is used to generate large volumes of cells and cell products such as outer-membrane vesicles under highly controlled conditions. This fermentor allows the end user precise control of culture growth to produce high-quality samples. Future generations of fermentors will be more highly instrumented, possessing sophisticated imaging and other analytical devices developed to analyze interactions among cells in biofilms under an array of conditions.

5.4.3.1. Requirement Examples

- Development of techniques to isolate single cells from a natural community and analyze them for the expression of targeted genes, proteins, and other products (e.g., using fluorescent tags produced from metagenome sequences).
- Evaluation of metabolic processes carried out by cells in structured microbial communities, using molecular tags for RNA, proteins, or other reported moieties to map individual cell populations within the community (see imaging discussion below).
- For unsequenced microbes, application of techniques such as intact biofilm polymerase chain reaction (IB-PCR) to construct 16S rRNA clone libraries once a community is formed. The rRNA sequences of each population present could be used to establish phylogenetic links to other known populations (Reardon et al. in press). Techniques such as FISH then could employ 16S rRNA sequence data to construct oligonucleotide probes to locate different populations within the biofilm and identify putative associations between colocalized populations.
- For many systems, capabilities to assay the distribution of properties among a population of extracted cells from structured communities, sediments, or soils. Providing invaluable information, flow cytometry will be a high-throughput method of choice. Novel cultivation approaches also might be combined with single-cell sorting techniques such as emerging microfluidic lab-on-a-chip devices to grow and study currently uncultivable members of microbial communities.

5.4.4. Development of Genomic Capabilities

To test hypotheses about function, genetic manipulation to generate mutants and specific constructs containing tags or reporter molecules is an essential requirement for systems biology research. Highly roboticized capabilities will be essential for high-throughput construction and screening at the genome level. Examples of required basic capabilities include nucleic acid isolation and analysis, sequencing and annotation, expression analysis, gene cloning and expression, fusion tagging of genes (Gaietta et al. 2002), general tools to manipulate members genetically, and cell sorting. Many of these capabilities will be present in other GTL facilities, the JGI, and in researchers' laboratories and may be incorporated into this facility as needed. Molecular microbiology and, in fact, all microbiology support capabilities will require a highly developed system for information management and integration.

5.4.5. Technology Development for Imaging and Spectroscopy

The Cellular Systems Facility will employ a broad range of imaging modalities to monitor the structure of microbial communities and image (spatially and temporally resolving) the many molecular species critical to community function. The imaging capabilities of the Cellular Systems Facility and the Molecular Machines Facility are complementary—Molecular Machines images intracomplex structure and cellular location, and Cellular Systems focuses on spatially and temporally mapping multiple processes through the lifecycle of a community of cells in a complex environment. Imaging modalities available for both facilities are presented in Table 1. Characteristics of Available Imaging Modalities, p. 182; Table 2. Attributes of Available Techniques for Cellular Systems Characterization, p. 183; and Fig. 1, p. 176, all of which list the primary probe methods available, techniques based on them, analytical characteristics, prospects for further development, and computing requirements. To be applicable, these methods must be chemically specific, perform measurements nondestructively, and be capable of functioning in concert with other techniques (Toner et al. 2005).

5.4.5.1. Analytical Characterization of Cellular Systems

Critical to addressing key scientific questions will be the novel application of existing and emerging technologies that characterize systems in a continuous and spatially resolved way. Analyses now conducted on bulk samples must transition to nondestructive processes capable of characterizing systems ranging from a microbial community through multiple processes within a single living cell. Also, the power to view multiple systems with high spatial and temporal resolution must be augmented with the ability to identify, track, and manipulate living microbes in the presence of other strongly interacting species. To achieve this, many classic imaging techniques must be coupled with methods that can detect specific molecules or processes. Physical, chemical, and biological variables must be identified and tracked. Furthermore, the power to observe systems in action will need to be enhanced by the ability to interact with these systems.

Developing the capability to view biological systems in great detail will enable new high-throughput approaches to studying cellular systems. Each cell in a culture, consortium, or community presents a unique reflection of the biological response to the overall system's changing state. Each provides a set of multilevel outputs in response to the effects of changing parameters (e.g., environmental insults, nutrient gradients, and temperature). To the extent that this parallel data stream can be captured in real time, biological experiments can be conducted in a high-throughput manner rather than running as several series of experiments to evaluate each possible response (e.g., cell division, movement, and protein shedding) for each type of environmental change.

To enable these advances, new technologies must address the special requirements for observing biological systems. Ideally, techniques will be nondestructive, noninterfering, and compatible with the analysis of heterogeneous, living systems. They will need to document the state of each cell (or many cells or cell types) as time and environmental conditions change. Furthermore, physical and chemical information must be mapped onto community structure while detailing changes at the molecular scale. These analyses will necessitate the development of new software to provide intelligent processing of data. Ultimately, such tools will

Table 1. Characteristics of Available Imaging Modalities

	Technique	Unique Characteristics	Future Prospects	Bioinformatics
Visible Light (possible: 50 nm practical: 300 nm)	TIR Absorbance Scattering NLO Adaptive optics FRET Structured light illumination	Noninvasive In situ Wide range of time + length scales Functional analysis Coordinated release of caged molecules Microsurgery, microablation Characterization of individual cells and communities (biofilms)	Better probes, lanthanite dyes, quantum dots, nanoparticles, tetracysteine tags, genetically encoded nanoparticle sensors More versatile excitation sources Better detectors	3D visualization (online, offsite) Pattern recognition (spatial, spectral) Multiscale, multimethod data fusion
X Ray (20 nm)	Tomography Spectroscopy Microprobes	Thick, hydrated samples Whole cells Clean spectrum Organic functional group metal redox spectroscopy Molecular localization in ultrastructural context Characterization interactions	More versatile excitation sources Better detectors	
EM (0.3 nm)	Tomography Molecular microscopy: Single particle Cryo	Whole cells or sections High-resolution molecular localizations in ultrastructural context Correlation with fluorescence	More versatile excitation sources Better detectors	
Force Imaging	AFM tapping	Cell wall imaging Imaging of protein, nucleic-acid components	Better tips (higher-aspect ratio: Carbon nanotubes)	
Force (manipulation, perturbation)	Optical tweezers Magnetic tweezers	Mechanical characteristics (cell wall) Thermodynamics and kinetics of transient interactions Characterization of the molecular-machine mechanochemistry Correlated mechanical properties	Combined single-molecule fluorescence, optical tweezers	

help elucidate the large-scale biochemical organization that characterizes community structure. Such new analytical approaches will be essential for assessing the community's physiological and phylogenetic makeup and for testing predictions derived from theoretical models.

A number of these scientific needs will require fundamental new developments in imaging technology—a transformational goal for GTL biology. Revolutionary advances will be essential for determining the dynamics of communities and their functions under various environmental conditions, defining the physical structure of cells and communities, detecting and tracking extracellular and intercellular molecules to define cell states, and, ultimately, understanding how molecular events are communicated in space and time.

Table 2. Attributes of Available Techniques for Cellular Systems Characterization

Scale of Analysis	Information Needed	Technoques for Structure and Imaging	Static Characterization Techniques	Dynamic Characterization Techniques
Proteins	Components Abundance Structure	X-ray crystallography (angstrom) Raman spectroscopy (angstrom) Neutron crystallography (angstrom) X-ray spectroscopy (sub-angstrom) Electron microscopy (SEM, TEM, STEM, tomography) Electron crystallography	Infrared spectroscopy Raman spectroscopy NMR (nuclear magnetic resonance) spectroscopy Microsampling Microfluidics Fluorescence Scattering	Infrared spectroscopy Raman spectroscopy NMR spectroscopy Microsampling Microfluidics Fluorescence Scattering
Molecular Machines	Components, active sites Function, role, interchangeability, stressed behavior	X-ray crystallography, Raman spectroscopy, neutron scattering, X-ray scattering, EM, multi- and hyperspectral fluorescence	Infrared spectroscopy Raman spectroscopy NMR spectroscopy Microsampling Sensors	Pump-probe spectros. Microsampling Sensors Labels (quantum dots, organic fluorescence) Laue X-ray crystallography
Cellular	Components Active sites Function role Interchangeability Communication Stressed behavior	X-ray microscopy Scanning probes Scanning probe microscopy (SPM) Atomic force microscopy (1.0 nm) Scanning near-field optical microscopy (NSOM or SNOM) Scanning tunneling microscopy Chemical force microscopy Electrostatic force microscopy Magnetic force microscopy Electron microscopy (SEM, TEM, STEM, tomography) Far-field vibrational imaging (>10 microns) Optical microscopy	Mass spectrometry NMR spectroscopy Probes Raman spectroscopy Neutron spectroscopy Infrared spectroscopy SPM PH meter Microsampling Sensors Multi- and hyperspectral fluorescence Optical microscopy (one or multiphoton, scanning optical tomography; 200 nm in conventional mode; 5 nm in FRET/FLIM modes, FISH, CARS, SHM)	Raman spectroscopy X-ray microscopy Scanning probes Mass spectroscopy NMR spectroscopy Probe spectroscopy Infrared spectroscopy SPM Microsampling Sensors Fluorescence Labels
Communities In lab In field	Components, active site, function, role, activators, interchangeability, stressed behavior How to communicate?	Far-field vibrational imaging (>10 microns) Optical microscopy (one or multiphoton, scanning optical tomography) NMR imaging Light-scattering spectroscopy Ultrasound	Infrared spectroscopy Raman spectoscopy NMR spectroscopy Microsampling Sensors	Pump-probe spectroscopy Microsampling Stop-flow chromotography Sensors Labels

5.4.5.1.1. Examples of Analytical Requirements

Intracellular Structure. Intracellular protein, RNA, and metabolite localization and kinetics of localization.

- Proteomics on replicate communities.
- Fine-scale cell ultrastructure.
- New multimodal capabilities for dynamic imaging of targeted intracellular molecules and their interactions (including machines) in individual cells and cell assemblies [e.g., with antibody labeling and electron microscopy (EM)].

Community Structure. Analytical instrumentation and techniques for determining overall community structure and identifying and characterizing spatial and temporal variations in metabolites, signaling and regulatory molecules, and the physicochemical environment within communities (see discussions under the Molecular Machines Facility, beginning on p. 143).

- Probes for the in situ measurement of extracellular metabolites in real time.
- Imaging and spectroscopy of population structure, gene expression, and metabolites in cell aggregates and subpopulations within communities.
- Characterization of cells in mixed communities by multispectral imaging of key cellular chromophores, possibly moving to on-the-fly cell-sorting platforms.
- Measurement of elemental distribution, oxidation state of elements, and biomolecules within and among communities.
- Quantitative imaging of metabolite (and signaling molecule) flux between cells in close proximity to or in contact with each other—one of the most critical needs for understanding how microbial communities function.
  - Analysis and assessment of the makeup and role of extracellular polymers in community structure, function, and stability.
  - Detection and frequency of genetic exchange, recombination, and evolution within communities.
  - Determination of macroscopic transport of water, solutes, and macromolecules and their relationship to microbial function.
- Characterization of interface physical and chemical properties.

Identified Development Needs

- Advanced chemical and biological probes, including engineered microorganisms, tagged biomolecules, and chemical sentinels that will help characterize microbial communities.
- Advanced tools for imaging characterization for use in the laboratory as well as in the field.

A variety of imaging and microspectroscopic techniques are emerging to meet these challenges. In general, imaging relates spatially dependent information. Characterization of additional dimensions, however, will be essential for relating system activity. Some commonly used imaging techniques include:

- **Short-Wavelength Techniques.** Analyses with electrons and X rays typically provide the highest spatial resolution. Although commonly associated with ultrastructural analyses, short-wavelength techniques are being extended for analyses of whole cells at atmospheric pressures. Additionally, X rays are useful for mapping trace metals, while spectroscopic measurements can provide chemical identification.
- **Optical Microscopies.** The current standard for live-cell imaging, these tools are ideal for studying dynamics across a broad range of time scales and are sensitive down to the single-molecule level. A number of physical scales can be assessed, and emerging techniques and new labels are improving the sensitivity and resolution of optical microscopy.

- **Long-Wavelength Techniques.** A variety of these procedures including vibration, magnetic resonance, and terahertz-based imaging can provide essential information on chemical structure, identity, and spatial arrangement. For example, vibrational signatures are molecularly specific and can produce direct chemical information without additional labels.
- **Other Techniques.** A broad range of unconventional imaging approaches are making an impact on biological studies. Most notable, the family of instruments comprising scanning probe microscopy enables molecular-scale resolution; and chemical, electrical, and physical properties can be measured simultaneously. Emerging tools based on optical and magnetic trapping are allowing measurement of mechanical properties while micro- and nanoscale structures permit sensing of chemical, physical, and biological attributes.

Clearly, many current imaging and microspectroscopic techniques possess significant attributes and provide information relevant to the study of biological systems. Significant advances still are needed to adapt many of these tools to the characterization of microbial cellular systems much smaller than eukaryotic cells. Advanced instrumentation, improved biocompatibility, new approaches for targeting and delivery of tags, and improved labels are but a few of the significant challenges that face imaging technologies. More significant, no technique alone can provide the broad range of information needed to understand community structure and system function. A combination of methods will be essential to extend the depth of information required.

5.4.5.1.2. Monitoring and Interacting with Cellular Systems

To enable effective systems-level studies, the ability to monitor systems in action must be enhanced with selective construction, manipulation, and interaction with the system. Only then can efficient experimental evaluations and effective iterations be achieved with pursuits in theory, modeling, and simulation. This integration will be a culminating product of the facility and an essential tool for studying microbes, consortia, and microbial communities.

Advanced cultivation systems that allow for precise control, manipulation, and monitoring of environmental conditions must be compatible with advanced imaging technologies. Chemical gradients will need to be controlled and monitored precisely while temporally measuring molecular-scale properties. Genetically defined organisms must be carefully arranged into ordered microbial communities, perhaps through molecular-scale patterning techniques resulting from nanotechnologies. Such highly defined systems will require integration with sensing capabilities and the ability to activate biomolecular networks remotely. The capacity for simultaneously imaging and specifically targeting reagent release or activation, as currently used in biomedical applications, is within reach for GTL systems biology studies. The creation of such compound, multifunctional instruments will enable the collection of information needed to understand and exploit complex biological systems.

5.4.5.1.3. Technology Development Progress and Benefits

5.4.5.1.3.1. Advanced Optical Methods – Laser or Synchrotron Based

- Optical spectroscopic methods can be used as tools for noninvasive characterization and monitoring of dynamic behavior.
- Measurements of absorption and in vivo fluorescence can be used to monitor the presence and relative concentration of optically active biochemical species.
- Light-scattering spectroscopy can probe the size distribution of community structures.
- Vibrational (infrared and Raman) spectroscopy is a technique for studying the composition of biological materials without perturbing or labeling the sample. Biological components (e.g., lipids, proteins, nucleic acids, and carbohydrates) and biofilm and microbial surfaces (e.g., minerals and polymers) have unique vibrational spectra based on their chemical structures.

- Use of these methods will provide new information on the following:
  - Large-scale (1- to 10-micron) biochemical organization.
  - Composition and distribution of extracellular polymer matrices.
  - Concentration and distribution of nutrients, metabolites, signaling molecules, and other macromolecules.
  - Interactions of biofilms and microbial communities with supporting surfaces.

Because vibrational spectromicroscopy is noninvasive, it can be performed on dynamic living systems in combination with other techniques. If synchrotron radiation is used as the photon source, a dynamic system can be studied directly on surfaces of geological materials (see Fig. 4.1, p. 27, Report on Imaging Workshop 2002).

Significant progress already has been made using confocal and two-photon fluorescence microscopy. The specificity of these techniques is provided by the exogenous chromophore targeted through an affinity reagent or fusion tag to a particular protein. The resolution is on the order of a micron and slightly higher for two-photon than for confocal microscopy. Delivering chromophores to remote regions within a community or cell is a particular challenge. Additionally, the identification of probes that maintain activity in diverse environments is required (see Fig. 4.2, p. 28, Report on Imaging Workshop 2002).

All these techniques can be used in an imaging arrangement to monitor changes in community behavior in real time. Improvements are needed in such areas as spatial resolution, the ability to provide quantitative information, and data-acquisition speed. Additionally, advanced light-microscopy techniques can be developed for high-resolution 2D and 3D mapping. Often with specificity to particular components associated with imaging, these techniques include surface-plasmon resonance, surface-enhanced Raman spectroscopy, imaging of second-harmonic generation, optical-coherence tomography, and coherent anti-Stokes Raman scattering.

5.4.5.2. Imaging Macromolecular Complexes

Many types of imaging technologies can be employed to identify and spatially and temporally localize macromolecular complexes and their interactions within a dynamic community environment. Some specialized techniques have specific applications to the analysis of macromolecular complexes in situ in live, fixed, or frozen cells or ex situ. The strengths of imaging techniques typically include detection sensitivity and the ability to identify complexes in cells. Imaging techniques are applicable to all classes of complexes. In many cases, however, the identities of one or more components of the complex must be known to prepare tagged probes for imaging analysis. This requirement limits the application of imaging to full identification of protein complexes. Currently, most imaging techniques are relatively slow; automation, however, is providing faster sample throughput, and improved computational tools are enhancing data acquisition and analysis. Imaging techniques relevant to identification and characterization of protein complexes are summarized below, with additional information on other imaging tools in Table 2, p. 183.

Tagged Localization. Used with visible, X-ray, or electron microscopies to identify sets of biomolecules labeled with tags. An in situ method applicable to live (visible light), fixed, or frozen cells, it also is applicable to tagged transient complexes and membrane-associated complexes. A limitation is that the complex must be labeled with a tag, requiring tag synthesis and introduction into cells. Spatial resolution in these modalities comes from the instrument response function of the exciting source (i.e., the exciting beam provides the resolution). More developed X-ray optics, more versatile excitation sources, and improved probes are needed. Lanthanide dyes, quantum dots, nanoparticles, tetracysteine-based ligands, and other probes are examples of some recently reported probes used with various imaging modalities.

Fluorescence Resonance Energy Transfer (FRET). Used to identify pairs of biomolecules labeled with tags as well as to provide information on biomolecule relationships. This in situ method is applicable to live cells, tagged transient species, and membrane-associated complexes. FRET is particularly good for structure and binding of extracellular ligands. Like other imaging techniques, it requires tag synthesis and introduction into cells.

Scanning Probe Microscopy. Identifies protein associations by scanning with a specific molecule attached to the tip, including transient molecules. The technique is capable of very high spatial resolution, depending on the length of probe time, and of single-molecule detection. It is most suited for the study of membrane-associated complexes with whole cells or for the study of isolated complexes. Like other imaging techniques, it requires that the identity of one component of the complex be known so a molecule can be attached to the tip as the probe molecule. The probe, for example, then can be used to identify interaction sites on a cell surface. The technique is labor intensive and slow. Identification is a one-at-a-time process unless multiprobe devices with individual probe molecules are employed. These multiprobe devices are under development to allow technique application in a highly parallel fashion. Computer modeling of protein folds would enhance data interpretation, and improved computation is needed for data visualization and manipulation.

5.4.5.3. Development Options

As previously mentioned, many techniques required for the facility have yet to be developed sufficiently to analyze microbes of less than one micron in complex and changing communities. Many potential options must be explored over the next few years to determine probe and detection modalities capable of providing necessary information under these demanding conditions. Options that may be explored regarding available techniques, their range of applicability, and information they might provide are shown in Table 2, p. 183, and Table 3. Cellular Systems Facility Technology Development Roadmap, p. 188. The bulk of intracomplex characterization of molecular machines will be carried out in the Molecular Machines Facility. The sidebar, The Super Imager, this page, details creation of super imagers comprising compound, multifunctional instruments that individually would include many of the capabilities listed. Many of these development issues are summarized in 6.0. Development Summary: Global, Crosscutting, and Long-Lead Issues, p. 191.

The Super Imager

The potential is to create compound, multifunctional instruments that individually include many of the following capabilities:

- **Mapping of molecular species such as RNA, proteins, machines, and metabolites through the use of fluorescent tags of various kinds**
- **Multiple excitation and detection wavelengths including both fluorescent and infrared absorption methods**
- **High-speed 3D imaging**
- **Nonlinear contrast imaging including second- and third-harmonic generation and coherent Raman scattering**
- **Lifetime mapping as sensitive probes of local environments**
- **Rotational correlation mapping for in situ analysis of protein structure and function**
- **Magnetic resonance imaging with 10-micron–scale analyses of metabolite concentrations and providing data on diffusion properties and local temperatures**
- **Acoustical imaging of the system's physical parameters with micron-scale resolution**
- **Atomic force microscopy (AFM) mapping of structures with added information provided by the controlled-interaction light with sharp metallic AFM tips to obtain optical resolutions of ~20 nm, one-tenth the diffraction limit**
- **High spatial resolution (nanometer scale) using X-ray and electron microscopies, including the use of special DOE facilities or perhaps the development of laboratory-based X-ray sources for imaging**

[Source: *Report on the Imaging Workshop for the Genomes to Life Program April 16–18, 2002* (Office of Science, U.S. Department of Energy, Nov. 2002); www.doegenomestolife.org/technology/imaging/workshop2002/]

5.4.6. Development of Computing Capabilities

Computational tools and infrastructure are required for efficiently collecting, analyzing, visualizing, and integrating large data sets to elucidate gene function and to model and simulate regulatory and metabolic networks, cells, communities, and ecosystems). These tools will support the development and validation of

Table 3a. Cellular Systems Facility Technology Development Roadmap

Technology Objectives	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
Technologies for Cultivation of Microbial Communities Precise control, manipulation, and monitoring of environmental conditions; interrogation Functional individual microbial cells in the context of characterized physiochemical environment Support for formation of structured communities	Requirements defined for analyzing individual cells within structured communities: Mixed microbial cultures • Suspended and structured • Biofilms Methods and approach to identify and track microbes and molecular complexes • Tagged probes » Increased variety of signals » Signal interpretation » Incorporation in cell • Arrays	Multiple flexible experimental systems to control and manipulate growth and conditions with multiplex measurements of activity including: • Chemostats • Microtechnologies • Remote sensing • Imaging • Surfaces to nucleate biofilms and other structures Multiple probes to identify community members Temporal monitoring of community structure and function	Integration of culturing capabilities within multiprobe instrumentation for simultaneous control, manipulation, and multimodal analyses of structured communities: • Nondestructive • Real time • Linked databases • Environmental, community, cellular, and molecular levels	Integrated, highly characterized, and real-time manipulatable structured microbial communities that simulate natural communities and niches: • Protocols • Extracted samples • Characterizations • Analytical images
Environmental Communities Sampling In situ measurements	Lab techniques extended to field use	Planned extension after operations begin		
High-Throughput Cultivation for Single-Cell Analysis Sampling techniques Controlled viable growth of single cells	Analysis from within structured communities, in microculture extracts, or in place: • Cell sorters • Lab on a chip and microfluidics • Single-cell analysis of "unculturable" environmental samples	Assessment of compatibility with analytical instrumentation and simulation fidelity of natural environments	High-throughput operational mode combining culturing techniques interfaced with multimodal, analytical, and manipulation modalities	Single cells prepared in conditions that simulate microniche environments in highly structured microbial communities such as biofilms (formerly unculturable)

(continued next page)

theories and models of community growth, function, and environmental response. New theory, algorithms, and implementation on high-performance computer architectures also are needed for modeling and simulating cellular systems. Enabling a broad range of biologists to access the large data sets and computational resources for discovery-based biology will require the development of web- and grid-based technologies (see 4.0. Creating an Integrated Computational Environment for Biology, p. 81, and Table 4. Computing Roadmap, p. 190).

Table 3b. Cellular Systems Facility Technology Development Roadmap

Technology Objectives	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
Temporal and Spatial Localization of RNAs, Machines, and Metabolites Analytical measurement contexts: • Environmental • Community • Intercellular • Intracellular	Requirements defined for multimodal measurements: • Environmental physicochemical variables • Intercellular biomolecules • Intracellular biomolecules • Metabolites • Community overall biochemical and biophysical functionality Examples of needed instrumentation with molecular specificity, sensitivity, and spatial and temporal resolution: • NMR for community-scale microscopy (e.g., metabolites, signaling molecules) • Small molecules in living cells • Gene expression in living cells • Proteins and machines in living cells, including dynamics and interactions • Biomolecular mapping microscopies [confocal, CryoEM, SPM (AFM, STM, others)] • Image-interpretation tools • Visualization • Computational systems • Databases	Modular analytical and imaging instrumentation and methods integrated with culturing, monitoring, control, and manipulation modalities to assess: • Viability of integrated approaches • Compatability with living systems • Intermodal interactions • Ability to meaningfully assess single cells • Data integration • Simulation and modeling integrated into experimental methods • Visualization of multimodal analyses and system monitoring and manipulation	Integrated culturing capabilities within multiprobe instrumentation for simultaneous control, manipulation, and multimodal analyses of structured communities: • Nondestructive • Real time • Linked databases	Characterizations of microbial communities in realistic environments at the environmental, community, cellular, and molecular levels: • Spatial • Temporal • Functional • Process • Molecular Databases and query tools Protocols QA/QC

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

Table 4. Computing Roadmap: Facility for Analysis and Modeling of Cellular Systems

Topic	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment
LIMS and Workflow Management Participate in GTL cross-facility LIMS working group	Available LIMS technologies Process description for LIMS system Crosscutting research into global workflow management systems Approaches to guiding experiment-based production protocols to inform how best to produce a protein as an AI system helps develop strategy for production	Prototype cellular systems LIMS system* Characterization design strategy system Workflow-management system for identification and characterization Process simulation for facility workflow	Cellular systems LIMS and workflow system Workflow integrated with other GTL facilities and experimental strategy systems
Data Capture and Archiving Participate in GTL cross-facility working group for data representation and standards	Data-type models* Technologies for large-scale storage and retrieval Preliminary designs for databases	Prototype storage archives Prototype user-access environments	Archives for key large-scale data types* Archives linked to community databases and other GTL data resources GTL Knowledgebase feedback
Data Analysis and Reduction Participate in GTL cross-facility working group for data analysis and reduction	Algorithmic methods for various modalities* Grid and high-performance algorithm codes Design for tools library Approaches for automated image interpretation in confocal light microscopy/FRET	Prototype visualization methods and characterization tools library* Prototype grid for data analysis, with partners Prototypes for automated image interpretation in confocal light microscopy and FRET Analysis tools linked to data archives	Production-analysis pipeline for various modalities* on grid and HP platforms Large-scale experimental data results linked to genome data Automated image interpretation in confocal light microscopy and FRET Repository for production-analysis codes Analysis tools pipeline linked to end-user problem-solving environments
Modeling and Simulation Participate in GTL cross-facility working group for modeling and simulation	Existing technologies explored for cell-system modeling and simulation Research methods for reconstruction of protein interaction, regulatory networks, metabolic pathways, and community interactions Mathematical methods for multiscale stochastic and differential-equation network models	Experimentally guided metabolic reconstruction Signaling and regulatory-network reconstruction and simulation Efficient modeling methods for community-interaction networks Mature methods for reconstructing protein-interaction and regulatory networks	Production pipeline and end-user interfaces for cellular and community-level combined network reconstruction and simulation Production codes for image time-series analysis
Community Data Resource Participate in GTL cross-facility working group for serving community data	Data-modeling representations and design for databases: In vivo protein expression and localization, cell models and simulations, community models and simulations, cellular and community methods and protocols	Prototype database End-user query and visualization environments Integration of databases with other GTL resources	Production databases and mature end-user environments Integration with other GTL resources and community protein-data resources
Computing Infrastructure Participate in GTL crosscutting working group for computing infrastructure	Analysis, storage, and networking requirements for cellular systems data Grid and high-performance approaches for large-scale data analysis for cellular and community networks and simulations and to establish requirements	Hardware solutions for large-scale archival storage Networking requirements for large-scale grid-based MS and image data analysis	Production-scale computational analysis systems Web server network for data archives and workflow systems Servers for community data archive databases

* Data types and modalities include MS, NMR, neutron scattering, X-ray, confocal microscopy, cryoEM, and process metadata. Large-scale experimental data results are linked with genome data, and feedback is provided to GTL Knowledgebase.

6.0. GTL Development Summary: Global, Crosscutting, and Long-Lead Issues

6.1. Coordinated GTL Program and Facility Development 192

6.2. Biology Drivers and Issues 192

6.2.1. Recalcitrant Proteins and Complexes 192
6.2.2. Biosample Growth and Culturing 192
6.2.3. Affinity Reagent Libraries 193
6.2.4. Characterization of Proteins and Complexes 193

6.3. Technology Drivers and Issues 193

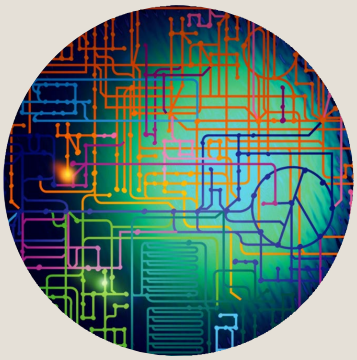
6.3.1. Technologies for Measurement of Proteins, Metabolites, and Molecular Machines 193
6.3.2. MEMS, Microfluidics, and Nanotechnology 193
6.3.3. Single-Cell Analysis 194
6.3.4. Imaging 194
6.3.5. Data-Quality Standards 194

6.4. Computing, Communications, and Information Drivers and Issues 194

6.4.1. Computational Methods for Experimental Data Analysis 194
6.4.2. Process Control, LIMS, Workflow Management 194
6.4.3. Data Architecture, Modeling, and Integration 195
6.4.4. Computing Hardware and Networking Infrastructure 195
6.4.5. Computational Models for Establishing Networks and Simulations 195
6.4.6. Genome Annotation 195
6.4.7. Computing, Communications, and Information 196

6.5. Other Issues 196

6.5.1. Ethical, Legal, and Social Issues (ELSI) 196
6.5.2. Technology Transfer 196
6.5.3. Industrial Involvement 196



GTL Development Summary: Global, Crosscutting, and Long-Lead Issues

6.1. Coordinated GTL Program and Facility Development

In the preparation of this roadmap, numerous crosscutting and long lead time issues have emerged that require a long-range perspective as well as a globally coordinated approach to technology development (see 1.3.6.2. Integrated Management and Development, p. 11). These include scientific, technological, computing, and governance issues across the GTL program and facilities. Key topics, drivers, and issues include the following:

6.2. Biology Drivers and Issues

Some technical advances needed in the biology underlying GTL production and analytical technologies are discussed below.

6.2.1. Recalcitrant Proteins and Complexes

Drivers: Production and analysis methods are needed for membrane, secreted, transient, and large proteins; proteins having unusual cofactors; those normally found in complexes; and others.

Issues: What should we know about these proteins and complexes? What are the most promising technologies for production and analysis?

- For all proteins, the relative efficacy of in vivo vs in vitro technologies must be determined.

6.2.2. Biosample Growth and Culturing

Drivers: High-fidelity measurements demand well-defined biosamples.

Issues: New experimentation will require intricate design and control, including simulation and modeling and real-time environmental and biological characterization.

- Techniques must encompass complex biology such as biofilms, mixed cultures, environmental samples, and "unculturables."

Note: Long-lead items are those for which the acquisition time (including development and procurement) is longer than the time allotted for a given facility construction project. Crosscutting items are critical to more than one of the GTL facilities and elements.

6.2.3. Affinity Reagent Libraries

Drivers: Comprehensive coverage of affinity reagents is a key product of the Protein Production and Characterization Facility and a critical analysis tool for the whole GTL enterprise.

Issues: Current libraries of affinity reagents, such as single-chain variable domains of antibodies, are not appropriate for high-throughput analyses and downstream applications in GTL. Many possible molecular scaffolds could be used as foundations for reagent libraries, and the usefulness and limitations of these molecular scaffolds should be evaluated. More exploration of novel libraries and development of improved affinity reagent libraries also will be needed.

6.2.4. Characterization of Proteins and Complexes

Drivers: Subsequent utilization and analyses will require a minimal set of characterizations for each protein and complex.

Issues: Community involvement will help to generate specific requirements for the characterization of proteins and molecular machines. These include biophysical characterization such as measurements of size, shape, stoichiometry, and organization, as well as biochemical assays designed to screen for functional activity.

6.3. Technology Drivers and Issues

GTL is highly dependent upon and should stimulate next-generation analytical and imaging approaches for measuring the parameters of microbial molecular and cellular systems. Many of these technologies are undergoing rapid advancement, and some will impact multiple GTL facilities. Requirements for successful facility operations, technical challenges and gaps, and the roadmap for acquiring necessary technologies should be defined.

6.3.1. Technologies for Measurement of Proteins, Metabolites, and Molecular Machines

Drivers: GTL will set requirements for measuring the presence and absolute quantity of many types of molecules in microbial systems. Technologies that can accomplish these measurements should be explored and evaluated regarding their potential for achieving the prerequisite detectivity, sensitivity, accuracy, dynamic range, and throughput to support facility operations. Critical considerations in proteomics, machines, and microbial-system analytical protocols must be resolved based on science goals for the research programs and technology capabilities and limitations.

Issues: Investigators need small sample volumes, measurement reproducibility, sensitivity, dynamic range, high sample quality, low cost, and the achievement of high throughput with technologies that currently are largely manual.

- Challenges include quantitative methods for mass spectrometry, reproducibility, dynamic range, robustness, and global coverage for gene products and regulatory molecules; sample preparation for different molecular classes and cellular fractions; stoichiometry of molecular machine partners; and high-throughput operations.
- Other technologies including arrays also must be investigated.

6.3.2. MEMS, Microfluidics, and Nanotechnology

Drivers: Technology miniaturization has the potential to make huge impacts in protein characterizations and other experimentation by reducing the amount of material and reagents needed. The nature and scope of GTL facilities could be affected dramatically through broad application of such technologies. Micro- and nanotechnologies have the potential to provide new functionalities in detection, manipulation, and analysis of biological systems. These technologies include microfluidics and microelectromechanical systems (MEMS).

Issues: Concerted effort is required to define and develop the use of microtechnologies for all facets of GTL science.

6.3.3. Single-Cell Analysis

Drivers: Many key questions and challenges will require a single-cell capability, including population heterogeneity for proteomics, culture issues, sample minimization, validation testing for presence of machines, in situ analysis, analysis of cellular specialization in communities, and temporal and spatial resolution of cellular systems processes as the ultimate test of systems models.

Issues: Single-cell analysis has challenges in instrumentation, robustness, detection limits, dynamic range, and handling.

6.3.4. Imaging

Drivers: New imaging modalities are required for localization and validation of complexes, dynamics, docking, intercellular communication, extracellular matrix, and metabolite distribution.

As technologies mature to enable the examination of molecular machines in vivo, high-throughput and automated image-acquisition and analysis capabilities will be needed.

Issues: Technical challenges include consistent and benign label incorporation to preserve the functionality of proteins and machines and provide for data acquisition and interpretation, multimodal imaging, high-throughput considerations, chemical analyses, sensitivity, and spatial and time resolution.

6.3.5. Data-Quality Standards

Drivers: Standards are the critical foundation in experimental design, data capture and analysis, and informatics and computational approaches in a data-intensive environment.

Issues: Issues extend across all facilities and throughout the research programs—including data definition, integration, "error" expression, and reference points.

6.4. Computing, Communications, and Information Drivers and Issues

To be successful, GTL must coordinate the analysis of vast amounts of data, share metadata about experimental processes and workflow, manage the combined output of joint experiments, and provide common gateways for the user community to access the data, models, and simulations of microbial systems. The program also must provide for shared hardware, tools, and network infrastructure. This will require long lead times and a coordinated research and development approach.

Key computing, communication, and information needs include the following.

6.4.1. Computational Methods for Experimental Data Analysis

Drivers: Vast amounts of data will be produced by many different methods. These data must be analyzed quickly enough to keep up with data production, using algorithms sufficient to extract information needed by researchers.

Issues: Challenges include advancement in algorithm design to more accurately extract and quantitate observations from raw data (e.g., mass spectrometry, NMR, scattering, expression analysis). Other challenges are the development of methods for large-scale distribution and management of analysis processes such as computing grids; configurable analysis tool pipelines linked to integrated data resources; and environments for researchers to facilitate large-scale analysis processes.

6.4.2. Process Control, LIMS, Workflow Management

Drivers: Operations at the GTL facilities will be coordinated using an integrated workflow process. This workflow environment is significantly more complex than those used for sequencing facilities. Large-scale

experiments will require development of a shared experimental strategy, a uniform pipeline with strongly coupled measurements, sharing of process metadata, and electronic investigator collaboration and coordination.

Issues: The GTL facilities will require a laboratory integrated management system (LIMS), electronic notebooks; collaboratory environments; process optimization; dynamic process scheduling; and sample archiving, tracking, prioritization, and storage. These technologies must be developed in a consistent fashion across the facilities using common technologies and data standards.

6.4.3. Data Architecture, Modeling, and Integration

Drivers: High-throughput processes will generate massive amounts of information that must be shared across facilities and with experimental planners and the community. Collected knowledge of all facility experiments and measurements needs to be captured as endurable data. Reduced data incorporated into models of biological systems and simulations based on these models must be managed as part of the data structure. All aspects of the data infrastructure must be integrated across the program.

Issues: Global database development, the most complex and important element of GTL, needs a representative working group to define and model data types, explore data management and user access technologies, and establish working standards.

6.4.4. Computing Hardware and Networking Infrastructure

Drivers: Analysis, modeling, and simulation from large data sets will require computing and networking capabilities and capacities well beyond existing infrastructures and those proposed for the next generation of computing platforms. This information-intensive undertaking will need terascale communications; distributed (grid) approaches to capacity computing problems; and environments for petascale, numerically intensive, physics-based simulations.

Issues: New hardware architectures will require extensive acquisition lead time. A working group should specify performance specifications that drive new architectures and communication technologies.

6.4.5. Computational Models for Establishing Networks and Simulations

Drivers: Methods and mathematical approaches will be used for modeling, simulating, and visualizing complex types of cellular processes and interactions. Substantial enhancement and maturation in mathematical methods and theory are required before systems with realistic complexity can be considered.

Issues: Existing modeling has not dealt with levels of complexity and uncertainty or the magnitude of data with which we will be working. Stochastic effects of such systems should be better represented. Also, the robustness and sensitivity of system models to data errors or omissions need to be better understood. The preference is to use models to drive experimental designs.

6.4.6. Genome Annotation

Drivers: GTL research is based on the availability of fully finished and annotated (nondraft) genomes. Annotation is the foundation for experimental planning to coordinate activities of the four facilities and provide significant initial information about protein function and significance in each genome.

Issues: Many enhancements in annotation are needed, especially those related to recognition of interacting proteins, regulatory signals and structures, infrastructure for large-scale genome analysis processes, and databases for microbial genomes.

6.4.7. Computing, Communications, and Information

Working groups on the following topics should be established to examine issues of global value to GTL:

- Microbial genome annotation and data management
- LIMS and workflow management
- Data-analysis algorithms and large-scale processing
- Data infrastructure, data modeling, databases, and data standards
- Regulatory, metabolic, and cell modeling
- Hardware, grid, and networking infrastructure

6.5. Other Issues

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

GTL is largely a microbiology program, but it encompasses many scientific activities that might be expected to impact society in a number of ways. DOE is committed to stressing the close coordination of ELSI studies with the ongoing science.

6.5.2. Technology Transfer

Transfer is key to infusing new technologies into the facilities and to rapidly making innovative concepts and technologies commercially available for broader biological research and DOE mission applications.

6.5.3. Industrial Involvement

Industrial vendors and developers will be important in the rapid development of new instruments and modalities for facilities and their subsequent introduction to the broader community. Clear policies and procedures involving industry need to be developed and put in place to facilitate this process.

Appendix A. DOE Mission: Energy Security

A.1.1. The Energy Challenge 198

A.1.2. The Role of Biology and Biotechnology in America's Energy Future 199

A.1.3. GTL's Vision for Biological Energy Alternatives 201

A.1.4. Ethanol from Biomass 203

A.1.4.1. Cellulose Degradation and Conversion 203

A.1.4.2. Bioethanol Research Targets for GTL 205

A.1.4.2.1. Gaps in Scientific Understanding 205

A.1.4.2.2. Scientific and Technological Capabilities Required to Achieve Goals 207

A.1.5. Biohydrogen Production 208

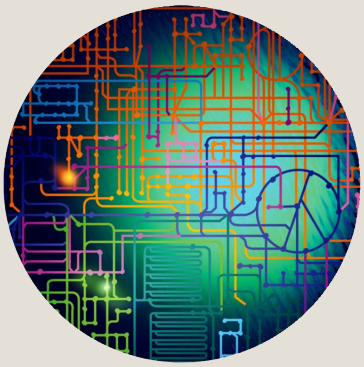
A.1.5.1. Biophotolysis of Water 209

A.1.5.2. Biohydrogen Research Targets for GTL 209

A.1.5.2.1. Gaps in Scientific Understanding 210

A.1.5.2.2. Scientific and Technological Capabilities Required to Achieve Goals 212

A.1.6. Summary 214



DOE Mission: Energy Security

Develop Biofuels as Major Energy Source

The Department of Energy's (DOE) strategic energy goal is to "protect our national and economic security by promoting a diverse supply and delivery of reliable, affordable, and environmentally sound energy" (*DOE Strategic Plan* 2003).

GTL supports this goal by providing the fundamental scientific knowledge needed to develop biological technologies that produce clean, renewable, carbon-free, and carbon-neutral alternatives to fossil fuels (see Mission Science Goals and Challenges, below right). These biotechnical advances will expand and improve our current portfolio of options and ultimately help lead to a secure, sustainable energy future for the United States and the world.

A.1.1. The Energy Challenge

Meeting projected increases in energy demand while decreasing dependence on foreign sources of energy defines America's energy challenge. From 2003 to 2025, U.S. energy demand is projected to increase by 35%, much greater than the projected increase in domestic production (Annual Energy 2005). Making up this projected shortfall without increasing imports will require investments in science and technologies that will improve conservation and efficiency and expand the domestic energy supply system. A primary goal of the national energy policy is not only to increase domestic supply but also to broaden our range of options in ways that will reduce vulnerabilities to supply disruptions and protect the environment (Reliable 2001).

Another key factor in America's energy challenge is rising carbon dioxide (CO_2) emissions. CO_2 is the most abundant greenhouse gas (GHG) in the atmosphere, and, based on projected energy use between 2003 and 2025, U.S. CO_2 emissions could increase almost 40% (Annual Energy 2005). With accelerated growth in fossil-fuel consumption projected for developing regions of the world, by 2025 annual global CO_2 emissions could be 55% higher than in 2001 (International Energy 2004). In 2002, global energy use emitted about 7 gigatons of CO_2 into the atmosphere. Several long-term projections estimated that CO_2 emissions could be as high as 30 GtC/year by 2100 (Nakicenovic et al. 2000). Stabilizing the concentration of CO_2 at any level requires that global CO_2 emissions must peak eventually and begin a long-term decline, ultimately falling to virtually zero.

Mission Science Goals and Challenges

Mission Science Goals: Understand the principles underlying the structural and functional design 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.

Challenges: Analyze thousands of natural and modified variants of such processes as cellulose degradation, fermentative production of ethanol or other liquid fuels, and biophotolytic hydrogen production.

A variety of breakthrough energy technologies will be needed to significantly reduce CO_2 emissions. To illustrate the scale of the CO_2 emissions challenge, Table 1. How Big is a Gigaton?, this page, provides examples of the types of technological actions required to reduce emissions by 1 GtC per year (Pacala and Socolow 2004).

Strategies for understanding the impacts of energy use on climate change and for developing technologies that will ensure economic prosperity while reducing GHG emissions are provided under the guidance of several government agencies through the Climate Change Science Program (CCSP 2003) and the Climate Change Technology Program (CCTP, www.climatetechnology.gov) (see Appendix F. Strategic Planning for CCSP and CCTP, p. 249).

A.1.2. The Role of Biology and Biotechnology in America's Energy Future

Biology played a key role in producing the fossil fuels so critical to meeting today's world energy demand. Fossil fuels were once living biomaterials synthesized eons ago by photosynthetic and biochemical processes. A series of fortuitous geological events trapped these materials beneath the sediments of ancient seas, and, over millions of years, the right mix of heat, pressure, and other factors transformed the biomaterials into fossil fuels.

With biotechnological innovations, biology once again can play an important role in producing high-energy fuels. Plants and photosynthetic microorganisms are masters at harvesting chemical energy from sunlight—a virtually inexhaustible supply of energy. By harnessing their photosynthetic and other biochemical capabilities, biological systems can be used to satisfy a greater portion of energy demand.

Applying biology to build a new U.S. bioenergy industry can benefit this nation's energy security, economy, and environment in many different ways. Biofuels, especially ethanol from plant materials (biomass), have the potential to reduce our dependency on foreign oil in the transportation sector and diversify our energy-technology portfolio. As renewable alternatives that can be harvested on a recurring basis, bioenergy crops (e.g., poplar trees and switchgrass) and agricultural residues (e.g., corn stover and wheat straw) can provide American farmers with important new sources of revenue. Consumption of biofuels produces no net CO_2 emissions, releases no sulfur, and has much lower particulate and toxic emissions than fossil fuels (Greene et al. 2004). In addition to ethanol, other biobased energy alternatives include biodiesel, methanol, hydrogen, and methane (see sidebar, Biological Energy Alternatives, p. 200).

Biomass currently is used to meet only 3% of U.S. energy consumption (Annual Energy 2005). In 2004, the U.S. produced 4 billion gallons of ethanol from corn grain, enough to meet about 2% of U.S. gasoline

Table 1. How Big is a Gigaton?

Today's Technology	Actions that Provide 1 Gt/year of CO_2 Mitigation
Coal Plants	Replace 1000 conventional 500-MW plants with "zero-emission" power plants
Geologic Sequestration	Install 3700 sequestration sites the size of Norway's Sleipner Project
Nuclear	Build 500 1-GW plants
Efficiency	Deploy 1 billion cars running at 40 mpg instead of 20 mpg
Wind	Install 150× current U.S. wind generation
Solar photovoltaics	Install 10,000× current U.S. solar PV generation
Biomass fuels from plantations	Globally convert open land >15× the size of Iowa's farmland to biomass production
Storage in new forests	Reforest open land >40× the size of Iowa's farmland

consumption (Homegrown 2005; Mann 2004). Ethanol from biomass has promise for meeting a significantly larger portion of U.S. gasoline demand, but higher production costs, technical difficulties, and inefficiencies in biomass conversion currently prevent ethanol from being cost-competitive with gasoline. Another concern has been the uncertainty in determining how much land must be dedicated to growing bioenergy crops to make a real difference in oil demand and how this would impact current agricultural and forestry practices. A recent report prepared for the U.S. Department of Agriculture and Department of Energy (DOE) has projected that relatively modest changes in the use of farmlands and forests could produce more than 1.3 billion dry tons of biomass per year, enough to reduce current oil demand by about

Biological Energy Alternatives

Most biological processes that produce energy require solar energy either directly or indirectly via photosynthesis, a complex biochemical pathway in which solar energy is used to drive the chemical conversion of low-energy inorganic molecules such as water and carbon dioxide into energy-rich organic molecules. The organic products of photosynthesis are used to build biomass (proteins, fats, carbohydrates, and cellulose) and store chemical energy needed to drive cellular processes. The biomass of photosynthetic organisms can be used directly as a burnable fuel or converted to such other high-value energy sources as ethanol, biodiesel, methanol, hydrogen, or methane.

Liquid Fuels

Ethanol: Currently the most widely consumed biofuel in the United States, used as a substitute or octane booster for gasoline. A gallon of this biofuel has about 2/3 the energy content of gasoline. Some 3 billion gallons of ethanol were produced from cornstarch in 2004, equaling about 2% of U.S. gasoline consumption (Homegrown 2005; Mann 2004). Inefficiencies in the conversion of biomass (e.g., agricultural residues, plant stems and leaves, grasses, trees, and municipal wastes) to ethanol prevent yields that could meet a larger portion of gasoline demand.

Methanol: High-octane liquid fuel that has about half the energy density of gasoline. Engine modifications are required to improve cold starts and prevent corrosion. In the United States, about a billion gallons of methanol are produced each year, primarily from methane, but methanol also can be thermochemically derived from biomass gasification. Methanol could be a future source of hydrogen for fuel cell vehicles.

Biodiesel: Diesel fuel substitute or extender obtained from chemically reacting organically derived oils and fats (e.g., excess soybean oil and restaurant greases) with alcohol to form ethyl or methyl esters. In its pure form, biodiesel reduces fuel economy and power by about 10% when compared with diesel. Biodiesel blends perform similarly to diesel and can be used in unmodified engines. Only about 30 million gallons of biodiesel are produced each year in the United States today—a tiny fraction of the billions of gallons of diesel consumed each year (National Biodiesel Board).

Gaseous Fuels

Hydrogen: Potential energy source that can be released from the breakdown of biomass by microorganisms or produced directly from water and sunlight via photobiological processes that do not require biomass as an intermediate. Much research is needed, however, before we can use these systems for clean, renewable hydrogen production. Currently, most hydrogen is derived from steam reformation of nonrenewable natural gas and used primarily for industrial chemicals production. Only a small fraction is used as an energy carrier. Each year in the United States, about 9 million tons of hydrogen are produced, enough to power 20 to 30 million hydrogen cars or 5 to 8 million homes (National Hydrogen Energy Roadmap 2002).

Methane: Main chemical component of the fossil-fuel natural gas, which currently makes up about 20% of the U.S. energy supply. Microorganisms naturally produce methane during biomass degradation. Extensive infrastructure already is in place for widespread distribution and use. Organic materials in agricultural, municipal, and industrial wastes could be used as feedstock for biomethane production; however, high production costs and incomplete biological conversion (as much as 50% of organic matter is not used) are major limitations.

one-third (Biomass as Feedstock 2005). As research improves efficiencies in both agricultural production and biomass conversion, land and sunlight availability in the United States should be sufficient to produce enough biofuels to meet domestic transportation-related demand without disrupting agricultural land use for food and fiber crops.

In addition to reducing our dependence on oil, biofuels also have great potential for decreasing greenhouse gas emissions associated with fossil-fuel consumption. Figure 1. Potential Role of Biotechnology in the Global Energy System, p. 202, presents the results of an economic analysis exploring conditions under which markets for commercial biofuels could develop (Edmonds et al. 2003). In this figure, two potential scenarios for global energy consumption in the 21st Century are compared: A reference case in which innovations in energy technology take place without constraints on CO_2 emissions and a CO_2-stabilization case in which emissions are limited. In the stabilization case, biomass becomes a major component of the energy-technology portfolio, and by 2100 biomass usage is greater than that of all current fossil fuels (oil, natural gas, and coal) combined. A transition to such large-scale use of biofuels and biotechnologies could create a new bioenergy industry potentially worth trillions of dollars over the 21st Century.

Before biomass and biotechnologies can compete successfully with established energy sources for market share, basic research is needed for a more complete understanding of the biological processes underlying biofuel production. Applying this understanding in innovative ways will enable the development of breakthrough technologies. Since it can take 30 to 50 years for an energy technology to go from research to large-scale commercial deployment, this basic research is needed today.

A.1.3. GTL's Vision for Biological Energy Alternatives

GTL will provide a systems-level understanding of biological processes for developing and deploying large-scale, environmentally sound biotechnologies to produce biofuels and other high-value chemical products that reduce dependence on foreign energy sources and enhance national economic prosperity.

A national vision for bioenergy and biobased products was defined by the Biomass R&D Technical Advisory Committee (BTAC): "By 2030, a well-established, economically viable bioenergy and biobased products industry will create new economic opportunities for rural America, protect and enhance our environment, strengthen U.S. energy independence, provide economic security, and deliver improved products to consumers" (Vision for Bioenergy 2002). BTAC, established as a result of the Biomass Research and Development Act, is responsible for advising the Secretary of Agriculture and the Secretary of Energy on issues relevant to biomass research and development. BTAC also coordinates partnerships among government agencies, industry, researchers, and other groups with interests in biomass R&D (U.S. Congress 2000).

GTL supports this national vision by providing a detailed understanding of the microbial processes that mediate the production of biofuels (see sidebar, Mission Science Goals and Challenges, p. 198). Our limited understanding of many of these processes presents fundamental scientific challenges that must be overcome before we can develop and deploy successful bioenergy technologies. In addition to advancing biofuel production, the capabilities and understanding of microbial systems provided by GTL will be applicable to the biotechnological development of other commercial chemical processes. Techniques used to design microbial systems for biofuel production could be used to develop other microbial systems optimized to convert biomass to biodegradable plastics and other chemical products currently derived from fossil fuels. Insights from GTL research could benefit several research areas supported by DOE's Office of Energy Efficency and Renewable Energy (EERE) (see sidebar, DOE Activities Complementary to GTL Research, p. 203).

The rest of this chapter will explore current science and technology gaps and research capabilities needed to overcome key challenges in two areas of applied research in bioenergy: Ethanol from biomass and biohydrogen.

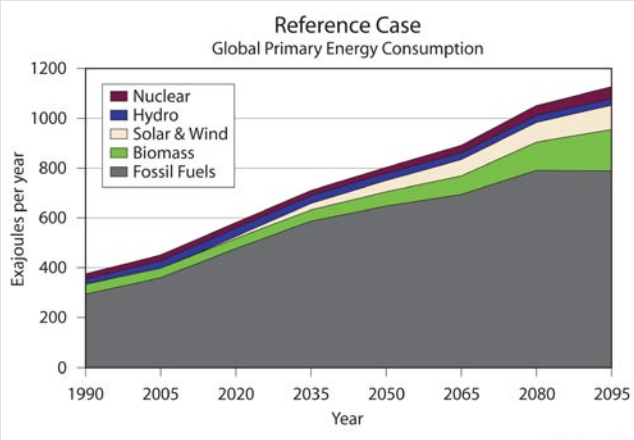


Fig. 1A. Reference Case

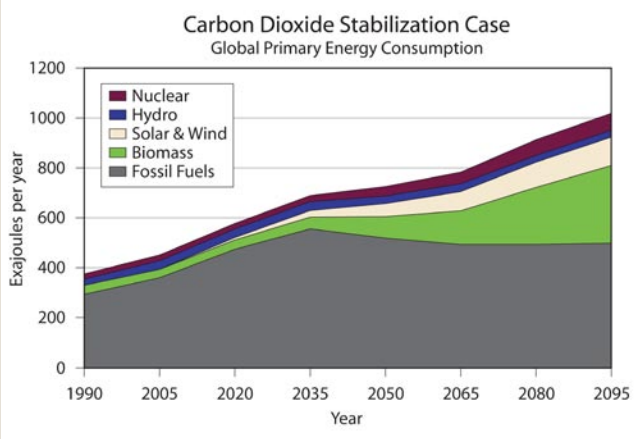


Fig. 1B. Carbon Dioxide Stabilization Case

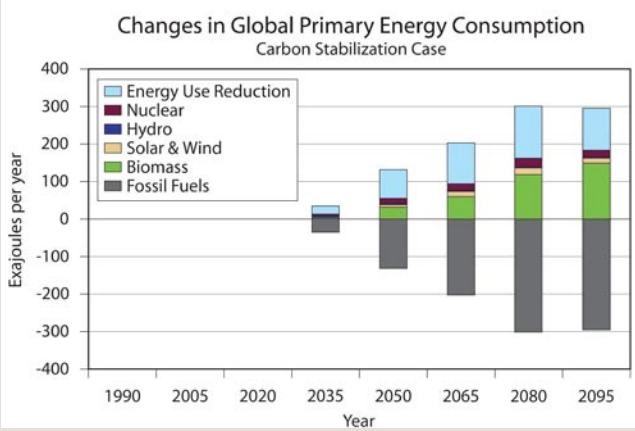


Fig.1 C. Changes in Global Primary Energy Consumption

*The case of 550 ppmv was chosen to illustrate the types of changes that might occur; currently, no scientific basis exists for preferring any particular CO_2 concentration.

Fig. 1. Potential Role of Biotechnology in the Global Energy System. These diagrams show results of an economic analysis that considered competition among energy technologies in the 21st Century and explored conditions under which biological energy sources could develop (Edmonds et al. 2003). Fig. 1A presents the MiniCAM B2 reference case (Edmonds et al. 2004). In this scenario, the world's economic activity and number of inhabitants continue to grow, with the population reaching 9.4 billion by 2100. Energy technologies continue to improve; however, strategies to address global environmental challenges (such as mitigating greenhouse gas accumulations) are not a priority.

Fig. 1B shows another possible energy-consumption scenario in which a global commitment has been made to stabilize long-term atmospheric CO_2 concentration at 550 ppmv (about double the preindustrial level of 280 ppmv); the current level is around 380 ppmv.* Placing limits on CO_2 emissions provides an incentive for developing noncarbon-emitting energy technologies and reducing energy consumption through conservation and improvements in energy efficiency. Over the century, increased biofuel consumption combined with reductions in energy use would displace hundreds of exajoules of fossil-fuel energy (Fig. 1C), and by 2100 biofuels would equal roughly all fossil-fuel usage today (coal + oil + natural gas). By decreasing fossil-fuel use in the stabilization case, hundreds of gigatons of CO_2 emissions would be avoided (Fig. 1D).

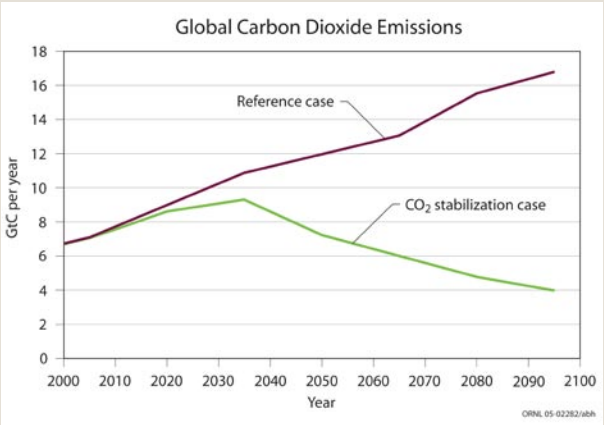


Fig. 1D. Global Carbon Dioxide Emissions

A.1.4. Ethanol from Biomass

A.1.4.1. Cellulose Degradation and Conversion

Understanding the conversion of biomass to ethanol begins with understanding the structural and chemical complexity of the three primary polymers that make up plant cell walls: Cellulose, hemicellulose, and lignin (see Fig. 2. Cellulose Structure and Hydrolysis Challenges, p. 204). Depending on plant species and cell type, the dry weight of a cell wall typically consists of about 35 to 50% cellulose, 20 to 35% hemicellulose, and 10 to 25% lignin (Saha 2004). Cellulose is the most abundant biomaterial on earth. Each cellulose molecule is a linear polymer of glucose residues. Depending on the degree of hydrogen bonding within and between cellulose molecules, this polysaccharide is found in crystalline or paracrystalline (amorphous) forms. Cellulose exists within a matrix of other polymers, primarily hemicellulose and lignin. Hemicellulose is a branched sugar polymer composed of mostly pentoses (five-carbon sugars) and some hexoses (six-carbon sugars). Lignin is a complex, highly cross-linked aromatic polymer that is covalently linked to hemicellulose, thus stabilizing the mature cell wall. These polymers provide plant cell walls with strength and resistance to degradation, which also makes these materials a challenge to use as substrates for biofuel production.

Enzymes such as cellulases, hemicellulases, and other glycosyl hydrolases synthesized by fungi and bacteria work together in a synergistic fashion to degrade the structural polysaccharides in biomass. These enzyme systems, however, are as complex as the plant cell-wall substrates they attack. For example, commercial cellulase preparations are mixtures of several types of glycosyl hydrolases, each with distinctly different functions (exocellulases, endocellulases, exoxylanases, endoxylanases, cellobiases, and many others). Optimization of these enzymes will require a more detailed understanding of their regulation and activity as a tightly controlled, highly organized system.

DOE Activities Complementary to GTL Research

Office of Energy Efficiency and Renewable Energy (EERE)

www.eere.energy.gov

EERE Biomass Program: www.eere.energy.gov/biomass/. The Biomass Program supports the research and development of advanced technologies that transform biomass into biofuels, biopower, and high-value bioproducts. Through partnerships with industry, the Biomass Program is fostering a new domestic bioindustry that will use liquid-based biofuels to reduce U.S. dependence on foreign oil. The program has five core R&D activities: (1) Biomass Feedstocks, which develops technologies to provide biomass feedstock supplies to biorefineries; (2) Sugar Platform, which studies and optimizes the chemical and biological processes that break down biomass into raw sugar components; (3) Thermochemical Platform, which uses gasification, pyrolysis, and hydrothermal processes to convert biomass to intermediate products; (4) Products, which concentrates on chemical and biological processes that convert Sugar Platform and Thermochemical Platform outputs to final products such as fuels and chemicals; and (5) Integrated Biorefineries, which uses technical successes in the other four R&D areas to establish an integrated, market-ready biorefinery capable of employing biomass to make a range of such high-value bioproducts as fuels, chemicals, and biopower. GTL will play an important role in providing a better understanding of current microbial processes and discovering new microbial capabilities relevant to the Sugar Platform and Products research areas.

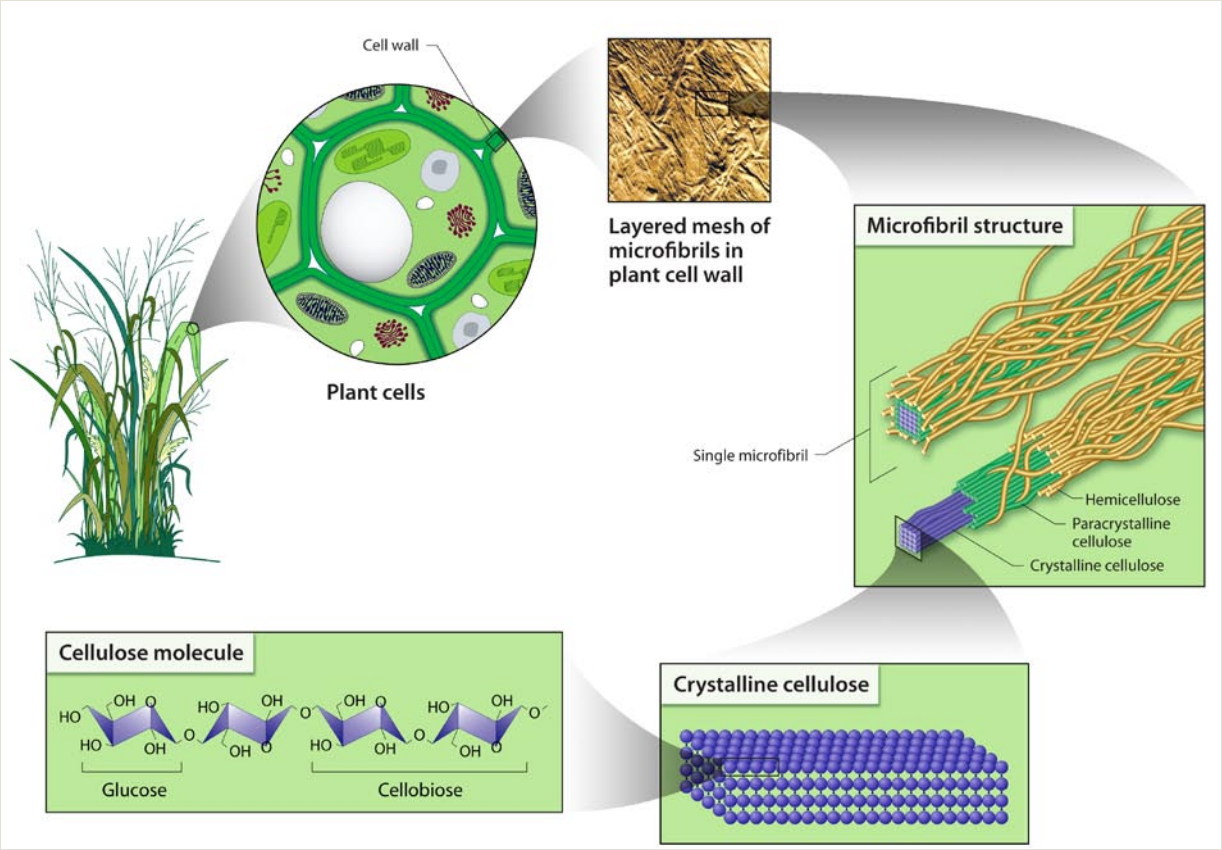
EERE Hydrogen Production: www.eere.energy.gov/hydrogenandfuelcells/hydrogen_production.html. EERE's Hydrogen, Fuel Cells, and Infrastructure Technologies Program aims to research and develop low-cost, highly efficient hydrogen-production technologies from diverse domestic sources. GTL science could benefit two related research areas: (1) biological and biomass-based production for improving efficiencies of anaerobic fermentation systems and (2) photolytic production of hydrogen by green algae.

Fig. 2. Cellulose Structure and Hydrolysis Challenges. Within the plant cell wall, chains of cellulose molecules associate with other polymers to form linear structures of high tensile strength known as microfibrils. Layers upon layers of microfibrils make up the cell wall.

Each microfibril is about 10 to 20 nm in diameter and may consist of up to 40 cellulose chains. A microfibril's crystalline and paracrystalline (amorphous) cellulose core is surrounded by hemicellulose, a branched polymer composed of a mix of primarily pentose sugars (xylose, arabinose), and some hexoses (mannose, galactose, glucose). In addition to cross-linking individual microfibrils, hemicellulose also forms covalent associations with lignin, a rigid aromatic polymer. Lignin is not pictured since its structure and organization within the cell wall are poorly understood. Pretreatment of biomass with enzymes or acids is necessary to remove the surrounding matrix of hemicellulose and lignin from the cellulose core prior to hydrolysis.

The crystallinity of cellulose presents another challenge to efficient hydrolysis. The high degree of hydrogen bonding that occurs among the sugar subunits within and between cellulose chains forms a 3D lattice-like structure. The highly ordered, water-insoluble nature of crystalline cellulose makes access and hydrolysis of the cellulose chains difficult for the aqueous solutions of enzymes. Paracrystalline cellulose lacks this high degree of hydrogen bonding, thus giving it a structure that is less ordered.

Each cellulose molecule is a linear polymer of thousands of glucose residues. Cellobiose, which consists of a pair of glucose residues (one right side up and one upside down) is the repeating unit of cellulose. [Microfibril portion of this figure adapted from J. K. C. Rose and A. B. Bennett, "Cooperative Disassembly of the Cellulose-Xyloglucan Network of Plant Cell Walls: Parallels Between Cell Expansion and Fruit Ripening," *Trends Plant Sci.* 4, 176–83 (1999).]



The biochemical conversion of biomass to ethanol currently involves three basic steps: (1) thermochemical treatments of raw lignocellulosic biomass to make the complex polymers more accessible to enzymatic breakdown; (2) production and application of special enzyme preparations (cellulases and hemicellulases) that hydrolyze plant cell-wall polysaccharides to a mixture of simple sugars; and (3) fermentation, mediated by bacteria or yeast, to convert these sugars to ethanol. A more complete understanding of enzymes and microbes involved in biomass conversion to ethanol is needed to overcome many current inefficiencies in the production process (see Table 2. Cellulosic Ethanol Goals and Impacts, this page; and Table 3. Cellulosic Ethanol Challenges, Scale, and Complexity, p. 206).

A.1.4.2. Bioethanol Research Targets for GTL

Improving Cellulase Systems. GTL will accelerate the development of optimal cellulase systems by providing resources for screening thousands of natural and modified enzyme variants, enabling the high-throughput production and functional analysis of these enzymes, elucidating regulatory controls and essential molecular interactions, and developing models for analyzing the structure and activity of natural and engineered enzyme systems.

Enabling the Development of Integrated Bioprocessing. A long-term target for GTL research is integrated bioprocessing, the conversion of biomass to ethanol in a single step. Accomplishing this requires the development of a genetically modified, multifunctional organism or a stable mixed culture capable of carrying out all biologically mediated transformations needed for the complete conversion of biomass to ethanol.

A.1.4.2.1. Gaps in Scientific Understanding

Without improving our understanding of microbial processes essential to bioethanol production, developing and improving technologies based on this understanding will be difficult. Biotechnology innovation requires basic research that explores a greater variety of enzymes and microorganisms, analyzes enzymes as systems,

Table 2. Cellulosic Ethanol Goals and Impacts

Factors	Today	Interim	Long-Term*
Billion gallons Fossil fuel displaced** CO_2 reduced	4 2% 1.8%	20 10% 9%	30 to 200 15 to 100%*** 14 to 90%
Feedstock****	Starch (14% energy yield)	Waste cellulose	Cellulosic energy crops (>37% energy yield)
Process	Starch fermentation Little cellulose processing	Acid decrystallization: Transition to enzymes Cellulases Single-sugar metabolism Multiple microbes Some energy crops	Enzyme decrystallization and depolymerization Cellulase and other glycosyl hydrolases Sugar transporters High-temperature functioning Multisugar metabolism Integrated processing Designer cellulosic energy crops Carbon sequestration through plant partitioning
Deployment	Large, central processing	Large, central processing	Distributed or centralized, efficient processing plants
Other impacts: Energy dollars spent at home, third crop for agriculture, land revitalization and stabilization, habitat, soil carbon sequestration, yield per acre roughly tripled (cellulose over corn starch).			

*Enabled by GTL.
**Current U.S. consumption of gasoline is about 137 billion gallons per year, which corresponds to about 200 billion gallons of ethanol (Greene et al. 2004) because a gallon of ethanol has 2/3 the energy content of a gallon of gasoline.
***Assumes improvements in feedstocks, processes, and vehicle fuel efficiency.
****Adapted from Smith et al. 2004.

and determines how certain factors influence biomass degradation or ethanol production. Several fundamental scientific questions in need of further investigation include:

- **What is the extent of natural diversity among biomass-degrading and ethanologenic organisms?** Over the last 30 years, most research devoted to ethanol production from cellulose has focused on fungal systems (primarily *Trichoderma reesei*) for the breakdown of cellulose into sugars coupled with the sugar-fermentation processes of yeast (*Saccharomyces cerevisiae*) (Demain et al. 2005). A deeper understanding of a greater variety of cellulolytic and ethanologenic systems is needed. Bacterial species in diverse physiological groups (e.g., bacteria with various tolerance levels for oxygen, temperature, and salt concentrations) are known to hydrolyze cellulose; thus a wide range of natural habitats could be explored for novel cellulotyic activities in bacteria.
- **How do soluble enzymes act on an insoluble crystalline substrate?** The hydrolysis of crystalline cellulose is the rate-limiting step in biomass conversion to ethanol because aqueous solutions of enzymes have difficulty acting on this insoluble, highly ordered structure. Cellulose molecules in their crystalline form are packed so tightly that enzymes and even small molecules such as water are unable to permeate the structure.
- **How do different biomass-degrading enzymes work together as a synergistic system?** Cellulases and hemicellulases are secreted from cells as free enzymes or as large, extracellular complexes known as cellulosomes. The collective activity of these enzyme systems is much more efficient than the individual activity of any isolated enzyme; therefore, to truly understand how these enzymes function, they must be studied as systems rather than individually or a few at a time. In addition, these systems eventually must be analyzed under laboratory conditions more representative of real-world environments. For example, laboratories often use purified cellulose as the substrate for enzyme analysis rather than more heterogeneous, natural lignocellulosic materials, and this can provide erroneous conclusions about natural enzyme activity.
- **Why are ethanologenic organisms less efficient at using certain sugar substrates?** A varied mix of hexoses (e.g., glucose, mannose), pentoses (e.g., xylose, arabinose), and oligosaccharides are released from the hydrolysis of lignocellulosic materials, and no microorganism is capable of fermenting all these sugars. The most widely studied ethanologenic microbes (e.g., yeast) prefer to use glucose as a substrate. Even when yeast cells are modified genetically to use xylose, they ferment all glucose before switching to the much slower xylose fermentation. Conversion rates can vary greatly depending on such factors as the type of sugar substrate being fermented, environmental conditions (e.g., pH, temperature), and the concentrations of certain products from other metabolic pathways.
- **How effective are sugar transporters at translocating different sugars across the cell membrane?** Sugar transporters are membrane-bound proteins that take up sugars from the environment and deliver them to

Table 3. Cellulosic Ethanol Challenges, Scale, and Complexity

Research and Analytical Challenges	Scale and Complexity
• Screening of databases for natural variants of cellulases (generally glycosyl hydrolases) and other enzymes or molecular machines in metabolic networks and characterization of variants • Analysis of modified variants to establish design principles and functional optimization • Modeling and simulation of cellulase, sugar transport, and multiple sugar-fermentation processes and systems • Integration of processing steps into single microbes or stable cultures	• Thousands of variants of all enzymes; screening of millions of genes, thousands of unique species and functions • Production and functional analysis of potentially thousands of modified enzymes, hundreds of regulatory processes and interactions • Models at the molecular, cellular, and community levels incorporating signaling, sensing, regulation, metabolism, transport, biofilm, and other phenomenology and using massive databases in GTL Knowledgebase • Incorporation of complete cellulose-degradation and sugar-fermentation processes into microbes or consortia—hundreds of metabolic, regulatory, and other interconnected pathways

the metabolic pathways inside cells. The inefficient transport of different sugar substrates by microbes can result in low product yield and is a major obstacle to the efficient conversion of biomass to ethanol. Our limited understanding of sugar transporters is due to a lack of adequate techniques for producing membrane proteins and studying their structure and function. Questions in need of investigation include: Can a glucose transporter transport other sugars, and, if so, how efficiently? Are some transporters better than others? Can transporters be modified for improved function?

- **Why do different enzymatic and microbial processes operate optimally at different temperatures?** Cellulases operate optimally at temperatures (>40°C) higher than those tolerated by ethanologenic organisms, so these two processes currently cannot be consolidated into a single process step. Thermophily (tolerance of high temperatures) improves the robustness of enzymes or microbes needed for industrial-scale processes and reduces the likelihood of culture contamination. The basis by which enzymes, pathways, and entire microbes are made thermophilic is understood poorly, and methods for inserting cellulolytic or fermentative pathways into thermophilic organisms are not well developed.
- **What are the requirements for producing and maintaining stable mixed cultures?** At a minimum, cultures used in bioethanol-production systems will need to be resistant or stable despite contamination by "outside" microbes or other potentially toxic materials or life forms. We currently do not understand in sufficient detail the dynamics of microbial consortia that carry out stable mixed processes such as aerobic and anaerobic digestion. Without this understanding, we will not be able to "design" or "engineer" such systems.
- **How can we improve systems for genetically engineering microorganisms involved in bioethanol production?** While many studies have expressed genes from cellulolytic organisms in *Escherichia coli* or other mesophilic organisms, systems for expressing foreign genes in cellulolytic or thermophilic organisms are in need of further development. Our current limited understanding of microbial regulation prevents the successful engineering of a microbe capable of versatile expression of lignocellulolytic enzymes, utilization of multiple sugars, and glycolysis.

A.1.4.2.2. Scientific and Technological Capabilities Required to Achieve Goals

Improving current understanding of bioethanol production will require a variety of new capabilities including techniques for surveying enzyme diversity; visualizing enzyme systems; efficiently producing enzyme systems and membrane proteins; cultivating microbial consortia; integrating transcriptomics, proteomics, and metabolomics; and genetically engineering microorganisms for integrated bioprocessing (see Table 3, p. 206). Specific needs include the following:

- **Ecogenomic approaches to explore the natural diversity of cellulases.** High-throughput sequencing and computational analysis of DNA from environments in which cellulose is widely available will lead to the discovery of genes for novel cellulase systems that could be used as templates for protein production.
- **Techniques to visualize cellulase systems in motion.** Advanced imaging techniques will provide new insights into how cellulases interact with crystalline cellulose and overcome current barriers to efficient cellulose hydrolysis (e.g., substrate accessibility, product or substrate inhibition, low product yield). Structural information and imaging from X-ray, nuclear magnetic resonance spectroscopy, scanning transmission electron microscopy, and other techniques will be needed to identify additional interactions between cellulases and other molecules needed for efficient function.
- **Large-scale production of cellulase enzyme systems, sugar transporters, and other proteins.** This will require improved methods for protein production and characterization. Currently, synthesis of sugar transporters and other membrane proteins is difficult, so analyzing the structure and activity of these proteins is challenging, if not impossible. High-throughput techniques and expression systems for efficiently producing membrane proteins, sets of different enzymes that work together, and enzyme complexes such as cellulosomes are in need of development. Access to validated expression systems for microorganisms with mission-relevant capabilities, including thermophilic, cellulolytic, and ethanologenic organisms, would help researchers spend less time on developing expression methods and more time on characterizing and improving proteins.

- **Methods to grow stable mixed cultures.** Improved experimental and modeling tools are needed to develop methods for producing a mixed microbial culture. The goal is to enable each population carrying out one part of the overall ethanol production process to perform stably.
- **Methods to integrate transcriptomic, proteomic, and metabolomic information.** Techniques that integrate information gathered from these global molecular measurements are essential to determining which genes are expressed and functionally active during cellulose utilization or ethanol fermentation and which metabolites influence the activity of enzymes involved in these pathways. As an insoluble substrate, cellulose cannot enter cells and induce the expression of genes involved in cellulose hydrolysis. Metabolic profiling could be used to identify which substrates or metabolites at what quantities activate or repress expression of key cellulolytic genes. In addition to illuminating regulatory strategies for cellulases and other coexpressed enzymes such as ligninases, these integrated omics approaches could be used to build regulatory and metabolic maps to guide genetic engineering. For example, these maps could be used to identify the best potential gene knockouts that redirect carbon flux from a particular sugar substrate toward ethanol fermentation and bypass competing pathways that produce other organic end products.
- **Methods to genetically engineer organisms for integrated bioprocessing.** *Clostridium thermocellum* is an anaerobic bacterium capable of both hydrolyzing cellulose and fermenting sugars to ethanol, but its yields are poor and conversion is slow. Improved methods for genetically modifying this and other cellulolytic microbes are needed. In one approach to developing an organism for integrated bioprocessing, a microbe naturally capable of hydrolyzing cellulose, such as *C. thermocellum*, is engineered to provide high product (ethanol) yields. In another approach, noncellulolytic microorganisms known to have high yields of ethanol are engineered to express cassettes of genes encoding cellulase enzyme systems. In either case, to achieve this ambitious goal of developing an organism capable of integrated bioprocessing, the current research paradigm must be altered to focus on understanding how microbial systems function and how their interacting pathways influence one another rather than focusing on only a few genes or enzymes.

A.1.5. Biohydrogen Production

Hydrogen is a promising energy carrier of the future: It can be derived from a variety of energy sources and used in fuel cells with high efficiency; "combustion" of hydrogen produces only water as a by-product, making it a nonpolluting, carbon-free energy alternative. The most common industrial methods for producing hydrogen include steam reformation of natural gas, coal gasification, and splitting water with electricity typically generated from fossil fuels. These energy-intensive industrial processes release carbon dioxide and other greenhouse gases and pollutants as by-products. Some microorganisms produce hydrogen naturally, and biotechnologies based on these microbial systems could lead to the development of clean, renewable sources of hydrogen. In a recent report on the hydrogen economy, however, the National Research Council (NRC) noted that "substantial, fundamental research needs to be undertaken before photobiological methods for large-scale hydrogen production are considered" (Hydrogen Economy 2004).

Several reviews have examined the potential of biological hydrogen production (Madamwar, Garg, and Shah 2000; Ghirardi et al. 2000; Melis and Happe 2001; Tamagnini et al. 2002; Levin, Pitt, and Love 2004; Nath and Das 2004; Prince and Kheshgi 2005). Although microorganisms produce hydrogen by different mechanisms, the step can be represented by the simple chemical reaction $2H^+ + 2e^- \text{<-->} H_2$. This reaction is known to be catalyzed by either nitrogenase or hydrogenase enzymes. Although alternative biological hydrogen production-pathways exist, each with its own set of advantages and disadvantages, the following discussion on biohydrogen production will focus on the challenges that must be overcome to improve one type of biological hydrogen production known as biophotolysis (see sidebar, Other Mechanisms for Biological Hydrogen Production, p. 209, and Table 4. Biophotolytic Hydrogen, p. 209).

A.1.5.1. Biophotolysis of Water

Under certain conditions, green algae and cyanobacteria can use water-splitting photosynthetic processes to generate molecular hydrogen (H_2) rather than fix carbon, the normal function of oxygenic photosynthesis (see sidebar, Photosynthetic Production of Hydrogen from Water, p. 210). Bidirectional hydrogenases in these organisms use electrons from the photosynthetic electron-transport chain to reduce protons to yield H_2. Biophotolysis holds potential for the scale of hydrogen production necessary to meet future energy demand. This approach to hydrogen production is promising because the source of electrons or reducing power required to generate hydrogen is water—a clean, renewable, carbon-free substrate available in virtually inexhaustible quantities. Another advantage of biophotolysis is the more efficient conversion of solar energy to hydrogen. Reengineering microbial systems for the direct production of hydrogen from water eliminates inefficiencies associated with carbon fixation and biomass formation. Theoretically, the maximal energetic efficiency for direct biophotolysis is about 40% (Prince and Kheshgi 2005) compared with a maximum of about 1% for hydrogen production from biomass (Hydrogen Economy 2004). Recognizing the important potential of biophotolysis, NRC has recommended that DOE "refocus its biobased program on more fundamental research on photosynthetic microbial systems to produce hydrogen from water at high rate and efficiency" (Hydrogen Economy 2004).

A.1.5.2. Biohydrogen Research Targets for GTL

Engineering Oxygen-Tolerant, Efficient Hydrogenases. Hydrogenases known to tolerate oxygen generally are not very efficient hydrogen producers. During biophotolytic hydrogen production, oxygen is released from the water-splitting reaction, thus engineering hydrogenases with sufficient activity and oxygen tolerance will be needed. Engineered hydrogenases then could be used in bioinspired nanostructures that maintain optimal conditions for hydrogen production.

Other Mechanisms for Biological Hydrogen Production

Nitrogenase-Mediated Hydrogen Production. In the absence of oxygen and presence of light, purple nonsulfur (PNS) photosynthetic bacteria such as *Rhodopseudomonas palustris* and *Rhodobacter sphaeroides* contain nitrogenase enzymes that can generate hydrogen under nitrogen-limited conditions. These microbes obtain the electrons they need to reduce protons to molecular hydrogen (H_2) from the breakdown of organic compounds. Certain species of cyanobacteria also contain nitrogenase enzymes capable of producing hydrogen as a by-product of nitrogen fixation.

Fermentative Hydrogen Production. A variety of bacteria such as *E. coli, Enterobacter aerogenes*, and *Clostridium butyricum* are known to ferment sugars and produce hydrogen using multienzyme systems. These "dark fermentation" reactions do not require light energy, so they are capable of constantly producing hydrogen from organic compounds throughout the day and night. Compared with other biological hydrogen-production processes, fermentative bacteria have high evolution rates of hydrogen. However, sugars are relatively expensive substrates that are not available in sufficient quantities to support hydrogen production at a scale required to meet energy demand.

Table 4. Biophotolytic Hydrogen: Goals and Impacts

- Sunlight and seawater, two resources in virtually limitless supply, can be used to produce the ultimate fuel and energy carrier, hydrogen. High-efficiency use of hydrogen in fuel cells can produce electricity directly with water as the by-product.
- This energy cycle is carbon free and can be developed as the complement to the electric grid for all energy applications—industrial, transportation, and residential.
- Development of biological photolytic processes to produce hydrogen at high rates and efficiency will enable the establishment of a hydrogen-economy strategy based on a renewable source.

Designing Microorganisms Optimized for Hydrogen Production. Photosynthetic microbes that have been genetically modified to produce hydrogen at high rates and efficiency from the biophotolysis of water could be grown in extensive farms of sealed enclosures (photobioreactors). Hydrogen would be harvested for use in energy applications, with oxygen released as a by-product.

A.1.5.2.1. Gaps in Scientific Understanding

Understanding biophotolysis well enough to model hydrogenase structure and function, regulatory and metabolic networks, and eventually entire organisms will stimulate the kind of biotechnological innovation needed to engineer the ideal organism to use in hydrogen bioreactors or the ideal enzyme-catalyst to use in bioinspired nanostructures for hydrogen production. But achieving this level of understanding will require basic research that investigates a greater range of hydrogen-producing enzymes and organisms, mechanisms of hydrogenase assembly, oxygen sensitivity of hydrogenase, electron-transfer rate limitations, and regulatory and metabolic processes that influence hydrogen production. Some specific issues relevant to these basic research needs follow.

- **What is the extent of natural diversity among hydrogenases and hydrogen-producing organisms?** A vast majority of organisms that contain hydrogenases have not been identified and probably cannot be cultured

Photosynthetic Production of Hydrogen from Water

Although microorganisms are capable of carrying out different types of photosynthesis, that found in plants, algae, and cyanobacteria is best understood. Photosynthesis in these organisms is a complex series of reactions that use light energy to drive electron transfer from water to carbon dioxide to yield carbohydrates.

Instead of using electrons harvested from water to synthesize carbohydrates from CO_2, under certain conditions green algae and cyanobacteria can use them to reduce protons and produce hydrogen gas (H_2). Molecular complexes involved in mediating electron flow from water to carbon-fixing or hydrogen-production reactions make up the photosynthetic electron-transport chain found in the thylakoid membranes of cyanobacteria and green algae. In eukaryotic green algae, thylakoid membranes are housed within a cellular organelle known as the chloroplast; in prokaryotic cyanobacteria, thylakoids are found in the cytoplasm as an intracellular membrane system (see Fig. A).

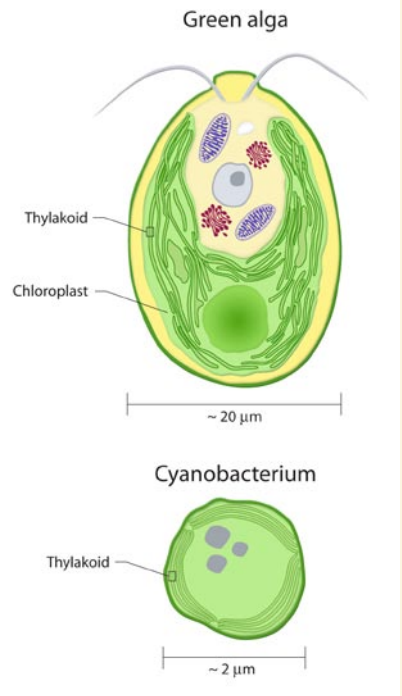


Fig. A. Thylakoids in Green Algae and Cyanobacteria.

An overview of steps involved in using light energy to produce carbohydrates or hydrogen is depicted in Fig. B and described below.

1. **Light Absorption by Photosystem II (PSII) Initiates the Photosynthetic Pathway.** PSII is a large molecular complex that contains several proteins and light-absorbing pigment molecules. The primary pigment molecules are chlorophylls and carotenoids, but cyanobacteria also have other pigments called phycobilins that absorb light at different wavelengths. The pigments are bound to proteins to form antenna complexes that absorb photons and transfer the resultant excitation energy to the reaction center of PSII, where energized electrons move to a small electron-carrier molecule. This molecule shuttles the excited electrons to the next complex in the photosynthetic electron-transport chain. To replace electrons lost in the transfer, the reaction center strips low-energy electrons from two water molecules, releasing four protons and an oxygen (O_2) molecule into the thylakoid space.

2. **Electron Transport Through the Cytochrome Complex Generates a Proton Gradient.** The electron carrier from PSII passes through the thylakoid membrane and transfers its electrons to the cytochrome complex, which consists of several subunits including cytochrome *f* and cytochrome b_6. A series of redox reactions within the complex ultimately transfer the electrons to a second electron carrier that acts as a shuttle to photosystem I (PSI). As electrons are transported through the complex, protons (H^+) outside the thylakoid are carried to the inner thylakoid space. The increase in proton concentration inside the thylakoid space creates a proton gradient across the thylakoid membrane.

3. **Light Absorption by PSI Excites Electrons and Facilitates Electron Transfer to an Electron Acceptor Outside the Thylakoid Membrane.** PSI is another large protein-pigment complex that contains light-absorbing antenna molecules and a reaction center. Light absorbed by the PSI reaction center energizes an electron that is transferred to ferredoxin (Fd), a molecule that carries electrons to other reaction pathways outside the thylakoid. The reaction center replaces the electron transferred to ferredoxin by accepting an electron from the electron-carrier molecule that moves between the cytochrome complex and PSI.

4. **Under Certain Conditions, Ferredoxin can Carry Electrons to Hydrogenase.** Normally, ferredoxin shuttles electrons to an enzyme that reduces NADP+ to NADPH, an important source of electrons needed to convert CO_2 to carbohydrates in the carbon-fixing reactions. Under anaerobic conditions, hydrogenase can accept electrons from reduced ferredoxin molecules and use them to reduce protons to molecular hydrogen (H_2).

5. **Dissipation of Proton Gradient is Used to Synthesize Adenosine Triphosphate (ATP).** ATP synthase couples the dissipation of the proton gradient generated in step 2 to the synthesis of ATP. Translocation of protons from a region of high concentration (thylakoid space) to a region of low concentration (outside thylakoid) releases energy that can be used to drive the synthesis of ATP from adenosine diphosphate (ADP) and phosphate (P). ATP is a high-energy molecule used to convert CO_2 to carbohydrates in the carbon-fixing reactions.

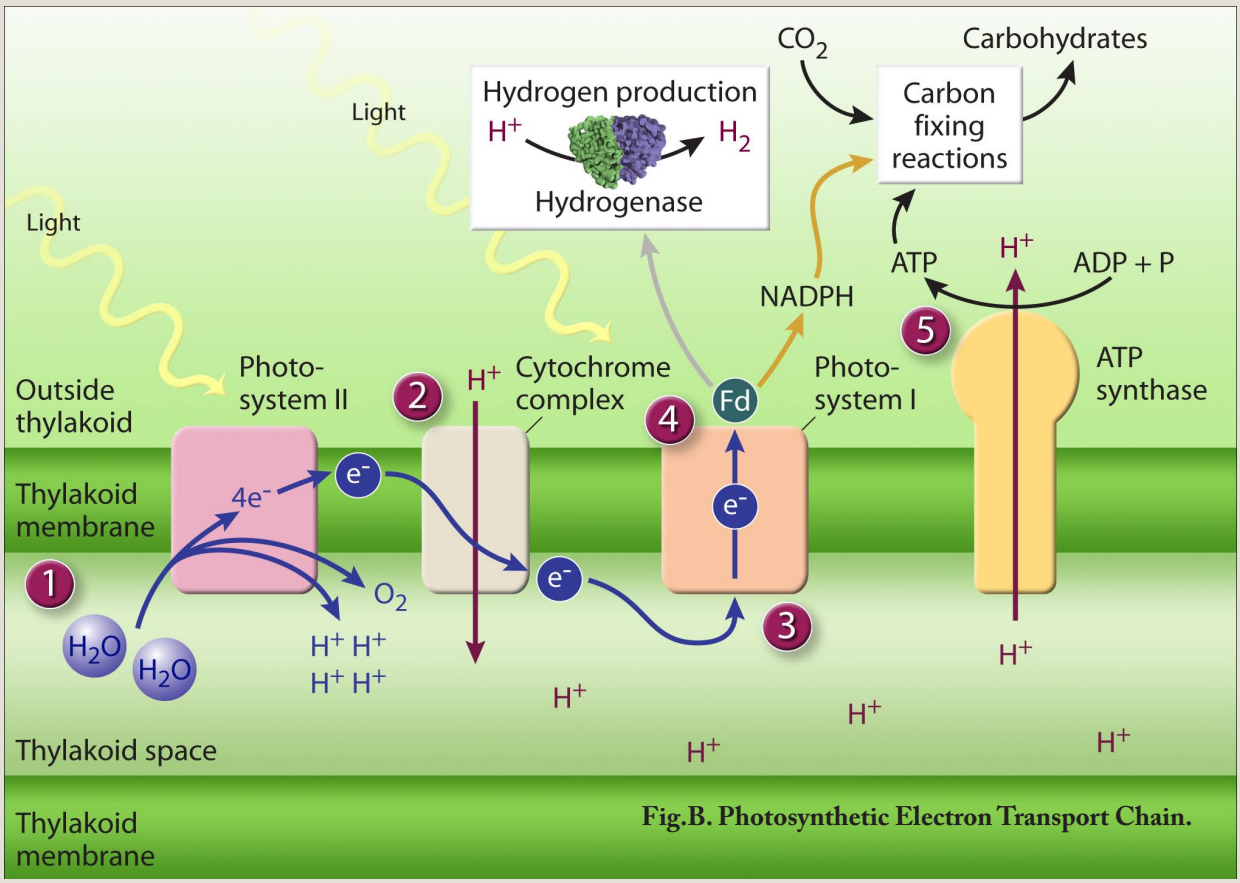


Fig.B. Photosynthetic Electron Transport Chain.

in the laboratory using current procedures. Studying hydrogenase enzymes involved in nonbiophotolytic pathways could provide structural or functional insights to guide the engineering of biophotolytic systems.

- **How are hydrogenases assembled, and how are metals incorporated into the active site?** Two major types of hydrogenases are defined by their biologically unique metallocenters: Nickel-iron (NiFe) and iron only (Fe). NiFe hydrogenases are found in many bacteria and some cyanobacteria. Fe hydrogenases are found in some bacteria and green algae. In green algae, hydrogenases are bidirectional (capable of catalyzing hydrogen oxidation or proton reduction to produce H_2); in cyanobacteria, hydrogenases are either bidirectional or they uptake enzymes. Although turnover is much higher for Fe hydrogenases, NiFe hydrogenases are more oxygen tolerant. The metallocenters of both NiFe and Fe hydrogenases form complexes with such unusual inorganic cofactors as carbon monoxide (CO) or cyanide (CN). Little is known about the assembly of an active hydrogenase, and several genes may be involved in the synthesis of cofactors required for activity. A better understanding of hydrogenase assembly will enable the engineering of enzymes with improved function.
- **How do we overcome the oxygen-sensitivity problem of hydrogenases?** The bidirectional Fe hydrogenases that catalyze the hydrogen-evolution reaction in biophotolytic systems are highly sensitive to oxygen, a product of the water-splitting reaction in the first step of the photosynthetic pathway. Oxygen sensitivity also makes hydrogenase isolation from cells and its subsequent analysis a challenge that will be met by new technologies.
- **What are the potential electron-transfer rate limitations associated with each step of the biophotolytic hydrogen production pathway?** Key factors that can impact the partitioning of electrons between hydrogenase and competing pathways include the buildup of a pH gradient across the photosynthetic membrane and variations in the concentrations of critical electron-transport carriers. Understanding how electron fluxes in an organism are regulated will aid the development of mechanisms for directing more electrons towards proton reduction and hydrogen production.
- **What are the regulatory and metabolic pathways that influence H_2 production?** A thorough examination of hydrogen metabolism in green algae and several different strains of cyanobacteria from diverse habitats will provide new insights into how hydrogen-production pathways are controlled. By understanding how an organism sustains and regulates hydrogen production, we will be able to determine which metabolic pathways contribute, how eliminating hydrogen-consuming reactions affects hydrogen metabolism and other cellular processes, and how organisms can be adapted to increase hydrogen yields.

A.1.5.2.2. Scientific and Technological Capabilities Required to Achieve Goals

Key capabilities needed to address many of the gaps in current understanding of biophotolytic hydrogen production include developing microbial hosts to produce hydrogenase enzymes, screening large numbers of enzymes for desired functionalities, large-scale molecular profiling to provide a global-view of hydrogen production, in vivo visualization of hydrogenase structure and activity, modeling of regulatory and metabolic networks, and metabolic engineering (see Table 5. Roadmap for Development of Biophotolyic Hydrogen Technologies, p. 213, and Table 6. Biophotolytic Hydrogen Production Challenges, Scale, and Complexity, p. 213). Specific needs include the following:

- **Suites of microbial hosts to produce hydrogenases from many different organisms.** Potentially thousands of enzymes from many different organisms will need to be produced and analyzed. Other requirements include methods for producing eukaryotic enzymes in simpler prokaryotic systems, designing host organisms that can provide the intracellular environment required for proper protein assembly and folding, and screening the proteins produced from these host organisms.
- **Methods to produce large numbers of enzymes to screen for desired hydrogenase properties.** With so much variability among natural hydrogenases and engineered variants, developing high-throughput capabilities for producing large numbers (perhaps hundreds of thousands to millions) of enzymes to screen for O_2 tolerance, H_2-production activity, spectroscopic examination, and structural analysis could accelerate the discovery of enzymes best suited for biotechnological applications.

- **Molecular profiling to provide a global view of cellular activity during hydrogen production.** Improvements in computational capabilities and large-scale molecular profiling techniques (transcriptomics, proteomics, metabolomics, measurements of metal abundance) are needed to obtain a global view of microbial hydrogen production. Systems-level analyses could guide experimental investigations by defining gene regulatory networks controlling the expression of genes involved in hydrogen production or cofactor synthesis and identify pathways activated or deactivated during hydrogen production for multiple organisms under varying conditions.
- **Methods to perform in vivo visualization and characterization of molecular machines.** Although crystal structures of some hydrogenases have been determined, this information provides only snapshots of enzyme structure. Advanced techniques for visualizing the different stages of hydrogenase assembly or monitoring hydrogenase activity in living cells will be critical to building predictive models that can be used to engineer hydrogenases optimized for biotechnological applications.
- **Support and techniques for systems-level studies to model and simulate regulatory and metabolic networks.** Studying hydrogenase function within the context of a network maintained by living cells is essential to understanding how this process is influenced by different pathways and environmental conditions.

Table 5. Roadmap for Development of Biophotolytic Hydrogen Technologies

Totality of Processes Must Be Optimized with a Number of Challenges for Each

Processes	Challenges	Deployment
Hydrogenases Regulatory pathways Charge transport Partitioning Multiple mechanisms	O_2 sensitivity Range of hydrogenases Primary and secondary pathways Electron transfer limits Reverse reactions Light capture	Photolytic organisms contained in bioreactors (closed flowing system with hydrogen and oxygen separations) Photosynthetic hydrogen production cassettes deployed in nanostructures
Development Strategy Explore natural range of hydrogenases for variability and design principles Explore mutations and other optimization strategies Understand regulatory and other ancillary pathways for systems optimization (e.g., buildup of protons in cytoplasm, alternative uses of reductants) Capture key functions for cell-free incorporation into nanomembranes		

Table 6. Biophotolytic Hydrogen Production Challenges, Scale, and Complexity

Research and Analytical Challenges	Scale and Complexity
• Database screening for and characterizing of natural variants of hydrogenases and other enzymes and molecular machines in the entire set of pathways that underlie this process • Analysis of modified variants to establish design principles for functional optimization of the overall process including oxygen sensitivity, reverse reactions, transport, light capture, and conversion efficiency • Modeling and simulation of photolytic systems to support systems design and optimization	• Screening of millions of genes, thousands of unique species and functions, and thousands of variants of all enzymes • Production and functional analysis of modified enzymes—potentially thousands of each, hundreds of regulatory processes and interactions • Models at the molecular, cellular, and community levels incorporating signaling, sensing, regulation, metabolism, transport, and other phenomenology and using massive databases in GTL Knowledgebase

Traditional in vitro biochemical methods that study hydrogenase activity one enzyme at a time in the laboratory do not provide sufficient information to understand enzymatic activities in living cells. Tools for monitoring hydrogenase activity in vivo and integrating diverse sets of experiment data are needed to build in silico models of a biophotolytic organism under H_2-producing conditions.

- **Metabolic engineering.** Metabolic engineering involves genetically modifying microorganisms to target and manipulate enzymatic, regulatory, or transport pathways that impact a particular microbial process such as hydrogen production. Models could guide metabolic engineering, for example, by identifying control points for manipulating the flow of electrons to hydrogenase or by predicting how cellular activity and hydrogen yields may be impacted by a variety of conditions. These conditions include the elimination of a particular metabolic pathway or the buildup of a pH gradient across the photosynthetic membrane.

A.1.6. Summary

The broad spectrum of analytical tools in the GTL facilities could be used to rapidly advance our current understanding of fundamental scientific issues that are impeding the development of biofuel production. Both ethanol from biomass and biophotolytic hydrogen represent bioenergy alternatives that could be produced renewably, domestically, and at a scale large enough to reduce our dependence on foreign energy sources. The potential economic and environmental advantages of developing bioenergy options cannot be realized, however, without pursuing answers to key fundamental scientific questions underlying critical R&D breakthroughs. The GTL program seeks to provide answers that can help biotechnology play a more prominent role in our energy future.

Appendix B. DOE Mission: Environmental Remediation

B.1.1. Environmental Remediation Challenge 216

B.1.2. The Role of Microbial Systems in Remediation 217

B.1.2.1. Benefits and Impacts 218

B.1.2.2. Establishing the Link Between Biology and Geochemistry 218

B.1.3. Using Genome Sequences as a Launch Point to Understand Communities 218

B.1.3.1. Modeling Microbial Metabolic Activities 220

B.1.3.2. Merging Metabolic and Field-Scale Models 220

B.1.4. GTL's Vision for Environmental Remediation and Restoration 221

B.1.4.1. Gaps in Scientific Understanding 221

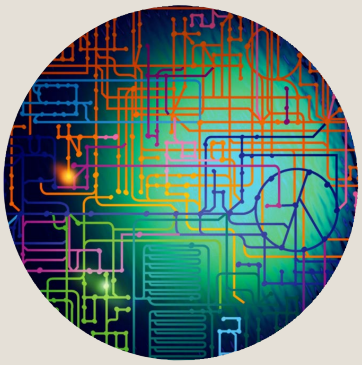
B.1.4.2. Scientific and Technological Capabilities Required to Achieve Milestones 222

B.1.4.2.1. Defining Microbial Communities and Their Potential 223

B.1.4.2.2. Measuring Microbial Processes and Responses 224

B.1.4.2.3. Microbe-Mineral Interactions 225

B.1.4.2.4. Modeling and Simulation Capabilities and Data Management 226



DOE has intractable contamination challenges at diverse sites around the country, making accurate prediction of contaminant behavior critical in determining the need for restoration and in suggesting stabilization or restoration strategies.

Understanding the complex interactions of microbes with contaminants and the subsurface environment—a GTL goal—will allow such predictions to be based on fundamental biological, geochemical, and hydrological properties of specific environments (see Mission Science Goals and Challenges, below right).

DOE Mission: Environmental Remediation

Develop Biological Solutions for Intractable Environmental Problems

B.1.1. Environmental Remediation Challenge

DOE is committed to remediating the large volumes of soil, sediments, and groundwater contaminated with metals, radionuclides, and a variety of organics at diverse defense production facilities and sites across the nation (see sidebar, A Legacy of Hazardous Waste, p. 217).

As an example of the problem's scope, about 5700 individual contaminant plumes, some quite extensive, are known to exist on DOE land (Linking Legacies 1997). Contaminated soils and sediments at the Nevada Test Site and Fernald, for example, are 1.5 and 0.71 million m^3, respectively. One plume at the Savannah River site extends over 7.8 km^2, and a plume of 18 km^2 exists at the Hanford site. In addition, unknown quantities of waste are buried at numerous sites. Without major breakthroughs in technology, projected costs for locating and characterizing contamination, restoring these sites, and disposing of wastes over the next 35 years are $142 billion (Closure Planning Guidance 2004). Although DOE has the goal of completing the remediation of 108 of 114 contaminated sites by 2025 (DOE Strategic Plan 2003), the 6 sites remaining to be addressed are the most challenging, and successful remediation will require development and deployment of innovative methods (see Table 1. Bioremediation: Goals and Impacts, p. 217).

Mission Science Goals and Challenges

Mission Science Goals: 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.

Challenges: Bioremediation will require understanding biogeochemical processes from the fundamental-molecular to community levels to describe contaminant-transformation processes coinciding with simulated changes in microbial-community composition and structure.

B.1.2. The Role of Microbial Systems in Remediation

Microbes found in the contaminated subsurface and other environments often have the metabolic capability to degrade or otherwise transform contaminants of concern to DOE. Currently, DOE environmental restoration is targeting soluble forms of toxic metals and radionuclides in soils and sediments that move through the groundwater. Subsurface microbes, through their interactions with each other and the geochemical environment, play a role in modifying the geochemistry of these subsurface environments, thereby affecting their chemical form and movement. Microbes can directly, or indirectly through their influence on sediment geochemistry, provide a potential cost-effective bioremediation strategy to immobilize contaminants. *Shewanella* and *Geobacter*, for example, are two types of microbes that can enzymatically transform toxic species such as Uranium(VI), which is soluble and moves in groundwater, to Uranium(IV), which is insoluble and precipitates as UO_2 (uraninite) (see sidebars, Microbial Transformation of Toxic Metals, p. 218 and BER Research Advancing the Science of Bioremediation, p. 219).

A Legacy of Hazardous Waste

For more than 50 years, the United States created a vast network of facilities for research and development, manufacture, and testing of nuclear weapons and materials. The result is subsurface contamination on more than 7000 sites at over 100 facilities across the nation, more than half of which contain metals or radionuclides and many with chlorinated hydrocarbons. Biologically based techniques can provide cost-effective restoration strategies for many of these sites.

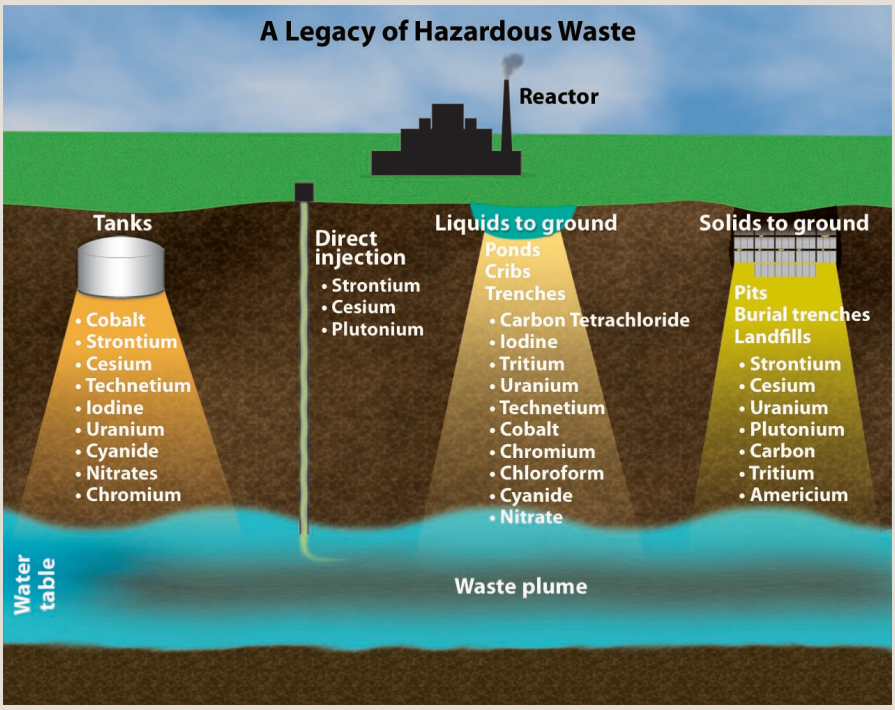


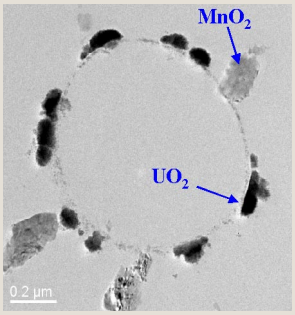
Table 1. Bioremediation: Goals and Impacts

- Understand and incorporate the effects of biological processes into computer models describing the fate and transport of contaminants in the environment. This knowledge could result in billions of dollars of savings by supporting decisions to take advantage of natural attenuation alternatives, use bioremediation for previously intractable problems, or improve the efficiency of conventional technologies.
- Develop new or improved bioremediation strategies and technologies. Potentially billions of dollars could be saved over traditional treatments. Bioremediation may offer solutions in previously intractable cases (i.e., where there was no solution at any price).
- Develop new suites of biosensors and performance assessment and monitoring techniques to track progress of environmental cleanup strategies and optimize operation of current cleanup techniques.

B.1.2.1. Benefits and Impacts

Although comparisons of the cost and effectiveness of metal and radionuclide bioremediation strategies with those of traditional remediation methods are not available, the cost savings for bioremediation of organics are estimated to range from 30 to 90%. In addition, in situ bioremediation, taking advantage of natural microbial populations in the subsurface, has the potential to reduce costs and increase the efficiency of groundwater treatment as compared to conventional pump-and-treat technology. Given that over 1 billion m^3 of water and 55 million m^3 of solid media at DOE sites in 29 states are contaminated with radionuclides (Linking Legacies 1997), potential savings accrued by use of innovative technologies are likely to amount to billions of dollars (Bioventing 1996; Patrinos 2005; Scott 1998).

Microbial Transformation of Toxic Metals



Naturally occurring, metal-reducing bacteria can influence the mobility of radionuclides such as uranium via enzymatic reduction processes. For example, *Shewanella oneidensis* reduces a wide range of organic compounds, metal ions, and radionuclides. Immobile precipitates are shown on the surface of this cross-section of *S. oneidensis*.

Research is needed to provide useful information to decision makers on whether remediation is necessary and practical, give an accurate prediction of contaminant mobility, and suggest bioremediation strategies. A biotreatment technique that works well at one site may perform poorly at another because we lack understanding of the unique interactions—in these "geologically powered dark ecosystems"—between the microbial community and subsurface geochemistry (Nealson 2005). Characterization and monitoring tools must be developed to gain that understanding. At a few sites of specific interest to DOE, less than 1% of the microorganisms have been collected, cultured, and characterized in any great detail, and only a small fraction of those have had their genomes sequenced. Even less is known regarding the interactions of microorganisms in communities. We have only begun to appreciate the existence of such systems, let alone understand them so we can take advantage of their diverse capabilities (Gold 1992; see The Microbial World, p. 13).

B.1.2.2. Establishing the Link Between Biology and Geochemistry

A key to successfully understanding these systems will lie in establishing the link between biology and geochemistry. An exacerbating challenge in establishing a meaningful science base for future applications is the inherent complexity of the subsurface and the distribution of contaminants. Most contaminated sites have extremely heterogeneous geology, hydrology, and resultant geochemistry. Subsurface processes are difficult to measure, control, and monitor, in part because samples from monitoring wells provide only single time-point data from a large three-dimensional area and because monitoring wells can disturb subtle interactions. Contaminants do not flow uniformly from a source point and can be transformed physically, chemically, or biologically to alter their state and mobility. Microbial and geochemical processes can, for example, both immobilize contaminants through redox (oxidation/reduction) processes and enhance transport through complexation. These natural and induced heterogeneities dramatically affect the distribution and makeup of microbial communities, resulting in countless niche environments and communities that must be understood and for which remediation strategies must account.

B.1.3. Using Genome Sequences as a Launch Point to Understand Communities

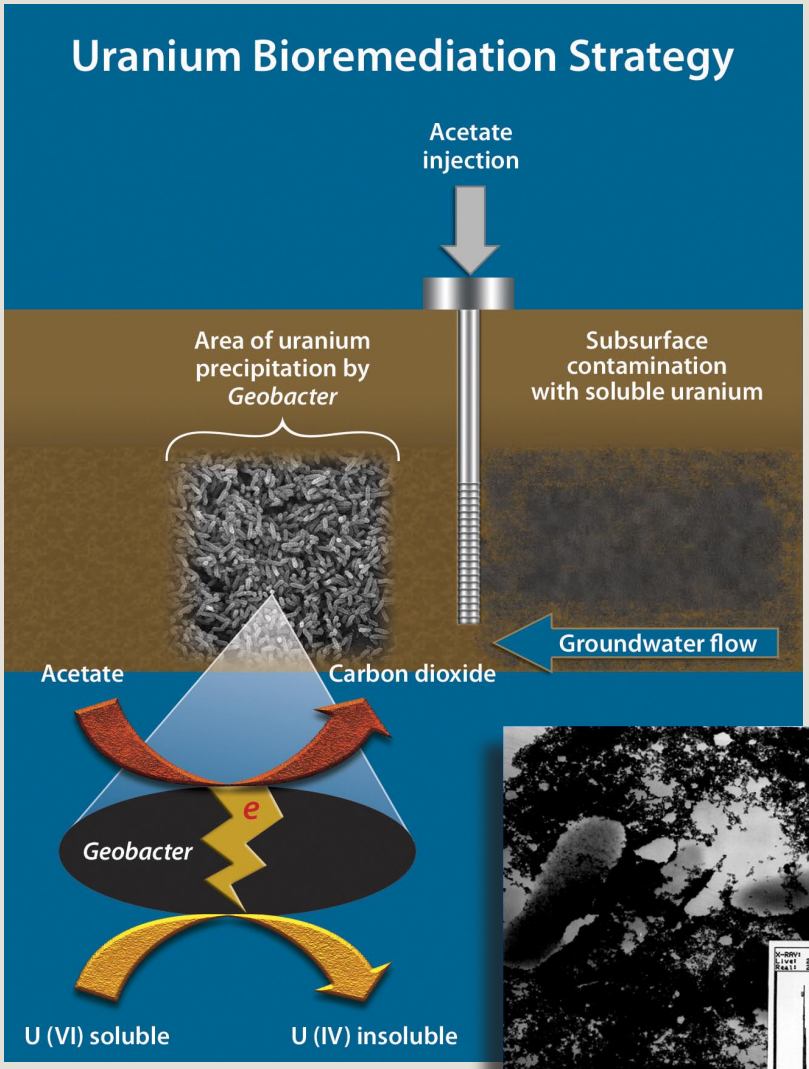
In this complex venue, we first must define the genomic potential of microbial communities. GTL uses genome sequences as a launch point for detailed, mechanistic investigations into microbial metabolism and other cellular subsystems to achieve a systems-level understanding. Having complete genome sequences for microbes known to catalyze important contaminant-transformation reactions provides an unprecedented

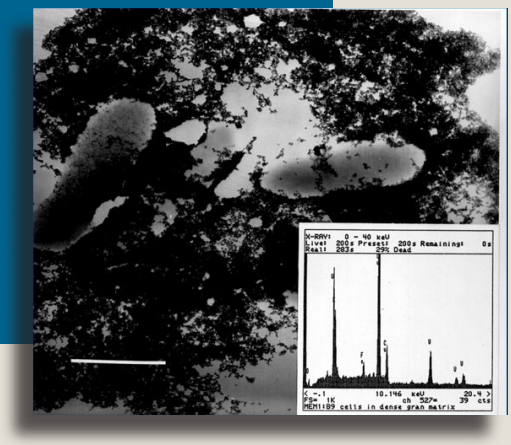
BER Research Advancing the Science of Bioremediation

Genome-Enabled Techniques Contribute to Model Development

Geobacter species can transform uranium from a soluble to an insoluble form, effectively removing it from groundwater and preventing its further mobility. Environmental Remediation Sciences Division (ERSD) investigations demonstrated that this reaction is coupled to growth, indicating that the process is part of cellular respiration. Further exploration into *Geobacter's* metabolic pathways showed that uranium reduction is linked to the oxidation of organic carbon compounds such as acetate. In ERSD programs, adding acetate over a 3-month period to a site where *Geobacter* and uranium were present resulted in the *Geobacter*-mediated precipitation of uranium. After 50 days, the responsible *Geobacter* species became a minor component of the entire microbial community, demonstrating a need to better optimize strategies for long-term bioremediation by these microbes.

The ultimate goal is for GTL environmental-remediation research to develop in silico models that can be used before initiating bioremediation to accurately predict the metabolic behavior of microorganisms involved. Such models will enable evaluation of multiple potential bioremediation strategies before resources and time are committed to field work. GTL and ERSD researchers have made substantial progress toward this goal. An in silico model of *G. sulfurreducens* developed from its genome sequence has accurately predicted its metabolic response to a variety of environmental conditions. With further development, a more generalized in silico model of the *Geobacter* species that predominate during in situ uranium bioremediation will be able to guide optimization of uranium cleanup at a wide range of DOE sites.





D. Lovley, Univ. of Mass., Amherst

Reference

R.T. Anderson et al., "Stimulating the In Situ Activity of *Geobacter* Species to Remove Uranium from the Groundwater of a Uranium-Contaminated Aquifer," *Appl. Environ. Microbiol.* 69, 5584–91 (2003).

opportunity to describe these processes, from molecules through systems. Genome sequences enable investigations of the identities and functions of individual genes that code for important contaminant-transformation reactions, and their expression can be related to field-scale models of transformation processes in the environment—an important advance from earlier, more qualitative descriptions. Whereas historically our studies have been limited to microbes that can be cultured in the laboratory, the combination of metagenomics with the production and characterization of proteins from genes allows new insights into microbial function.

Most current research focuses on understanding single microbial species having potential for environmental remediation and stabilization (e.g., *Shewanella, Desulfovibrio,* and *Geobacter*) in laboratory-based cultivation or field studies. Microbially mediated environmental processes, however, rarely are due to the activity of a single group of organisms—microbes, even those in contaminated environments, typically live in diverse communities. Little is known of the overall dynamics of these communities, and the complexity and spatial structure of energy-transfer reactions that occur across the microbe-mineral interface have only begun to be revealed (see sidebars, BER Research Advancing the Science of Bioremediation, p. 219; *Geobacter*, p. 74; and *Shewanella*, p. 70). Characterization of microbial communities generally has been investigated using single-gene (16S rRNA) sequencing surveys to gain phylogenetic insights. Only a few organisms (notably metal and sulfate reducers) have been sequenced and characterized to any extent. Obtaining detailed links between genome sequence and molecular mechanistic function will require the use of the most robust metagenomic techniques to assess the makeup of communities and their genomic potential.

B.1.3.1. Modeling Microbial Metabolic Activities

Mechanistic models of microbial communities are key requirements in constructing field-scale contaminant fate and transport models that extrapolate over very long time periods. These models must treat such aspects of microbial metabolism as reactions to growth, stress, and nutrient limitation (among many others) that can directly affect gene expression. Understanding changes in all these metabolic activities along the migration pathway or with respect to time requires mathematical models. These models accurately simulate microbial metabolism and are supported by data from biogeochemical and environmental measurements to viably reflect the dynamic interplay of microbes and environment (see sidebar, A Revolutionary Whole-Genome Perspective, p. 221). Results will form the basis for evaluating and modeling pathways of such cellular processes as signaling, regulation, and response to contaminants. Key elements of modeling scenarios include microbe-mineral interactions and resulting molecular structural and charge-transfer responses; microbial-community responses (e.g., signaling, motility, biofilm formation, and other structural responses); and ensuing community functionality.

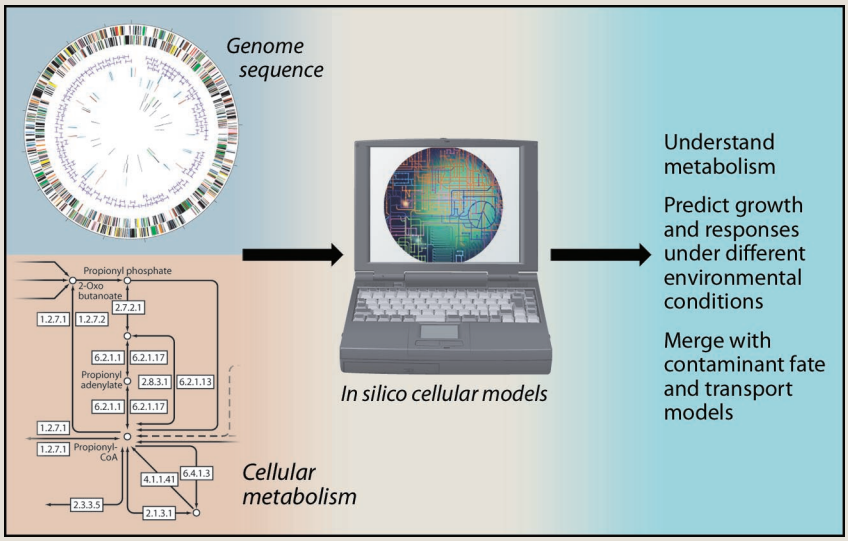
B.1.3.2. Merging Metabolic and Field-Scale Models

Eventually, GTL in silico genome-based microbial-metabolism models must be merged with Environmental Remediation Sciences Division (ERSD) field-scale models of contaminant fate and transport (see sidebar, Microbe-Mineral Interface in Contaminated Environments, p. 222). Some of these techniques already are being generated within environmental-restoration programs and GTL. GTL currently is focusing on quantitatively deciphering the molecular and biochemical pathways of several model microbes that catalyze contaminant-transformation reactions of interest to DOE. Additionally, complementary research within ERSD focuses on understanding the biogeochemical potential of subsurface microbes (see sidebar, Environmental Remediation Sciences Division Activities Complementary to GTL, p. 223). [For more information, see *Bioremediation of Metals and Radionuclides: What it is and How it Works*; 2nd ed., 2003).] To accomplish this linkage, more trained scientists are needed to determine the makeup of subsurface microbial communities and their interactions with the geochemical environment.

Meshing GTL's Approach to ERSD's Challenges

A Revolutionary Whole-Genome Perspective

Genomic information on cellular metabolism can be incorporated as physiological modules into computer (in silico) models to better understand metabolism, predict cell responses under different environmental conditions, and improve contaminant fate and transport models. Microbes already sequenced by DOE and under intense analysis in GTL have been detected in Environmental Remediation Sciences Division (ERSD) studies of in situ immobilization techniques in the subsurface of uranium-contaminated aquifers. These organisms are closely related to the GTL organisms on which existing physiological modules are based. As additional sequenced organisms become available, similar in silico models could be developed to more accurately model multispecies phenomena. These include syntrophic relationships, anaerobic degradation consortia, and shifts in the dominant terminal electron accepting process (called TEAP) observed in sediments. Although progress is being made, many challenges remain in placing fundamental physiological knowledge in the context of the dynamic flow and transport regimes characteristic of DOE sites.



B.1.4. GTL's Vision for Environmental Remediation and Restoration

GTL science will facilitate detailed, large-scale discovery and investigation of microbes and microbial ecosystems with important contaminant-transformation capabilities. These studies will expand knowledge about structure, function, metabolic activity, and the dynamic nature of microbial communities and their interaction with the geochemical environment. The information will aid in the prediction of microbe-mediated contaminant fate and transport by providing more reliable, science-based information upon which to base remediation decisions. Capabilities and information established by GTL in conjunction with ERSD programs also will enable the integration of predictive microbiology with remediation science, leading to more effective and reliable applications.

B.1.4.1. Gaps in Scientific Understanding

Details underlying successful field-scale models ultimately focus on dynamic microbe and microbial-system geochemical interactions and functionality but have their foundations in the molecular interactions and processes that GTL seeks to understand. The following outlines, in broad terms, the science required over the next 15 years to deliver an integrated physiology and genomics-based understanding of microbial metabolism to support the detailed modeling of microbially catalyzed contaminant transformation in the environment. Key questions that we must be able to answer include the following (see also 5.4.1.1. Probing Mixed Microbial Populations and Communities, p. 175):

- What is the makeup of microbial communities? We need to learn who is there, their physiological states, their individual contributions, how they relate to each other and the environment, and how metagenomic DNA sequence can be used to predict the function, behavior, and evolutionary trajectory of microbial communities.
- How do microbes identify, access, and modify their local geochemical environments to gain energy and nutrients and meet other metabolic requirements?
- What physicochemical environmental interactions control the dynamic makeup, structure, and function of microbial communities in the subsurface, and what is the resultant impact on contaminant transformation?
- How and why do contaminants impact microbial communities, and what are potential indicators of these impacts?
- How do molecular mechanistic processes in microbes and communities relate to macroscopic behaviors in field environments?

B.1.4.2. Scientific and Technological Capabilities Required to Achieve Milestones

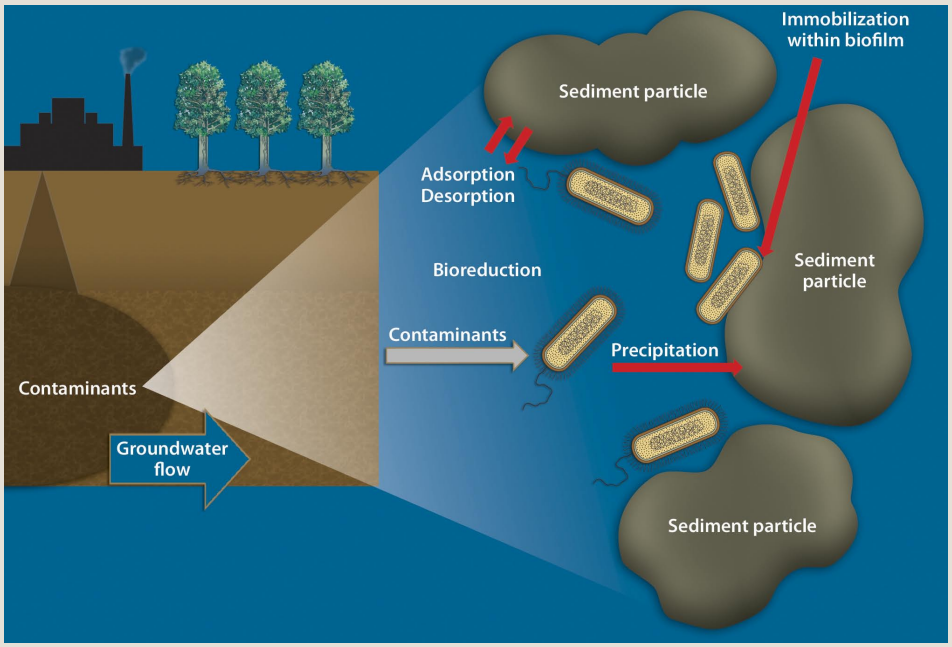
Key science and technology milestones for GTL and partner programs over the next 15 years are discussed in sections 1.4.2.1–1.4.2.4. Table 2. Bioremediation Challenges, Scale, and Complexity, p. 223, lists some research and analytical challenges with the scale and complexity of their scope.

Mission Challenge of Environmental Remediation Sciences Division

Microbe-Mineral Interface in Contaminated Environments

Biogeochemical processes driven by interactions at the microbe-mineral interface in soils and sediments influence contaminant behavior. These reactions occur at the level of the individual cell or groups of cells, forming an architecture assembled against the mineral surface. The cells interact with each other and the mineral surface, creating a dynamic, microscale domain that controls the kinetics of biogeochemically mediated reactions.

Before large-scale restoration strategies can be implemented, such processes must be better understood to predict contaminant transport in situ under natural and induced conditions. ("Induced conditions" refers to the addition of nutrients, oxygen, or other electron donors and acceptors to increase microbial activity.) Achieving this level of knowledge requires new tools to characterize and resolve fundamental phenomena at the molecular, cellular, and community levels. Microbial information must be integrated with the subsurface strata's geochemical characterization and the system's hydrological properties at the sediment mineral and pore scales.



B.1.4.2.1. Defining Microbial Communities and Their Potential

- **Metagenomic methods tailored to unique subsurface communities and environments to assess their general and specific makeup.** These studies will identify biochemical potential where possible and determine unknown genes through protein production and characterization and **other methods. Improving high-throughput** sequencing is allowing recovery of the genetic potential of single-strain microbes and whole communities (Venter et al. 2004; Tyson et al. 2004).
- **Refinement of the growing genome sequence databases and bioinformatics tools to identify and analyze genes found in environmental organisms.** New and faster methods of genome annotation ("super annotation") are required to capture fully the genetic potential of sequenced organisms.
- **Modeling and experimental capabilities to explore the physiology of full systems.** Current genomic descriptions of microbial metabolism for sequenced species are based on genes and the proteins they encode of known function or inferred from similar gene sequences found in other organisms such as *E. coli*. Genes of unknown function, which constitute a significant portion of most sequenced genomes, cannot be modeled. We must be able to produce and characterize these unknown proteins on demand. The sheer number of novel gene sequences to be investigated dwarfs currently available investigative techniques (such as knockout mutant characterization).

Environmental Remediation Sciences Division Activities Complementary to GTL

The Environmental Remediation Sciences Division (ERSD) of DOE's Office of Science seeks to understand microbial function in diverse environments and how these functions can be harnessed for restoration of contaminated DOE sites. Field-scale models for predicting contaminant fate and transport and designing remedial measures in complex heterogeneous environments depend on understanding biogeochemical reactions occurring in the subsurface at much smaller scales. This knowledge must be both spatially and temporally extrapolated. ERSD is positioned to identify microorganisms and processes by taking advantage of the genomic and proteomic systems biology tools offered by GTL. The goal is to use microbial capabilities for improving our understanding of the complex processes operating in the subsurface, placing them in the context of other simultaneous chemical and physical processes, and scaling the results to the field using advanced conceptual and mathematical models.

ERSD funds numerous laboratory- and field-based projects to evaluate the potential for subsurface microorganisms to immobilize or remobilize contaminant metals (including radionuclides) in situ. Significant progress has been made in detecting subsurface microorganisms associated with this process and in describing and modeling biogeochemical reactions mediated by microbes. A more complete understanding of microbial metabolism and community behavior will help determine the impact of microorganisms on contaminant fate and transport.

Table 2. Bioremediation Challenges, Scale, and Complexity

Research and Analytical Challenges	Scale and Complexity
• Analysis of microbial communities and their metabolic activities that impact the fate and transport of contaminants • Analysis of geochemical changes in subsurface environments due to microbial or chemical activity • Accurate conceptual and quantitative models for coupling and scaling microbial processes to complex heterogeneous environments	• Hundreds of different sites, millions of genes, thousands of unique species and functions • Functional analysis of potentially thousands of enzymes involved in microbe-mineral interactions; hundreds of regulatory processes and interactions; spatially resolved community formation, structure, and function; influence on contaminant fate • Models at the molecular, cellular, and community levels incorporating signaling, sensing, metabolism, transport, biofilm, cell-mineral interactions; incorporated into macromodels for fate and transport

- **Genome annotation to include functional genomics information such as cell response to environmental stimuli using functional gene arrays or expressed protein analyses.** These results will develop the basis for identifying and modeling biochemical pathways and regulatory networks within cells, including growth, stress, and metabolic responses to potential contaminants and other environmental factors.
- **Additional genome sequence generated for microbes and microbial communities and increased computing power to enable more intricate comparative genomics studies.** Also, more extensive protein structural analysis, protein networks analysis, and fold-recognition applications will be possible.
- **Method refinements to monitor growth and activities of microbial communities within the subsurface.** GTL will develop and standardize technologies for extraction of mRNA and protein from environmental samples. Information provided by these analyses forms crucial links between results obtained from current GTL research and the microbial processes as they occur in the environment.
- **Improved cultivation methods such as the microdroplet technique (Zengler et al. 2002; Keller and Zengler 2004) and others to capture a greater proportion of microbes associated with environmental samples.** These capabilities will enable us to study samples in the laboratory and compare them to the more thoroughly studied model organisms.

B.1.4.2.2. Measuring Microbial Processes and Responses

Breakthrough capabilities and technologies are needed to support measuring and modeling of microbially mediated contaminant-transformation processes and microbe-mineral interactions. These tools should permit a more complete mechanistic understanding of microbially mediated processes within the environment and provide solutions for microbiological and geochemical scaling issues needed for field-scale descriptions. Capabilities will be developed for modeling and measuring communities and single cells, both in isolation and within communities, to understand microbial systems. Examples include the following:

- **New, sensitive methods to measure the proteome, metabolome, and transcriptome of populations and communities of organisms.** Such measurements will involve significant developments in the area of high-throughput gene sequencing, protein identification and production, multivariable cultivation techniques, controlled cultivation and physiological analyses, and enzymatic analysis.
- **Methods to determine the biochemical basis for intracellular and intercellular interactions that contribute to the functionality of biofilms and other structured communities.**
- **Multifunctional imaging techniques to monitor biological, chemical, and physical changes simultaneously within environmental samples.**
- **Methods to investigate subsurface biogeochemical processes.** These require further development because most remain focused on either the geochemical or biological aspects of subsurface processes, to better integrate mechanistic aspects of both. As examples,
  - Several microspectroscopic techniques afford detailed analyses of mineral and biological composition and structure at high resolution and small scales and can be further developed for use with live biological analyses.
  - Kinetic data gathered from nutrient-enriched cultures grown in the lab often do not reflect processes observed in the field. Computer models of contaminant fate and transport within the environment must be developed to include robust biogeochemical modules. New techniques and capabilities are needed to apply laboratory data to complex field environments.
- **Techniques to examine and identify the composition of natural organic matter for exploring the metabolism of naturally derived substrates in environmental samples.** Only a subset of this material currently can be identified.

- **Data to characterize natural communities, including the following:**
  - Microbe-microbe interactions such as cell signaling, materials and energy transfer, gene transfer, and syntrophy to study phenomena occurring in biofilms. They remain poorly understood at a mechanistic level.
  - Functional microbial-community analysis to understand microbial positioning, including the potential for biofilm and other structured community development; their structure and relationship to function; and the molecular basis for microbial motility, competition, and niche exploitation. Little is known about the molecular basis for changes in community structure resulting from shifts in environmental conditions such as the introduction of contaminants. Detailed information about the impact of environmental changes on microbial-community structure leads to understanding of the community's functional stability and the net metabolic flux of electron donors and acceptors, including contaminant transformation. This knowledge will lead to better, more environmentally relevant descriptions of microbe-mediated processes.

B.1.4.2.3. Microbe-Mineral Interactions

Techniques will be developed for evaluating electron transfer and growth at mineral surfaces. Many microbes of interest to DOE respire metals (i.e., iron and manganese) that typically exist in solid-phase minerals. These organisms oxidize organic compounds and hydrogen and ultimately pass these electrons onto mineral surfaces, capturing energy for growth during the process. A crucial component of investigation is the ability to image and quantify processes occurring at the microbe-mineral interface, including the following:

- **The molecular basis for changes in cell-membrane structure to understand attachment, electron transfer, and mineral chemistry at the microbe-mineral interface.** More detailed analyses at the microbe-mineral interface will include molecular-level analyses of cell-membrane composition and dynamics and the interactions between membrane proteins and potential contaminants and mineral phases that can serve as electron acceptors.
- **Methods to measure microbial-community and environmental chemical and physical structures using methods that combine, for example, nuclear magnetic resonance (NMR) and optical imaging technologies.**
- **Further development of tomographic and spectromicroscopic techniques to provide information about chemical changes occurring on and, to a certain extent, within mineral surfaces, as well as about metals associated with microbial cells.**
- **Application of established and novel microscopies in combination with new sample-preparation and spectroscopic techniques to observe bacteria attaching to and growing on mineral surfaces in a noninvasive, real-time fashion.** These new techniques, in addition to infrared and micro-NMR imaging, will be needed to evaluate microbe-mineral associations under noninvasive conditions and in real time, a significant challenge.
- **Imaging abilities coupled with probes designed to evaluate metabolic processes as they occur within cells and communities of cells at a mineral surface in real time.** Coupled visual measurements of biological and geochemical parameters at the cellular scale will aid development of temporal and structural descriptions of microbial communities at mineral interfaces.
- **New suites of biosensors or probes to measure a variety of metabolic and geochemical processes down to the microbe-mineral scale.** These tools must be specific enough to relate biochemical changes to geochemical processes at high spatial resolution and extremely small scale. The sensors, along with required advances in cultivation techniques, will permit the investigation of microbially mediated processes, not only under controlled laboratory conditions but also within the environment. Once these processes can be measured and evaluated, new methods of kinetics analysis will be required for accurate testing and evaluation.

B.1.4.2.4. Modeling and Simulation Capabilities and Data Management

- **Novel analysis techniques to mate the results of GTL in silico models of microbial metabolism with field-scale contaminant fate and transport models created within DOE ERSD.** This is part of a classic problem in relating the temporal, spatial, and process scales of molecular mechanisms to the scales of macrosystems such as contaminant plumes that function over years and kilometers.
- **Computational techniques to correlate spatial information with geochemical properties and reactions at meaningful scales.**
- **Improved modeling techniques to enable incorporation of many more variables, particularly when attempting to simulate molecular processes at larger scales or under a variety of environmental conditions.** Current modeling capabilities and analysis tools are geared to mathematical representations that are too simple or too small (<10-state variables).
- **Significant improvements in data management and interpretation to use and distribute the enormous amounts of data generated by genome sequencing and associated systems biology.** Current bioinformatics techniques are limited primarily to sequence and structure analyses of genetic information. The vast amount of biological and environmental data presents equally pressing challenges for storing and managing access to these very large data sets. Organization is key, and centralization is envisioned in which input of and access to data can be standardized to common QA/QC protocols and routines. Metadata definitions and databases must be formalized and managed in a similar fashion. (For a more detailed discussion, see 4.0. Creating an Integrated Computational Environment for Biology, p. 81.)
- **Significant computer power and efficient computational methods to support visualization techniques, particularly important for protein structure and 3D simulation of processes in both the laboratory and environment.**

Appendix C. DOE Mission: Carbon Cycling and Sequestration

C.1.1. The Climate Change Challenge 228

C.1.2. The Role of Microbes 229

C.1.3. Microbial Ocean Communities 231

C.1.3.1. Photosynthetic Capabilities 231

C.1.3.2. Strategies for Increasing Ocean CO_2 Pools 231

C.1.3.3. GTL's Vision for Ocean Systems 233

C.1.3.3.1. Gaps in Scientific Understanding 233

C.1.3.3.2. Scientific and Technological Capabilities Required 233

C.1.4. Terrestrial Microbial Communities 235

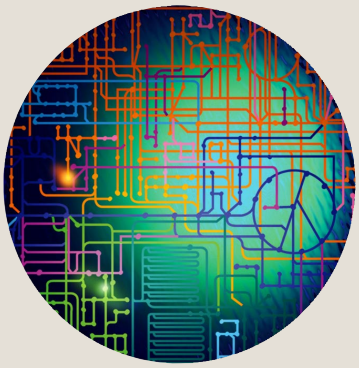
C.1.4.1. Influence on Plant Growth 235

C.1.4.2. Strategies for Increasing Stable Carbon Inventories 236

C.1.4.3. Terrestrial Systems Vision 236

C.1.4.3.1. Gaps in Scientific Understanding 236

C.1.4. 3.2. Scientific and Technological Capabilities Required 238



DOE Mission: Carbon Cycling and Sequestration

Understand Biosystems' Climate Impacts and Assess Sequestration Strategies

The United States is committed to understanding the factors that influence climate change, reducing uncertainties in assessments of climate change, and developing strategies to mitigate change. Microbes in the earth's oceans and soils play a major role in the cycling of carbon and other elements. GTL seeks to understand this role to be able to predict the impacts of climate change on microbes and their responses to the resulting ecosystem shifts. This knowledge also will provide the basis for developing and assessing strategies for ocean- and soil-based carbon sequestration.

C.1.1. The Climate Change Challenge

Atmospheric greenhouse gas (GHG) concentrations have been increasing for about 2 centuries, mostly as a result of human (anthropogenic) activities, and now are higher than they have been for over 400,000 years. As shown in Fig. 1. Simplified Representation of the Global Carbon Cycle, p. 229, about 6 billion tons (gigatons) of carbon are released into the air by human activity each year, three-quarters from the burning of fossil fuels and the rest from deforestation and other changes in land use, with a small amount from cement production. Although the effects of increased levels of CO_2 on global climate are uncertain, many agree that a doubling of atmospheric CO_2 concentrations, predicted for the middle of this century by the Intergovernmental Panel on Climate Change (IPCC), could have a variety of serious environmental consequences.

Global climate change is a long-term energy and environmental challenge requiring major investments in targeted research and development (see Mission Science Goals and Challenges below). Gaining a greater knowledge of how carbon cycles through ecosystems is a critical element of the national strategy to understand climate and potential changes that might occur due to anthropogenic greenhouse gases and to develop solutions to reduce future increases in CO_2 (the most important

Mission Science Goals and Challenges

Mission Science Goals: 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.

Challenges: We are just beginning to understand the genetic and functional diversity of ocean and terrestrial ecosystems. They potentially contain millions of microbial species organized in extensive communities. We must understand both the global and molecular mechanistic behaviors of these large systems.

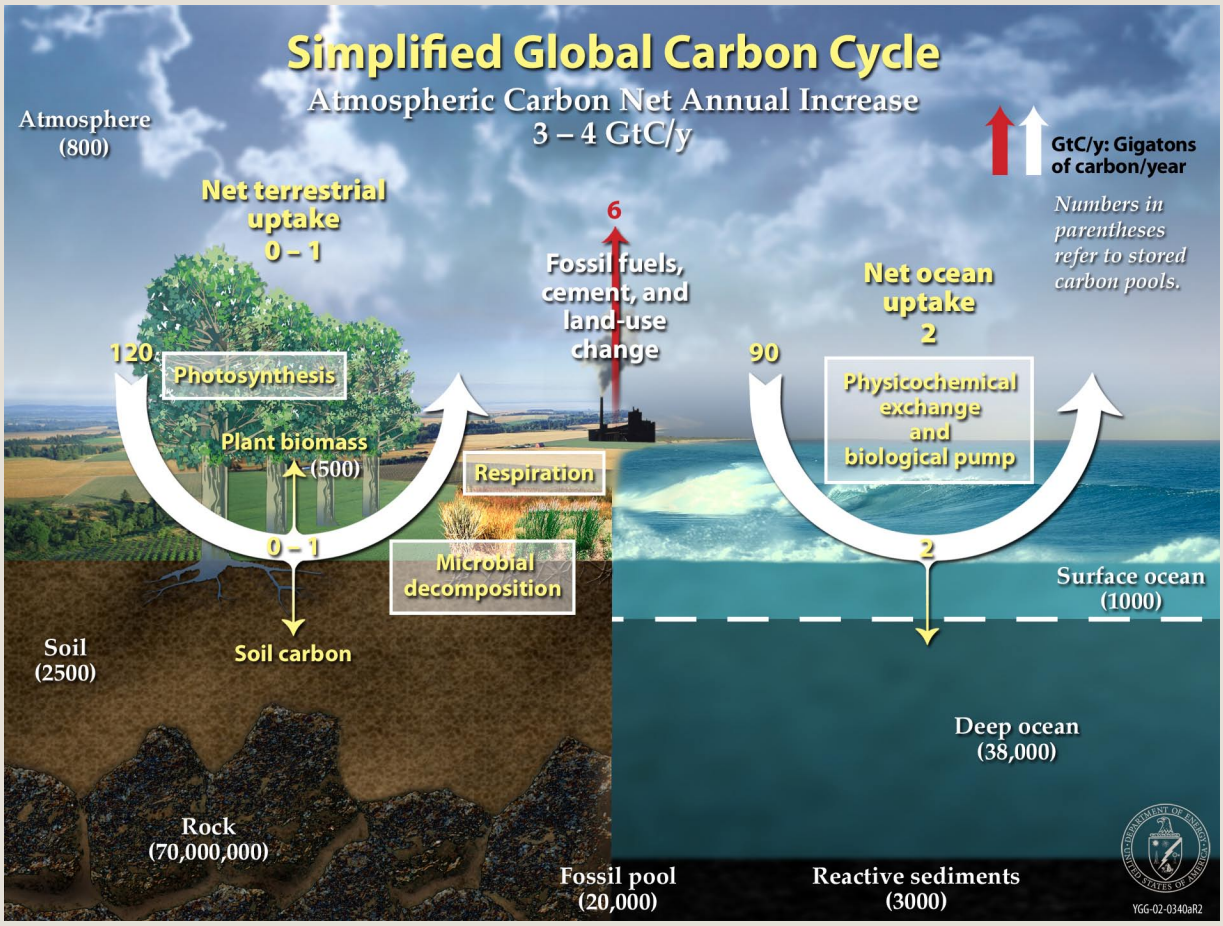


Fig. 1. Simplified Representation of the Global Carbon Cycle. The illustration depicts human-induced changes relative to the total cycle. [Graphic adapted from *Carbon Sequestration Research and Development* (1999).]

GHG) and other GHGs (see sidebar, CCSP Calls for Carbon Cycle Data, this page). Understanding how climate affects both natural and managed "pools" (e.g., forest, agriculture lands) of carbon stored in global ecosystems and how these carbon "sinks" influence atmospheric concentrations of CO_2 will be important in reducing uncertainty in climate models and in understanding the long-term sequestration capacity of those pools (Carbon Sequestration 1999).

C.1.2. The Role of Microbes

Natural processes also contribute to the storage and cycling of carbon (Fig. 1, this page). The stability and sequestration of the vast pools stored in oceanic and terrestrial environments depend, in part, on the microbial world.

CCSP Calls for Carbon Cycle Data

The U.S. Climate Change Science Program (CCSP) is a multiagency effort to understand the earth's climate and predict how it will evolve under various greenhouse gas scenarios. The CCSP calls for development of information on the carbon cycle to assist in evaluation of carbon-sequestration strategies and alternative response options, the understanding of key "feedbacks" including biological and ecological systems, improved knowledge about the sensitivity of ecosystems to climate variability and change, and incorporation of such knowledge into climate models. Understanding how carbon dioxide and other by-products of energy generation affect the global environment requires research into how carbon (and, to a lesser extent, nitrogen, phosphorous, oxygen, and iron) cycle through ecosystems (see Appendix F. Strategic Planning for CCSP and CCTP, p. 249).

APPENDIX C

According to the American Society of Microbiology (King et al. 2001), "Microbes, responsible for transforming many of earth's most abundant compounds, cannot be ignored in the search for scientific solutions to adverse global changes. . . . Both the ubiquity of microbes and the delicacy of environmental balances contribute to [the planet's] sensitivity to disturbances in the microbial world."

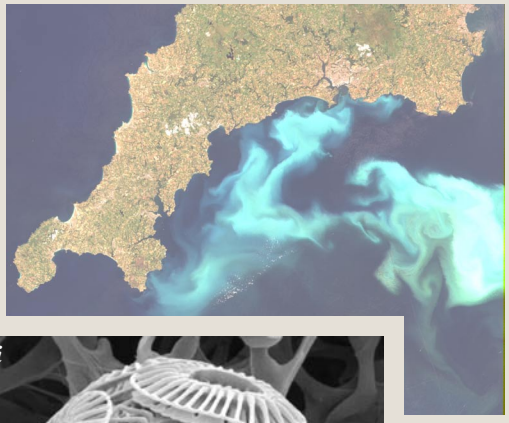
Microbial energy transfer and material processing in the biosphere have been transforming the earth for over 3 billion years (see sidebar, Planet-Transforming Microbes Cycling Carbon, this page) and influencing climate on a global scale (Staley et al. 1997). Extremely diverse ocean and terrestrial microbial communities serve fundamentally different roles in the carbon cycle as primary photosynthetic producers of biomass in the ocean biological "pump" and as carbon and nutrient managers and decomposers in terrestrial systems (see Table 9. Microbial Community Characteristics in Diverse Earth Environments, p. 39). Microbes cycle immense volumes of carbon in the process of recycling most of earth's biomass: They can fix CO_2 by light-driven (photoautotrophy) and geochemically driven (lithoautotrophy) reactions, generate methane, produce CO_2 as they decompose organic matter, precipitate carbonate minerals, and catalyze the polymerization of plant polymers into recalcitrant pools of carbon in soil.

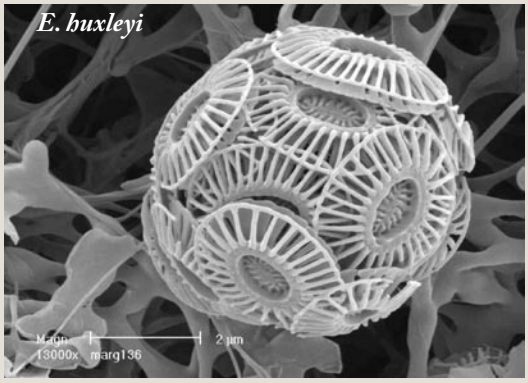
The DOE mission of global carbon management requires that we achieve a comprehensive understanding of terrestrial and marine microbial communities so we can learn the role that these communities play in carbon sequestration. We then must find ways to enhance their capabilities to develop microbe-based strategies for capturing and sequestering atmospheric CO_2 and to assess the potential effectiveness and adverse ecological impacts of proposed carbon-sequestration technologies. Microbial systems also have great potential as sensitive indicators of environmental change.

Natural cycles of carbon in the environment involve exchanges many times greater than anthropogenic emissions (see Fig. 1, p. 229). While anthropogenic emissions threaten to change the globe's climate gradually, secondary effects on natural ecosystems and disturbance of their much larger atmospheric exchanges could result in even larger shifts. Knowing the effects of anthropogenic emissions on natural cycles is important as part of the complete picture of carbon management on a global scale.

Planet-Transforming Microbes Cycle Carbon

Blooms of *Emiliania huxleyi*, captured by satellite, are shown just off the coast of the United Kingdom. Though microscopic, these carbon-cycling cocolithophores (bottom image) are present in such large numbers that they are visible from space—an indicator of their pervasiveness and thus influence on ocean ecosystems (top image). Their shells are made of calcium carbonate, and over the ages their deposits have created the "White Cliffs of Dover" on the southern coast of England. Understanding the planet-transforming capabilities of these and other ocean microbes—that is, how they affect ocean ecosystems by cycling carbon and other important elements—is a focus of the GTL program. *E. huxleyi*'s genome is being sequenced by DOE's Joint Genome Institute. For more information, see www.noc.soton.ac.uk/soes/staff/tt/.





Top – Landsat 7 photo of July 24, 1999, by S. Groom, Plymouth Marine Laboratory Remote Sensing Group; bottom – J. Young, Natural History Museum, London

Investigating and understanding these ecosystems require probing numerous complementary functionalities in thousands of species and millions of genes, involving hundreds of thousands of proteins. In brief, goals and challenges in this mission follow (see Table 1. Carbon Cycling and Sequestration: Goals and Impacts, this page).

C.1.3. Microbial Ocean Communities

C.1.3.1. Photosynthetic Capabilities

Microbial communities living near the surface layers of oceans are the primary photosynthetic organisms driving the biological pump. Absorbing CO_2 and sunlight to produce most oceanic organic materials, the organisms make up the foundation of the marine food chain. Photosynthesis of such phytoplankton as diatoms, dinoflagellates, and cyanobacteria converts about as much atmospheric carbon to organic carbon in the ocean as plant photosynthesis does on land. Large variations in phytoplankton abundance, therefore, can greatly impact the oceans' ability to take up atmospheric carbon.

Oceans currently have a net absorption of about 2 Gt of carbon per year, offsetting about 30% of carbon emitted to the atmosphere by the burning of fossil fuels. Understanding the interactions and dynamics underlying this natural CO_2 sink is necessary to explain past shifts in global climate and to predict future environmental changes. Microbes drive these processes by converting atmospheric CO_2 into organic matter, some of which remains in the oceans (see sidebar, Carbon Cycling in the Oceans, p. 232).

Dominant organisms in surface waters include such cyanobacteria as *Synechococcus* species and *Prochlorococcus marinus*, which capture CO_2 and light to carry out photosynthesis. *Prochlorococci* now are thought to be the most abundant photosynthetic organisms on earth. Eukaryotic diatoms such as the recently sequenced *Thalassiosira pseudonana* also live in surface waters and convert CO_2 and other nutrients into hard silicates. This process carries organically complexed carbon to ocean depths, thus converting its relatively rapid cycling in surface waters (where it is returned to the atmosphere) to a considerably slower one in ocean sediments.

C.1.3.2. Strategies for Increasing Ocean CO_2 Pools

Ocean carbon-sequestration strategies aim to increase the deep ocean inventory of CO_2. Two approaches typically are considered: Direct injection of a CO_2 stream into the ocean interior depths and iron fertilization to enhance photosynthesis by phytoplankton in the biological pump and thus increase carbon uptake. The potential effectiveness and adverse impacts must be evaluated for both approaches. According to the CCTP strategic plan, due for publication in 2005: "A research portfolio is required that seeks to determine, via experimentation and computer simulations, the ability of the world's oceans to effectively store anthropogenic CO_2 without negative environmental consequences" [CCTP, www.climatetechnology.gov].

DOE has sponsored genomic sequencing of several of these organisms (see sidebar, Microbial Genomes Yielding Clues to Global Climate Change, p. 233). Additionally, recent GTL-sponsored metagenomic approaches have involved researchers sequencing DNA fragments isolated from samples taken from ocean (and terrestrial) environments. These studies have for the first time allowed direct insights into the makeup and functionality of these natural systems, revealing an amazing diversity. Analyses from Sargasso Sea samples, for example, turned up more than a million previously unknown genes, including almost 800 rhodopsins (the

Table 1. Carbon Cycling and Sequestration: Goals and Impacts

- Improved understanding of key feedbacks and sensitivities of biological and ecological systems and accelerated incorporation into climate models will reduce uncertainties in assessments of climate change.
- Knowledge of the carbon cycle will allow evaluation of carbon-sequestration strategies and alternative response options.
- Development of sensors and monitoring techniques and protocols will allow use of these sensitive ecosystems as sentinels for the effects of climate change.

Carbon Cycling in the Oceans: Solubility and Biological Pumps

Ocean processes regulate the uptake, storage, and release of CO_2 to the atmosphere. The total exchange of carbon between atmosphere and oceans is controlled by two principal processes: The solubility (or physical) pump and the biological pump.

The solubility pump is driven by physical processes. The solubility of CO_2 in water increases with lower water temperature, and the colder water sinks. This gradient, from lower CO_2 concentration near the surface to higher concentrations below about 500 m, helps draw CO_2 from the atmosphere into the oceans. The solubility pump, in combination with ocean circulation, results in net CO_2 emissions at the equator and net CO_2 drawdown at high latitudes. Changes to ocean circulation or stratification due to increased global warming from increased greenhouse gases are predicted to result in decreased ocean uptake of CO_2 by the ocean solubility pump.

The biological pump, whose activities are just being revealed, refers to the composite of biological processes occurring in ocean surface layers. These begin with the microbial photosynthesis of CO_2 into organic matter (much like land plants) and end with either the conversion of organic matter to CO_2 at different depths or with the deposition of a small fraction of organic material into sediments on the ocean floor. The biological pump's efficiency is a function not only of carbon fixation but also of the depth at which the organic carbon is remineralized to CO_2. Current models, which rely on incomplete carbon-cycle models having little biological input, suggest that if the biological pump were turned off today, atmospheric levels of CO_2 would rise to 680 ppm (~400 ppm higher than preindustrial levels and about 300 ppm higher than current levels). The ocean's future activity as a carbon sink is uncertain, however, because of potential (and currently uncharacterized) feedbacks among global climate change, ocean circulation, and microbial communities that actively cycle carbon. These natural ocean carbon-sequestration processes extend beyond carbon to affect organic and inorganic pools of nitrogen, phosphorus, oxygen, and many other chemical species.

A key feature of ocean environments is the extremely slow recycling of mineral nutrients. Dead organisms from the photosynthetically active top of the water column sink into its depths and, ultimately, the ocean floor. They carry with them essential nutrients, mainly nitrogen and phosphorus, that are liberated in the darkness of the deep ocean. From there, upwelling currents take several thousand years to return the nutrients to warm surface waters. Consequently, primary production in the top of the water column is limited severely by the lack of mineral nutrients, whereas the nutrient-rich deep waters lack light energy for primary photosynthetic production.

light-absorbing antennae of microbes essential for photosynthesis) (Venter et al. 2004). Metagenomic studies of soil samples show an even greater amount of genetic material (Riesenfeld, Schloss, and Handelsman 2004). Results suggest that microbial communities have an extraordinarily wide range of mechanisms and pathways that could offer new applications to meet DOE carbon-management missions.

GTL will use these data as starting points for explorations into molecular processes underlying microbial photosynthesis. The goal is to ascertain fundamental principles of photosynthetic systems' molecular design. These principles will reveal the dynamics of carbon-assimilation pathways and those that degrade organic matter and ultimately either sequester carbon or return it to the atmosphere.

GTL research will enable us to begin identifying critical organisms and their capabilities and responses to stress. These data will provide the foundation for developing biological rate constants that can be incorporated into detailed models of carbon cycling. When these models are extended to the global ecosystem, the potential impact of carbon-cycle perturbations on climate-change models can be assessed. Ultimately, this knowledge will guide decisions about acceptable levels of change and hence acceptable atmospheric levels of GHGs.

In addition to elucidating carbon-cycling nuances, such detailed biological data can lead to the development of increasingly sophisticated microsensors that can detect changes in the levels of biomolecules (DNA, RNA, proteins, metabolites) and serve as indicators of microbial-community response to environmental stressors (see sidebar, Ocean Monitors, p. 234).

C.1.3.3. GTL's Vision for Ocean Systems

The GTL Knowledgebase ultimately will provide in silico models of microbial systems in oceans, with supporting data and experimental capabilities that can be used to inform policies and develop methods and applications relevant to DOE missions in carbon management.

C.1.3.3.1. Gaps in Scientific Understanding

Understanding carbon cycling and sequestration requires knowledge about the underlying mechanisms controlling microbiological systems. Investigations will include defining key players and their roles; determining how systems change as a function of climate, CO_2, nutrients, and biogeochemical cycles; and enabling predictions of atmospheric CO_2 and climate impacts on marine communities over time. Specifically, these analyses will enable us to begin exploring the following types of questions:

- What happens to carbon in the oceans, and how is it portioned among various life forms? How does this portioning vary in rate as a function of location, depth, salinity, nutrient availability, temperature, proximity to population centers and coastlines, currents, and seasons?
- How far do carbon and carbon dioxide migrate from their "points of entry" into the ocean, and what impacts their travel and processing?
- What are the elements of the biological pump?
- What happens to growth rates of phytoplankton as a function of carbon entry into the oceans in light of the variables noted above?
- What would happen to carbon absorption if growth rates for phytoplankton were altered either up or down?
- What are the dynamic community structures of ocean microbes, and how do they impact carbon processing?
- How reversible would be the effects of actions that we might take to alter ocean carbon sequestration (and on what time scales)?

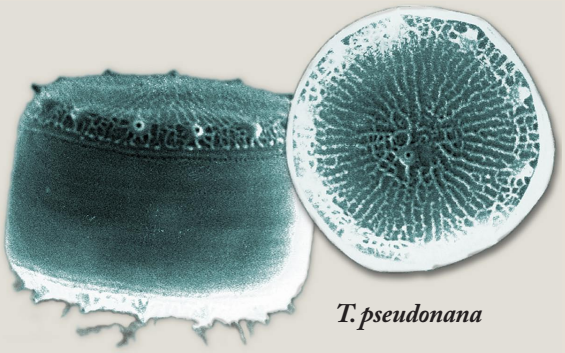
C.1.3.3.2. Scientific and Technological Capabilities Required

Defining communities and their genomic potential will require capabilities for rapid and accurate sequencing of single cells, key organisms, and environmental communities. Also needed are methods to perform comparative genomic analyses and capabilities for gene synthesis and manipulation. Specific needs include the following:

- Metagenomic approaches to aid in sifting through millions of genes and determining which proteins are produced by ocean communities and when.

Microbial Genomes Yielding Clues to Global Climate Change

Analyses of the first ocean microbes to be sequenced—a diatom and several cyanobacteria—are beginning to help investigators understand the physiological and genetic controls of photosynthesis and the cycling of carbon and nitrogen. The diatom *T. pseudonana* (images below) and species of *Prochlorococcus* and *Synechoccus* contribute to absorbing amounts of CO_2 comparable to all the world's tropical rain forests combined. GTL research on the molecular processes underlying the capabilities of these organisms can lead to more-accurate climate models and strategies for carbon sequestration. The diatom and three of the four cyanobacteria in these analyses were sequenced at the Joint Genome Institute and funded by DOE (see also Falciatore and Bowler 2002). [*Science* 306, 79–86 (2004); *Proc. Natl. Acad. Sci. USA* 100, 10020–25 (2003); *Nature* 424, 1037–42 and 1042–47 (2003)]



T. pseudonana

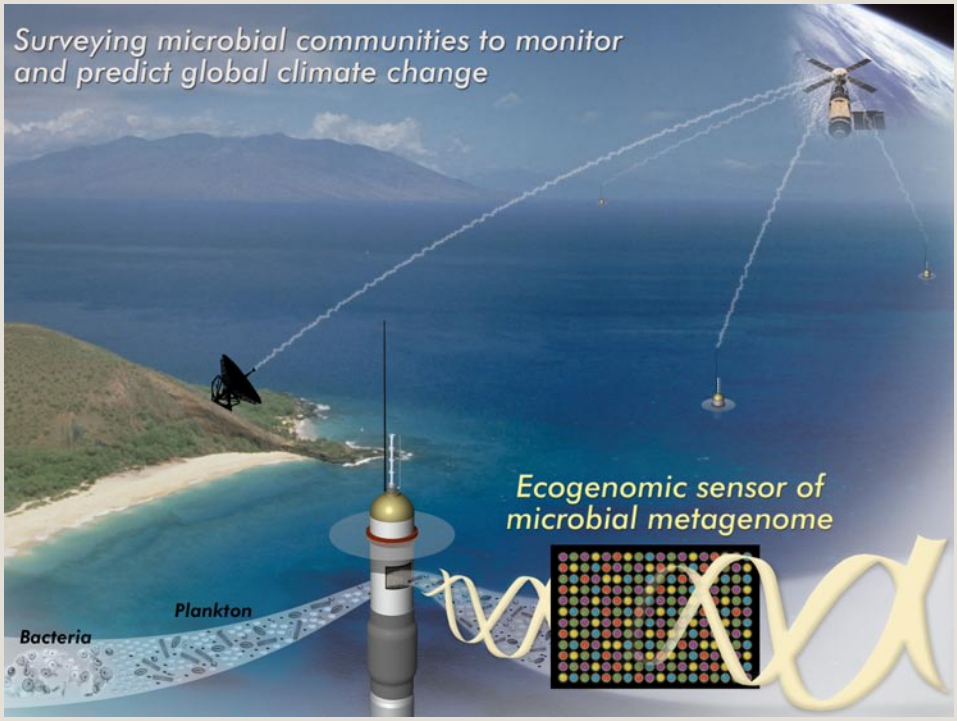
B. Palenik and D. Lee, Scripps Institution of Oceanography

- Capacity to make and study the proteins determined by the ocean's metagenome to understand ocean microbial functionality and processes. Because these microbes are essentially unculturable, protein analysis initially will be achieved only by synthesis directly from genome sequence. A high-throughput approach would permit simultaneous, highly parallel production and characterization tests on hundreds of thousands of proteins.
- Molecular tags (or affinity reagents) for proteins with established critical roles to use as probes for determining the structure and function of natural ocean ecosystems.
- New sampling and analysis tools to investigate the natural dynamics of relationships among microbial, biogeochemical, and physical processes.
- Technologies to measure environmental responses, including ecogenomic sensors of sentinel organisms; biochemical assays of cells, communities, and ecosystems; and environmental assays (see sidebar, Ocean Monitors, this page).
- Detailed studies of proteins, multimolecular machines, and metabolites to aid in understanding key microbial responses in terms of photosynthesis, transporters, and biomineralization processes; development of functional assays and technologies, including imaging, to measure system responses.
- Information on microbial mechanistic behaviors (cellular, community, ecosystem) for incorporation into more accurate climate models.
- Database of genes, pathways, microbes, and communities to explore the structure and function of ocean ecosystems, and, in particular, the roles of ocean microbes in carbon processing and their impact on global climate processes.

Ocean Monitors: Nanoscale Ecogenomic Sensors

Nanoscale environmental genomic sensors may one day be used to monitor microbial populations and their interactions with environmental processes, including those affected by climate change. The real-time approach envisioned by DOE for the National Oceanographic Partnership Program merges information from genome research programs with nanotechnologies and smart sensors.

The knowledge gained will enhance understanding of the genetic diversity and functions of microbial communities and help answer key questions about their influence on ocean and terrestrial biogeochemical cycles. Microbial sentinels of ecosystem changes may forewarn the approach of such events as red tide caused by an increase in *Pfisteria* species. (For more information on microbial sensing, see Klaper and Thomas 2004 and Belkin 2003.)



C.1.4. Terrestrial Microbial Communities

C.1.4.1. Influence on Plant Growth

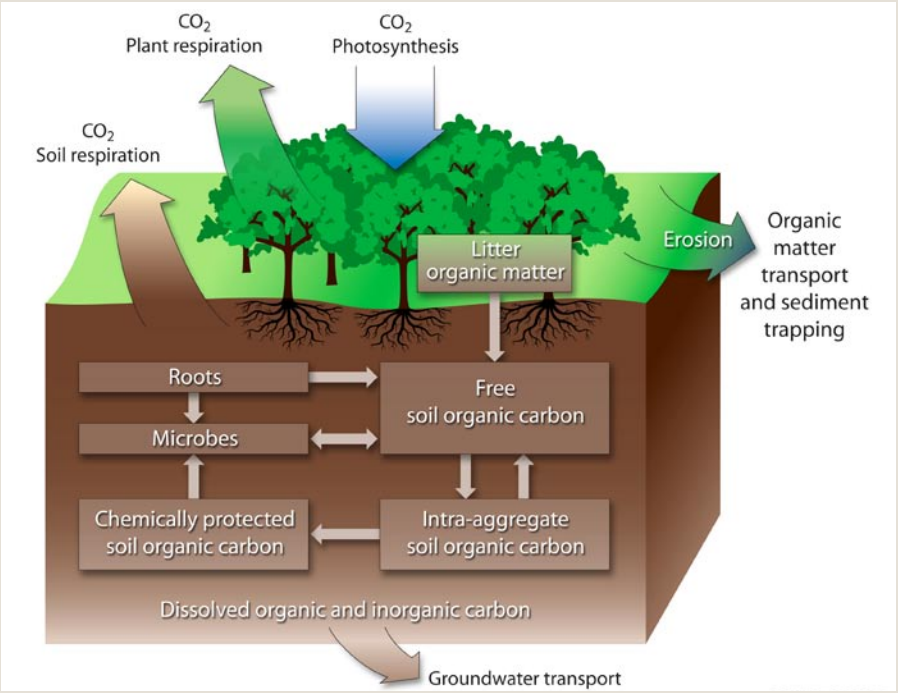
Terrestrial ecosystems absorb CO_2 directly from the atmosphere, mainly via plant photosynthesis. The carbon is stored in plant biomass and soil organic matter or respired back to the atmosphere. Terrestrial ecosystems can help reduce concentrations of CO_2 in the atmosphere by increasing carbon stores in biomass, soils, and wood products (see sidebar, Carbon Transformation and Transport in Soil, this page).

Some microbial populations influence carbon storage in plants by enhancing their growth through interactions with organic compounds around the root (the rhizosphere), by providing nutrients such as phosphorous and nitrogen, or by suppressing plant pathogens in the soil. Other microbial communities exert neutral or even harmful effects. A better understanding is needed of the molecular mechanisms that enable microbes to colonize root surfaces, interact with organic compounds in the rhizosphere, and cooperate with other organisms.

Microbes impact carbon storage in soils by transforming carbon in decaying plants into other forms of organic matter, with varying degrees of recalcitrance. Soils thus are a complex mixture of compounds having different residence times, with more-stable compounds being the most important for carbon sequestration because their turnover times can be hundreds to thousands of years. Soils contain about 75% of the carbon in the terrestrial ecosystem, and knowing more about the microbial processes taking place there will lead to a better understanding of long-term carbon storage in soils.

Carbon Transformation and Transport in Soil

These processes can result in sequestration of carbon in the soil as organic matter or in groundwater as dissolved carbonates, increased emissions of CO_2 to the atmosphere, or export of carbon in various forms into aquatic systems. [Source: *The U.S. Climate Change Science Program: Vision for the Program and Highlights of the Scientific Strategic Plan*, 2003, www.climatescience.gov/Library/stratplan2003/vision/ccsp-vision.pdf]



Carbon dioxide is emitted from soils through soil respiration, a result of the metabolic activity of plant roots and soil microbes decomposing plant material and soil organic matter. Most plant material entering the soil is respired relatively quickly as CO_2; a small fraction becomes humus, which remains in soils for a longer time. Soil respiration is a major component of the global carbon cycle, returning nearly 10 times as much CO_2 to the atmosphere as emissions from fossil-fuel combustion (Rosenberg, Metting, and Izaurralde 2004). The shift in the ability of microbes to respire carbon to the atmosphere during environmental stresses such as climate change (e.g., more carbon is released by decomposition when

stress causes plants to die) is a serious complicating factor in determining the permanency of these pools for sequestration. Physical influences such as agricultural tillage practices and fire contribute greatly to the amount of carbon released to the atmosphere from soils. As we understand microbial species and specific processes that create recalcitrant forms of carbon and those that metabolize carbon rapidly to carbon dioxide, we can manage terrestrial ecosystems in better ways, including low-till and no-till agriculture.

C.1.4.2. Strategies for Increasing Stable Carbon Inventories

Gaining a fundamental understanding of biological mechanisms of carbon cycling and sequestration in an ecological context can help us understand and predict effects of climatic change on key ecological processes. Genomics and, even more so, proteomics and metabolomics will become valuable tools for developing a biological systems understanding and reducing uncertainty about effects of future (potential) climate changes on the terrestrial biosphere's structure and function. They also will be useful for increasing the likelihood of successful human responses to such climate-change contributors as carbon sequestration and ecosystem management (see sidebar, Integrated Assessment Program, this page).

Augmenting natural microbial activities may be a promising option to optimize the inventories of stable carbon forms. DOE has sponsored successful field experiments that remove uranium from contaminated groundwater by stimulating the growth of particular microbial communities known to precipitate (and immobilize) that contaminant. We also can envision potentially altering some plants (notably cellulosic energy crops needed to produce bioethanol) in ways that stimulate them to produce larger fractions of more-recalcitrant organic matter that would lead to increased carbon sequestration in the terrestrial biosphere (see sidebar, Poplar Tree Offers Potential for Greater Carbon Storage, p. 237). An added benefit could be improved soil quality because of increased carbon. Natural carbon fluxes are large, so even small forced changes resulting from sequestration strategies can be very significant.

C.1.4.3. Terrestrial Systems Vision

GTL science will generate the knowledge to incorporate, for the first time, models describing the global ecosystem into climate models to provide foundations for a more robust science base for policy and engineering. It also will enable evaluation of potential biology-based strategies for terrestrial carbon sequestration. The national goal is to develop these policies and strategies substantially over the coming decade (see Table 2. Carbon Cycling and Sequestration Challenges, Scale, and Complexity, p. 237).

C.1.4.3.1. Gaps in Scientific Understanding

Understanding the global ecosystem and its climatic effects requires learning about key microbial processes involved in carbon and nitrogen cycling, maintaining soil fertility, and increasing soil carbon content. Understanding how microbes and their ecosystems respond to a variety of environmental factors will allow for more accurate assessments and predictions of carbon inventories in terrestrial systems and their impacts on climate change to enable more-effective strategies to manage these inventories.

Integrated Assessment Program

DOE's Integrated Assessment (IA) of Climate Change Research Program combines simplified representations of the entire global climate system, emphasizing greenhouse gas (GHG) emissions and actions that would affect emissions. Integrated models are used to assess the value of technologies that result from research; these include new biotechnological approaches, for example, that improve biomass conversion and generate hydrogen, an expected result of the knowledgebase generated in the GTL program. IA research provides a foundation for subsequent policy analysis or decision making.

The IA Program addresses several of DOE's research priorities. The most immediate issue of interest is the role of energy production and associated GHG emissions in global climate change. Closely related is the question of climate change impacts—by extension, energy production impacts—on humans and their environment. The program thus brings together a major approach (integrated assessment), a global environmental problem (climate change), and the scientific challenges of interdisciplinary and modeling of multiscale complex systems.

Carbon Cycling and Sequestration

As part of a broader science base for understanding effects of climatic change on terrestrial ecosystems, GTL systems biology will support studies on interactions among terrestrial ecosystems and on changes in atmospheric composition and the climate system. In particular, advanced hardware and software capable of rapidly sequencing genomes will provide the foundation for performing systems biology analyses and quantifying climate effects on key protein functions to understand the following:

- How do microbes contribute to carbon transformation in soils, and what is their potential for sequestering meaningful amounts of carbon (gigatons per year) in more stable forms? This knowledge will provide decision makers, including the public, with information on designing and evaluating options for responses to potential climatic effects of future carbon-based energy production.
- How do microbial genomes adjust mechanistically to climate change? This understanding will allow more realistic prediction of future climate-change effects (or explain effects of recent climate change) on the structure and functioning of ecosystems.
- What is the genomic-mechanistic basis for biological feedbacks to the climatic system brought about through the terrestrial carbon cycle? The potential exists for significant releases of CO_2 or CH_4 to the atmosphere in response to rising temperatures and changes in precipitation.
- With a "simple" understanding of the underlying biology of ecosystems, how can we develop a modeling framework to put systems biology information into a usable context for predicting feedbacks to climate and atmospheric CO_2?

Poplar Tree Offers Potential for Greater Carbon Storage

An international team including the DOE Joint Genome Institute recently sequenced the genome of the black cottonwood or poplar tree (*Populus*). This research could be used to improve tree breeding and forest management practices that would enable significant quantities of carbon to be sequestered by this and, eventually, other trees. In addition, a significant fraction of carbon associated with a stand of trees is in soil organic-matter pools rather than in aboveground biomass or living roots. The poplar genome sequence information might be used to develop ways to enhance both the production and translocation of organic compounds from leaves and shoots to roots and soil, where it might lead to long-term storage of carbon. In addition to carbon storage, poplar produces products and services of considerable value to humans and many ecosystems. Moreover, poplar trees are highly productive in many environments and have a wide ecological range or distribution.

Table 2. Carbon Cycling and Sequestration Challenges, Scale, and Complexity

Research and Analytical Challenges	Scale and Complexity
• Analysis of ocean and terrestrial microbial-community makeup and genomic potential • Analysis of carbon and other cycling processes » Photosynthesis and respiration in oceans » Storage and decomposition in soil: microbial, fungal, and plant communities • Modeling and simulation of microbe biogeochemical systems	• Thousands of samples from different sites, consisting of millions of genes, thousands of unique species and functions • Functional analysis of enzymes involved—potentially tens of thousands; hundreds of regulatory processes and interactions; spatially resolved community formation, structure, and function • Models at the molecular, cellular, and community levels incorporating signaling, sensing, metabolism, transport, biofilm, and other phenomenology into macroecosystem models

C.1.4. 3.2. Scientific and Technological Capabilities Required

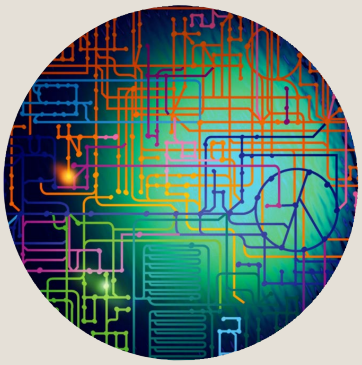
Defining communities and their collective genetic functional potential requires both single-cell and community sequencing (in situ and in vitro), systems biology studies, and the ability to relate microbial activities to soil processes. Capabilities to accomplish these goals include:

- Methods to understand processes by which carbon is transformed into long-lived forms and to design technical and management strategies for enhancing advantageous processes and mitigating negative responses.
- Methods to measure biomolecular inventories correlated with environmental conditions; characterizations of microbial-system interactions with soils, rhizosphere, and plants; and imaging of microbial functional activities (e.g., proteomes and metabolomes) at cellular and community levels—all to understand processes that impact production of GHGs (CO_2, methane, nitrous oxide, and dimethyl sulfide).
- Methods to detect and measure microbial responses to manipulation of plant inputs to the carbon cycle, to human inputs to soils, and to other environmental changes.
- Methods to use microbes as sentinels of climate-induced change in the environment. Research will determine the biomarkers that correlate with specific environmental parameters. Biomarker signatures include combinations of RNAs, proteins, metabolites, and signaling elements; community genomic makeup brought about by population shifts; and functional assays (Tringe et al. 2005).

Appendix D. GTL Meetings, Workshops, and Participating Institutions

GTL Meetings and Workshops 240

Participating Institutions 242



GTL Meetings and Workshops

2000–2004 Totals

Participants:	1486
Unique participants:	792
Meetings:	46
Institutions:	235

2005

- June 19–22 Biotech Industry Organization (BIO) GTL Symposium, Philadelphia, Pennsylvania
- June 13–14 Plant Genomics for Biofuels, DOE-BP Joint Meeting; Washington, D.C.
- Feb. 6–9 Contractor-Grantee Workshop III; Washington, D.C.*

2004

- June 14–16 Roadmap Planning Phase II: Technology Deep Dive; Arlington, Virginia
- June 6–9 Biotech Industry Organization (BIO) GTL Symposium; San Francisco, California
- June 4–6 AAM Colloquium on Systems Microbiology; Portland, Oregon
- April 21–23 World Congress on Industrial Biotechnology GTL Workshop and Symposium, Orlando, Florida
- March 3–4 Planning Study I, Program Science and Capability Needs for DOE Missions; Washington, D.C.
- Feb 29–March 2 Contractor-Grantee Workshop II; Washington, D.C.*

2003

- Sept. 10–11 GTL and Beyond: Data Standards Workshop; Berkeley, California
- Sept. 3 Genomes to Life Milestones Workshop; Crystal City, Virginia
- July 22–24 Data Management, Protein Folding, and Modeling and Simulation Workshops; Gaithersburg, Maryland*
- June 17–18 Characterization and Imaging of Molecular Machines Facility Workshop; Atlanta, Georgia*

*Workshop reports followed by an asterisk are on the web: www.doegenomestolife.org/pubs.shtml

- June 2–3 Facility User Interactions Workshop (Portals); Gaithersburg, Maryland
- May 29–30 Protein Production and Characterization Workshop; Argonne, Illinois*
- May 12–14 Bioinformatics Workshop for Proteomics; La Jolla, California*
- April 1–2 Global Proteomics Workshop; Santa Fe, New Mexico*
- Feb. 9–12 Contractor-Grantee Workshop I; Arlington, Virginia*
- Feb. 3 Scientific Workshop on Affinity Reagent Needs for Facility I; Chicago, Illinois

2002

- Dec. 3–4 Biological and Environmental Advisory Committee (BERAC); Washington, D.C.
- Oct. 14–15 Facilities Planning; Gaithersburg, Maryland
- Aug. 16–17 Facilities Planning; Chicago, Illinois
- June 19–22 Facilities Planning; San Francisco, California
- April 16–19 Computing Strategies; Oak Ridge, Tennessee
- April 16–18 Imaging Workshop; Charlotte, North Carolina*
- April 2 Keck Institute Meeting; Claremont, California
- March 18–19 Mathematics Workshop; Gaithersburg, Maryland*
- March 6–7 Computer Science Workshop; Gaithersburg, Maryland*
- Jan. 22–23 Computational Infrastructure Workshop; Gaithersburg, Maryland*

2001

- Dec. 10–11 Technology Assessment for Mass Spectrometry; Washington, D.C.*
- Nov. 27 BERAC Presentation and Discussion of GTL Payoffs; Washington, D.C.
- Oct. 24–25 Energy and Climate Mission Payoffs; Chicago, Illinois
- Sept. 9–10 Science Mission Payoffs; Washington, D.C.
- Sept. 6– 7 Visions for Computational and Systems Biology Workshop; Washington, D.C.*
- Aug. 7–8 Computational Biology Workshop; Germantown, Maryland*
- June 23 Role of Biotechnology in Mitigating Greenhouse Gas Concentrations; Arlington, Virginia*
- Jan. 25–27 Genomes to Life Roadmap Planning; Germantown, Maryland

2000

- Dec. 1 Preliminary Report to BERAC on Roadmap; Washington, D.C.
- Nov. 30 Roadmap Drafting; Oak Ridge, Tennessee
- Nov. 14–15 Microbial Cell Project Workshop; Chicago, Illinois
- Nov. 7 Roadmap Strategy Meeting; San Diego, California
- Oct. 29–Nov. 1 Roadmap Planning; San Diego, California
- Oct. 6 Roadmapping Meeting; Denver, Colorado
- June 1 Subcommittee Report Approved by BERAC; Washington, D.C.
- March 2 BERAC Subcommittee; Washington, D.C.
- Jan. 1 BERAC Subcommittee; Washington, D.C.

APPENDIX D

Participating Institutions*

Abbott Laboratories
Advanced Life Sciences
Affymetrix
Agilix Corporation
Alfred P. Sloan Foundation
American Association for the Advancement of Science
American Type Culture Collection
Ames Laboratory
ApoCom Genomics
Applied Biosystems
Argonne National Laboratory
Arizona Court of Appeals
Arizona State University
Athenix
Baylor College of Medicine
Beckman Research Institute, City of Hope
Bell Labs
BIATECH
Bio-Technical Resources
Boston University
Brookhaven National Laboratory
California Institute of Technology
Carnegie Mellon University
Celera Genomics
CNRS UMR Ecologie Microbienne
Colin Gordon and Associates
Columbia University
Conkling Fiskum and McCormick, Inc.
Cornell College of Veterinary Medicine
Cornell University
Courant Institute
Dalhousie University
Dana-Farber Cancer Institute
Defense Advanced Research Projects Agency
Diversa Corporation
Duke University
DuPont Central Research and Development
Ehime University
Energy Sciences Network
Environmental Protection Agency
Fellowship for Interpretation of Genomes
Flad and Associates
Florida State University
Food and Drug Administration
Fred Hutchinson Cancer Center
Gene Logic, Inc.
Gene Network Sciences
General Electric
geneticXchange
GeneXPress
Genomatica, Inc.
GENOSCOPE
Georgetown University
Georgia Institute of Technology
GlaxoSmithKline
Harvard Medical School
Harvard University
HDR Inc.
Hebrew University
Howard Hughes Medical Institute
IBM Corporation
Indiana University
Inovise Medical
InPharmix Incorporated
Institute for Biological Energy Alternatives
Institute for Systems Biology
Institute for Genomic Research, The
Integrated Genomics, Inc.
J. Craig Venter Science Foundation

*2000–2004

Jackson Laboratory
Johns Hopkins University
Joint Genome Institute
Keck Graduate Institute
Keio University
Kosan Biosciences Incorporated
Lawrence Berkeley National Laboratory
Lawrence Livermore National Laboratory
Linus Pauling Institute
Los Alamos National Laboratory
Louisiana State University
Marine Biological Laboratory
Marshfield Medical Research Foundation
Massachusetts Institute of Technology
McGill University
Medical University of South Carolina
Merck Research Laboratories
Metrigen, Inc.
Michigan State University
Microsoft Corporation
MITRE Corporation
Molecular Sciences Institute
Monsanto Company
Montana State University
Monterey Bay Aquarium Research Institute
MP Biomedicals, Inc.
National Academies
National Academy of Sciences
National Cancer Institute
National Cancer Institute, Center for Cancer Research
National Center for Genome Research
National Center for Supercomputing Applications, Keck Genome Center
National Energy Research Scientific Computing Center
National Heart, Lung, and Blood Institute
National Institute of Environmental Health Sciences
National Institute of General Medical Sciences
National Institute of Standards and Technology
National Institutes of Health
National Renewable Energy Laboratory
National Science Foundation
National Water Research Institute (Canada)
Natural Resources Defense Council
Naval Surface Warfare Center, Dahlgren Division
NeoGenesis
New England Complex Systems Institute
New York University
NimbleGen Systems Inc.
North Carolina State University
Northern Arizona University
Northwestern University
Novagen
Novartis Research Foundation
Novation Biosciences
Oak Ridge Institute for Science and Education
Oak Ridge National Laboratory
Office of Management and Budget
Office of Naval Research
Ohio State University
Old Dominion University
Oregon Health Sciences University
Oregon State University
Oxford GlycoSciences
Pacific Northwest National Laboratory
Pennsylvania State University
Perkins and Will
Pfizer Inc.
Pittsburgh Supercomputing Center
PolyLC Inc.
Protometrix (Invitrogen)
Purdue University
Quantum Intelligenece
RIKEN
Roche
Rockefeller University
Rutgers University
Salk Institute
San Diego Supercomputer Center
Sandia National Laboratories
Sanger Centre
Science
Scripps Institution of Oceanography
Scripps Research Institute
SmithKline Beecham Pharmaceutical
Software Technology Group, Inc.

APPENDIX D

SoundVision Productions
Southwest Parallel Software
SRI International
St. Jude Children's Research Hospital
Stanford University
Stanford University School of Medicine
State University of New York, Stony Brook
Stony Brook University
Structural Genomics Consortium
Teranode Corporation
Texas Tech University
Thomas Jefferson National Accelerator Facility
U. S. Department of Agriculture
U. S. Department of Energy
U. S. House of Representatives, Science Committee
U. S. Senate, Energy and Natural Resources Committee
Uniformed Services University of the Health Sciences
University of Buffalo
University of California, Berkeley
University of California, Davis
University of California, Los Angeles
University of California, San Diego
University of California, San Francisco
University of California, Santa Barbara
University of California, Santa Cruz
University of Chicago
University of Cincinnati College of Medicine
University of Colorado
University of Connecticut
University of Connecticut Health Center
University of Delaware
University of Florida
University of Georgia
University of Illinois
University of Illinois, Chicago
University of Illinois, Urbana-Champaign
University of Iowa
University of Maryland
University of Maryland Biotech Institute
University of Massachusetts
University of Massachusetts, Amherst
University of Miami
University of Minnesota
University of Missouri, Columbia
University of North Carolina
University of Notre Dame
University of Pennsylvania
University of Pittsburgh
University of Southern California
University of Southern Mississippi
University of Tennessee
University of Tennessee Health Science Center
University of Tennessee, Knoxville
University of Texas
University of Texas, Austin
University of Texas, Houston
University of Texas Medical School
University of Utah
University of Virginia
University of Washington
University of Wisconsin, Madison
University of Wyoming
Uppsala University
Utah State University
Vanderbilt University
Vertex Pharmaceuticals
Virginia Tech
VizX Labs
Wadsworth Center
Washington State University
Washington University (St. Louis)
Wayne State University
Weyerhauser
Whitehead Institute for Biomedical Research
Whitehead Institute/MIT Center for Genome Research
Windber Research Institute

Appendix E. GTL-Funded Projects

Program Projects 246

Communication 247

Bioinformatics, Modeling, and Computation 247

Environmental Genomics 247

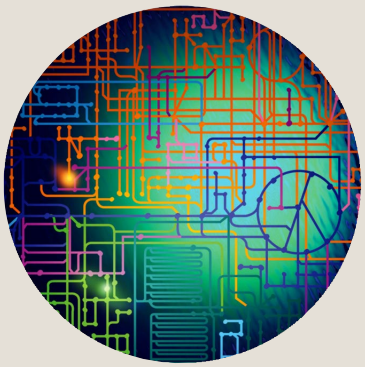
Microbial Genomics 247

Technology Development and Use 247

- Imaging, Molecular, and Cellular Analysis 247
- Protein Production and Molecular Tags 248
- Proteomics and Metabolomics 248

Ethical, Legal, and Social Issues 248

Computing and Education 248



GTL-Funded Projects*

Additional information:

- 3.3. Highlights of Research in Progress to Accomplish Milestones, p. 55
- See "GTL Research" on GTL web site (www.doegenomestolife.org)

Program Projects

Harvard Medical School

- Microbial Ecology, Proteogenomics, and Computational Optima

J. Craig Venter Institute

- Reconstruction of a Bacterial Genome from DNA Cassettes and Sargasso Sea Metagenomics

Joint Genome Institute

- DNA Sequencing for Genomics:GTL

Lawrence Berkeley National Laboratory

- Rapid Detection of Stress Response Pathways in Metal- and Radionuclide-Reducing Bacteria; with Sandia National Laboratories (SNL), Lawrence Livermore National Laboratory (LLNL), and Oak Ridge National Laboratory (ORNL)

Oak Ridge National Laboratory and Pacific Northwest National Laboratory

- Genomics:GTL Center for Molecular and Cellular Systems; with SNL, LLNL, and Argonne National Laboratory (ANL)

Sandia National Laboratories

- Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Heirarchical Modeling; with ORNL

Shewanella Federation

- The *Shewanella* Federation: Environmental Sensing, Metabolic Response, and Regulatory Networks in the Respiratory Versatile Bacterium *Shewanella;* with Pacific Northwest National Laboratory (PNNL); BIATECH; Boston University; University of California, Los Angeles (UCLA); ORNL; Michigan State University; University of Southern California; Baylor University; and Genomatica

*Projects funded between GTL's 2002 inception and July 2005. Some listed projects are not funded currently, and GTL-supported workshops and conferences are not included.

University of Massachusetts, Amherst

- Analysis of the Genetic Potential and Gene Expression of Microbial Communities Involved in the In Situ Bioremediation of Uranium
- Harvesting Electrical Energy from Organic Matter; with ANL

Communication

- Genome Management Information System; ORNL

Bioinformatics, Modeling, and Computation

- A Conceptual and In Silico Model of the Dissimilatory Metal-Reducing Microorganism, *Geobacter sulfurreducens*; University of Massachusetts, Amherst
- Animal Gene Regulatory Networks; California Institute of Technology
- Cofunding for BioSpice Projects; Defense Advanced Research Projects Agency (DARPA)
- Computation Hypothesis Testing: Integrating Heterogeneous Data and Large-Scale Simulation to Generate Pathway Hypotheses; Gene Network Sciences
- Computational Resources for GTL; Keck Graduate Institute
- Computing Frontiers: Prospects from Biology; National Academy of Sciences
- Development of Advanced Tools for Data Management, Integration, Analysis, and Visualization Through a Comprehensive Systems Analysis of the Halophilic Archaeon; Institute for Systems Biology
- Development of Bioinformatics and Experimental Technologies for Identification of Prokaryotic Regulatory Networks; Brown University
- Identification and Characterization of Prokaryotic Regulatory Networks; Washington University

Environmental Genomics

- Application of High-Throughput Gel Microdroplet Culturing to Develop a Novel Genomics Technology Platform; Diversa Corp.
- Genome-Facilitated Analyses of Geomicrobial Processes; PNNL, LLNL, LBNL
- Growth of Uncultured Microorganisms from Soil Communities; Northeastern University and ORNL
- Proteogenomic Approaches for the Molecular Characterization of Natural Microbial Communities; University of California, Berkeley

Microbial Genomics

- Dynamics of Cellular Processes in *Deinococcus radiodurans*; Henry M. Jackson Foundation
- Genome-Wide Analysis of *Prochlorococcus marinus* Protein-Protein Interactions; LLNL
- Global Characterization of Genetic Regulatory Circuitry Controlling Adaptive Metabolic Pathways; Stanford University
- Metabolic Engineering of Light and Dark Biochemical Pathways in Wild-Type and Mutant Strains of *Synechocystis* PCC 6803 for Maximal, 24-Hour Production of Hydrogen Gas; Oregon State University
- Metabolic Functional Analysis of Bacteria Genomes; Oregon State University
- Molecular Basis for Metabolic and Energetic Diversity; University of Wisconsin
- Rapid Reverse Engineering of Genetic Networks Via Systematic Transcriptional Perturbations; Boston University
- *Rhodopseudomonas palustris* Microbial Cell Project; ORNL and Ohio State University
- Whole Genome Transcriptional Analysis of Environmental Stresses in *Caulobacter crescentus*; LBNL

Technology Development and Use

Imaging, Molecular, and Cellular Analysis

- Development of a Hybrid Electron Cryotomography Scheme for High-Throughput Protein Mapping in Whole Bacteria; Brookhaven National Laboratory (BNL)
- Dynamic Spatial Organization of Multiprotein Complexes Controlling Microbial Polar Organization, Chromosome Replication, and Cytokinesis; Stanford University

- Electron Tomography of Microbial Cells; LBNL
- Microscopies of Molecular Machines: Structural Dynamics of Gene Regulations in Bacteria; LBNL
- New, Highly Specific Vibrational Probes for Monitoring Metabolic Activity in Microbes and Microbial Communities; LLNL
- Probe of Single Microbial Proteins and Multiprotein Complexes with Bioconjugated Quantum Dots; Georgia Tech Research Corporation
- Real-Time Gene Expression Profiling of Live *Shewanella oneidensis* Cells; Harvard University and PNNL
- Single Cell Imaging of Macromolecular Dynamics in a Cell; LBNL
- Use of Near-Infrared Probes of Microscopic Functional Analysis of Microbial Consortia Including Hard-to-Culture Microbes; ANL

Protein Production and Molecular Tags

- Center for Genomics and Proteomics Research Program; UCLA
- Chemical Methods for the Production of Proteins and Protein and Peptide Reagents and for the Characterization of Protein Complexes; University of Chicago
- Combined Informatics and Experimental Strategy for Improving Protein Expression; University of Maryland Biotech Institute
- Development of Genome-Scale Expression Methods; ANL
- Development of Multipurpose Tags and Affinity Reagents for Rapid Isolation and Visualization of Protein Complexes; PNNL
- Functionalized Nanotubes of Enzyme Immobilization (Pilot Project); PNNL
- High-Throughput Biophysical Analyses of Purified Proteins; BNL
- Integrated Approach to Functional Genomics; Los Alamos National Laboratory (LANL)

Proteomics and Metabolomics

- Application of High-Throughput Proteomics Structural Studies of Essential Proteins from *D. radiodurans* and *S. oneidensis*; PNNL
- Development of High-Throughput Proteomics Production Operations; PNNL
- Growth and Metabolism of Individual Bacterial Cells Utilizing Nanosims; LLNL
- Metabolomic Functional Analysis of Bacterial Genomes; LANL
- Microbial Communities; PNNL
- New Technologies for Metabolomics; LBNL
- Technology for Ultrahigh-Resolution Localization of Gene Transfer; PNNL

Ethical, Legal, and Social Issues

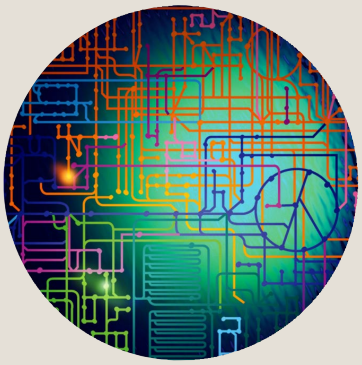
- DNA Files; SoundVision Productions
- Science Literacy Workshop; SoundVision

Computing and Education

- BACTER (Bringing Advanced Computational Resources to Environmental Research) Institute; University of Wisconsin, Madison
- Center for Computational Biology; University of California, Merced; Renssalear Polytechnic Institute; and LLNL
- Institute for Multiscale Modeling of Biological Interactions; Johns Hopkins University, University of Delaware, and LANL

Appendix F. Strategic Planning for CCSP and CCTP

CCSP Goal 250
Improve Quantification of the Climate-Changing Forces and Related Systems 250
CCTP Goals 250
Reduce Emissions From the Energy Supply 250
Capture and Sequester Carbon Dioxide 251
Reduce Emissions of Other Greenhouse Gases 251
Enhance Capabilities to Measure and Monitor GHG Emissions 251



Strategic Planning for CCSP and CCTP

Strategic planning for Climate Change Science Program (CCSP) and Climate Change Technology Program (CCTP) included analysis of hundreds of technologies and methods to meet national and global goals. Biological and environmental research studies (including those on the roles of microbes in ocean and terrestrial environments) are critical to understanding and predicting climate changes and developing technologies for biobased fuel production and climate mitigation and adaptation.

Knowledge from GTL research will be the foundation for developing microbial strategies that support CCSP and CCTP goals outlined below.

CCSP Goal

- **www.climatescience.gov**

Improve Quantification of the Climate-Changing Forces and Related Systems

- Improve understanding of key "feedbacks" and sensitivities of biological and ecological systems and accelerate incorporation into climate models to reduce uncertainty.
- Develop information on the carbon cycle to assist in evaluation of carbon-sequestration strategies and alternative response options.

CCTP Goals

- **www.climatetechnology.gov**

Reduce Emissions From the Energy Supply

- Use microbes or microbial enzymes in nanostructures for the photosynthetic production of hydrogen and other high-energy fuels.
- Use microbes for production of biofuels from biomass, in situ bioprocessing of fossil fuels, and design of improved biomass feedstocks.

Capture and Sequester Carbon Dioxide

- Understand the role of microbes in the cycling of carbon in terrestrial and marine environments.
- Use this understanding to determine efficacy and impacts of ocean and terrestrial sequestration strategies.
- Incorporate microbial processes into systems for capturing carbon dioxide from the atmosphere.

Reduce Emissions of Other Greenhouse Gases

- Understand microbial contributions to nitrous oxide and methane emissions from natural biogeochemical cycles.
- Understand the communities of microbes in the digestive systems of livestock that release methane.

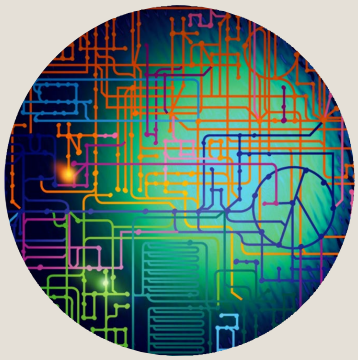
Enhance Capabilities to Measure and Monitor GHG Emissions

- Develop ecogenomic sensors for detecting changes in ocean and soil microbial communities.
- Quantify net emissions of GHGs from microbial processes.

APPENDIX F

Appendix G. Microbial Genomes Sequenced or in Process by DOE

Carbon Sequestration 254
Energy Production 257
Bioremediation 257
Cellulose Degradation 260
Biotechnology and Applied Microbiology 261
Microbial Consortia 262
Technology Development, Pilot Projects 262



Microbial Genomes Sequenced or in Process by DOE

Uncovering Potential Applications Relevant to DOE Missions

Program Manager

Daniel Drell, 301.903.4742
Daniel.Drell@science.doe.gov

For updates to list, see http://microbialgenome.org/brochure.pdf

Joint Genome Institute Microbial Sequencing

http://genome.jgi-psf.org/mic_home.html

For information on JGI's Community Sequencing Program, see p. 53.

Microbes and microbial consortia sequenced or in process for various DOE missions include those listed below (last updated July 15, 2005).

Carbon Sequestration

‡***Aureococcus anophagefferens*** (algae, ~32 Mb): Brown tide-forming pelagophyte, forms coastal blooms, reduces trace metals; can sequester substantial amounts of carbon.

†***Azotobacter vinelandii* AvOP** (bacteria, 4.5 Mb): Aerobic, fixes nitrogen; found in soils worldwide; has nitrogenases incorporating molybdenum and vanadium (in addition to iron); relevant to energy use and carbon sequestration.

‡***Bradyrhizobium sp.* strain BTAi** (bacteria, 9.2 Mb): Versatile photosynthetic; carbon dioxide- and nitrogen-fixing symbiont of legumes; nodule-forming on roots and stems; aids plant in carbon processing.

‡***Calyptogena magnifica*** (clam) proteobacterial symbiont (bacteria, est. ~ 4 Mb): Isolated from deep sea vents, sulfur oxidizing, nitrogen fixing; fixes carbon dioxide via possibly novel pathway; carbon sequestration.

†***Chlamydomonas reinhardtii*** (eukaryotes, ~100 Mb): Green alga, photosynthetic, widespread in environment, 17 chromosomes, widely used model system.

Photosynthetic Green Sulfur Bacteria: Sequester carbon via photosynthesis; produce hydrogen when cocultured with sulfate-reducing bacteria.

- ‡***Chlorobium limicola* DSMZ 245(T)** (2.4 Mb): Nonmotile, rod shaped; type strain for all green sulfur bacteria.
- ‡***Chlorobium phaeobacteroides* DSMZ 266T** (~2.2 Mb): Rod shaped; does not use nitrogen or sulfide.
- ‡***Chlorobium phaeobacteroides* MN1 Black Sea** (2.2 Mb): Photosynthetic in very low light, with chlorophylls that absorb 1 photon every 5 hours.
- ****Chlorobium tepidum*** (2.1 Mb): Photosynthetic; may play important role in earth's overall carbon cycle.

- ‡***Chlorobium vibrioforme f. thiosulfatophilum* DSMZ 265(T)** (2.5 Mb): Curved rod shape.
- †***Chloroflexus aurantiacus* J-10-fl** (3 Mb): Modern version of organism; needs no oxygen for photosynthesis; uses unique pathway to fix carbon dioxide.
- ‡***Chloroherpeton thalassium*** (~3 to 3.5 Mb): Most taxonomically divergent of green sulfur bacteria.

***Chloroflexi* Bacteria (7 Strains, est. ~5 Mb each):** Gram-negative, filamentous anoxygenic phototrophs; useful in carbon sequestration, biofuels.

- ‡***Candidatus Chlorothrix halophila***: Marine and hypersaline biofilms; produces bacterial chlorophylls (BChl) a and c and chlorosomes.
- ‡***Chloroflexus aggregans* DSMZ 9485:** Motile, grows at 55°C in both light and dark.
- ‡***Chloronema sp.* strain UdG9001**: Motile, photoautotrophic; isolated from Little Long Lake, Wis.; grows in iron-rich environments.
- ‡***Heliothrix oregonensis***: Bright orange colored, motile; grows optimally at 45° to 55°C; forms monolayers on top of microbial biofilms.
- ‡***Herpetosiphon aurantiacus* DSM 785**: Orange colored, isolated from Birch Lake, Minn.; hydrolyzes starch, does not produce BChls.
- ‡***Roseiflexus castenholzii* DSM 13941**: Motile, red to reddish-brown colored; has BChl-a but not BChl-c or chlorosomes.
- ‡***Roseiflexus sp.* strain RS-1**: Isolated from high-temperature biofilms in Octopus Spring, Yellowstone.

◊***Colwellia* 34H** (bacteria, 5.3 Mb): Psychrophile, important in carbon and nutrient cycling in polar marine environments.

‡***Crocosphaera watsonii* WH8501** (cyanobacteria, 3.6 to 5 Mb): Marine, unicellular; confined to waters from 26 to 32°C; temporally segregates carbon-dioxide fixation from nitrogen fixation.

‡***Emiliania huxleyi* 1516** (marine algae, ~5 Mb): Marine coccolithophorids; plays role in global carbon cycling and sulfur transformation.

‡***Frankia* Cc13** (bacteria, ~8 to 10 Mb): Actinomycetes, Group II, fixes nitrogen; forms major nitrogen-fixing symbiosis in temperate soils; promotes formation of woody-biomass energy source.

‡***Frankia sp.* EAN1pec** (bacteria, ~10 Mb): Group III, ubiquitous, fixes nitrogen, forming major nitrogen-fixing symbiosis in temperate soils; promotes formation of woody-biomass energy source; grows well, shows metal resistance.

‡***Jannaschina sp.* CCS1** (bacteria, 4.5 to 5 Mb): Member of *Roseobacter* clade; contributes to oceanic anoxygenic phototrophy, a mode of light-driven energy acquisition.

***Micromonas pusilla ssp.* Eukarya (2 strains, ~15 Mb each):** Abundant in oceans, very small (1 to 3 microns in length); significant planktonic primary producers (carbon-dioxide fixers) in size class; carbon sequestration.

- ‡***M. pusilla* NOUM17(RCC 299)**: Equatorial Pacific isolate.
- ‡***M. pusilla* CCMP1545**: England West Coast isolate.

‡***Moorella thermoacetica*** ATCC39073 (bacteria): Fixes carbon dioxide in absence of oxygen; can grow on hydrogen, carbon dioxide, or carbon monoxide as sole carbon source; acetogenic.

Bacteria Involved in Nitrification Affecting Climate Change: Oxidize ammonia; can degrade chlorinated aliphatic hydrocarbons; give insight into basis of biogeochemical nitrogen cycle.

- ‡***Nitrobacter hamburgensis*** (~3 Mb): Found in soil; model organism for biochemical, structural, and molecular investigations; has carboxysomes.
- ‡***Nitrobacter winogradskyi* Nb-255** (~3 Mb): Widely distributed, also nitrite oxidizing; can grow with several metabolic modes and anoxically by denitrification; can fix carbon dioxide.
- ‡***Nitrosococcus oceani*** (~3 Mb): Gamma proteobacterium that oxidizes ammonia (others are beta proteobacteria).
- ****Nitrosomonas europaea* ATCC19718** (2.2 Mb): Aids incorporation of carbon dioxide into biomass.
- ‡***Nitrosomonas eutropha*** (~3 Mb): Physiologically diverse; can oxidize nitrous oxide while reducing either ammonia or hydrogen; important in wastewater treatment systems; potential for remediation of high ammonia concentrations in waters.

- ‡***Nitrosospira multiformis* Surinam** (~3 Mb): Well-studied, typical of those seen in soil environments.

*******Nostoc punctiforme* ATCC29133** (bacteria, 10 Mb): Fixes carbon dioxide and nitrogen; produces hydrogen; survives acidic, anaerobic, and low-temperature conditions.

‡***Ostreococcus*** (eukaryotes, est. 8 to 10 Mb): Fast-growing, ubiquitous; important in marine carbon fixation.

‡***Pelodictyon luteolum* DSMZ 273(T)** (bacteria, 3.0 Mb): Rod-shaped photosynthetic GSB cells that can form yellow-green hollow microcolonies.

‡***Pelodictyon phaeoclathratiforme* BU-1, DSMZ 5477** (bacteria, 3 Mb): Gas vesicle containing green sulfur bacterium cells that can form 3D net-like microcolonies.

*******Prochlorococcus isolate* NATL2** (prokaryotes, 1.7 to 2.4 Mb): Ocean carbon sequestration.

******Prochlorococcus marinus* MED4** (bacteria, 1.7 Mb), ******Prochlorococcus marinus* MIT9313** (bacteria, 2.4 Mb), and *******Prochlorococcus marinus* MIT9312** (bacteria, ~2.4 Mb): All ecotypes abundant in temperate and tropical oceans; important in ocean carbon cycling; absorb blue light efficiently; **MIT9313** is adapted to lower-light conditions (lower ocean depths) and **MIT9312** to higher-light conditions nearer the surface.

‡***Prosthecochloris aestuarii SK413*, DSMZ 271(t)** (bacteria, 2.5 Mb): Nonmotile, spherical to ovoid green sulfur bacteria; nitrogen-fixing marine strain; high salt requirement.

***Rhodopseudomonas palustris* Bacteria (5 strains, ~5.5 Mb each):** Metabolically versatile, can produce hydrogen, fix carbon dioxide, biodegrade organic pollutants and plant biomass; biofuels.

- ‡***R. palustris* BisA53:** Isolated from Dutch site; grows well on benzoate, tends to aggregate.
- ‡***R. palustris* BisB5:** Isolated from Dutch contaminated site; smaller, more motile form; fewer rosettes than sequenced **CGA009.**
- ‡***R. palustris* BisB18:** Isolated from Dutch site; slower growing than **CGA009**.
- *******R. palustris* CGA0009**: Biodegrades under both aerobic and anaerobic conditions.
- ‡***R. palustris* HaA2**: Unable to grow on benzoate; isolated from Haren site.

*******Rhodospirillum rubrum* ATCC11170** (bacteria, 3.4 Mb): Phototrophic; grows in various conditions, including aerobic and anaerobic; fixes nitrogen, grows on hydrogen; model for photosynthesis.

‡***Roseobacter* strain TM1040** (bacteria, ~4.5 Mb each): Isolated from dinoflagellate; fixes carbon in marine surroundings.

‡***Sphingopyxis alaskensis* RB2256** (bacteria, 3.2 Mb): Makes up large proportion of oceanic biomass; major contributor to global carbon flux; can bioconcentrate trace metals.

*******Synechococcus elongates* PCC7 942** (cyanobacteria, 2.4 to 2.7 Mb): Carbon fixation; photosynthesis in fresh waters.

‡***Synechococcus sp.* C9902 (coastal)** and **Cc9605 (oligotrophic)** (bacteria, ~2.4 Mb each): Fixes carbon dioxide; globally distributed; important in carbon fluxes in marine environment.

******Synechococcus* WH8102** (bacteria, 2.4 Mb): Photosynthetic; important to ocean carbon fixation; genetically tractable.

†***Thalassiosira pseudonana*** (eukarya, ~25 Mb): Ocean diatom, major participant in biological "pumping" of carbon to ocean depths.

*******Thiobacillus denitrificans* ATCC23644** (bacteria, ~2 Mb): Fixes carbon; oxidizes sulfur and iron; involved in bioremediation.

‡***Thiomicrospira crunogena*** (bacteria, 2 Mb): Marine gamma proteobacterium isolated from East Pacific; found in deep sea vents; grows rapidly (doubling time, ~1 hour); carbon-concentrating mechanism similar to cyanobacteria; sulfur oxidizing; fixes carbon dioxide; can grow in low to absent oxygen conditions; desulfurylates coal; strips sour gas (hydrogen sulfide) from petroleum.

‡***Thiomicrospira denitrificans*** (~1.6 Mb): Marine epsilon proteo-bacterium found in hydrothermal vents but also in oxygen-containing, anoxic ocean-transition regions; uses reverse TCA cycle for carbon fixation; sulfur oxidizing; fixes CO_2; can grow in low to absent oxygen conditions; desulfurylates coal; strips sour gas (HS) from petroleum.

†***Trichodesmium erythraeum* IMS101** (bacteria, 6.5 Mb): Key nitrogen-fixing microbe; plays major role in tropical and subtropical oceans.

Energy Production

****Anabaena variabilis ATCC29413** (cyanobacteria, 7 to 10 Mb): Filamentous heterocyst-forming; fixes nitrogen and carbon dioxide; produces hydrogen.

‡***Caldicellulosiruptor saccharolyticus*** (bacteria, 4.3 Mb): Versatile biomass-degrading, hydrogen-producing thermophile; biofuels.

◊***Carboxydothermus hydrogenoformans*** (bacteria, 2.10 Mb): Gram positive; converts carbon monoxide and water to carbon dioxide and hydrogen.

‡***Clostridium phytofermentans*** (bacteria, ~5 Mb): Degrades plant polymer cellulose, pectin, starch, and xylan to produce ethanol and hydrogen.

‡***Clostridium beijerinckii* NCIMB 8052** (bacteria, 6.7 Mb): Produces solvent; converts biomass to fuels and chemicals; potential for alternate energy production.

****Methanobacterium thermoautotrophicum* Delta H** (archaea, 1.7 Mb): Produces methane; plays role in earth's overall carbon cycle.

†***Methanococcoides burtonii* DSM6242** (archaea, 3 Mb): Extremophile adapted to cold (less than 5°C); produces methane.

****Methanococcus jannaschii* DSM2661** (archaea extremophile, 1.7 Mb): May identify high-temperature, high-pressure enzymes; produces methane.

‡***Methanosaeta thermophila* P_T(DSM6194)** (archaea, ~ 3Mb): Widely distributed in environment; metabolizes acetates into methane; potential producer of biofuel.

†***Methanosarcina barkeri* Fusaro** (archaea, 2.8 Mb): Lives in cattle rumen; digests cellulose and other polysaccharides to produce methane; very oxygen sensitive; grows in variety of substrates.

‡***Methanospirillum hungateii* JF1** (bacteria, 2.8 Mb): Methanogen; system for studying multispecies microbial assemblage composed of metabolically diverse microorganisms functioning as a single catalytic unit.

‡***Methylobacillus flagellatus* KT** (proteobacteria, 3.1 Mb): Bioremediation; cycling of one-C compounds; environmentally benign bioprocessing into feedstocks.

‡***Pichia stipitis* CBS 6054** (fungi, 12 Mb): Ferments xylose to ethanol; potential to oxidize products of lignin degradation and play a role in cellulose degradation as endosymbiont of beetles; converts biomass to ethanol.

‡***Syntrophomonas wolfei* Göttingen DSM 2245B** (bacteria, 4.5 Mb): Methanogenic and syntrophic; potential hydrogen producers; useful in bioremediation; system for studying multispecies microbial assemblage of metabolically diverse microorganisms functioning as single catalytic unit.

‡***Syntrophobacter fumaroxidans* MPOB** (bacteria, 3.3 Mb): Methanogenic proprionate oxidizer; uses fumarate as electron acceptor; can produce hydrogen and formate; syntrophic (i.e., part of bacteria community).

◊***Thermotoga neopolitana* ATCC49045** (bacteria, ~1.8 Mb): Combines with oxygen to produce hydrogen.

Bioremediation

‡***Acidiphilium cryptum* JF 5** (bacteria, 2.46 Mb): Reduces iron and iron oxides in very acid conditions (pH 2.2 to 5); possible bioremediation of metals in acid environments.

†***Acidithiobacillus ferrooxidans*** (bacteria, 2.9 Mb): Used in mining industry to sequester iron and sulfide.

‡***Acidobacterium sp.*** (bacteria, two Group 1 strains, one Group 3 strain, est. ~4 Mb each): Ubiquitous in soil, including those contaminated with chromium, zinc, other metals, and PCBs.

‡***Alkaliphillus metalliredigenes*** (bacteria, ~4 Mb): Reduces iron, other metals, uranium under alkaline conditions (optimal growth, pH 9.6).

‡***Anaeromyxobacter delahogenans* 2CP-C** (bacteria, 3.38 Mb): Reduces metal (iron, uranium, others); degrades aromatic and halogenated hydrocarbons.

‡***Arthrobacter sp.* strain FB24** (bacteria, ~2.4 Mb): Resists metal (reduces chromium, lead); degrades hydrocarbon; resists radiation; widely distributed in soils.

‡***Burkholderia ambifaria*** (bacteria, 4.7 Mb): Genomovar VII; smallest *Burkholderia* genome, biocontrol agent.

‡***Burkholderia ambifaria* AMMD** (bacteria, ~7.2 Mb): Ubiquitous rhizosphere colonizer and member of the *Burkholderia cepacia* complex; nitrogen fixer, organic-pollutant degrader; bioremediation.

†***Burkholderia xenovorans (formerly Burkholderia fungorum)* LB400** (bacteria, 8 Mb): Outstanding degrader of polychlorinated biphenyls (PCBs).

‡***Burkholderia vietnamiensis* G4** (bacteria, ~8 to 10 Mb): Genomovar V; degrades trichloroethylene; colonizes rhizosphere.

****Caulobacter crescentus*** (bacteria, 4.01 Mb): Potential for heavy-metal remediation in waste-treatment plant wastewater.

‡***Chromohalobacter salexigens* DSM 3043** *(*formerly *Halomonas elongatee)* (bacteria, 4 Mb): Most-halotolerant eubacteria known; displays metal resistance; degrades aromatic hydrocarbons and toxic organics; high halotolerance, suggesting applications in extreme environments.

†***Dechloromonas*** RCB (bacteria, 2 Mb): Oxidizes iron. Converts perchlorate to chloride; anaerobically oxidizes benzene to carbon dioxide.

****Dehalococcoides ethenogenes*** (bacteria, 1.5 Mb): Degrades dangerous solvent trichloroethene to benign products.

‡***Dehalococcoides sp.* strain BAV1** (bacteria, 2 Mb): Detoxifies many dichloroethene isomers; potential for bioremediating organic-compound contamination; isolated from Michigan site.

‡***Dehalococcoides sp*. strain VS** (bacteria, 1.5 Mb): Detoxifies many dichloroethene isomers; potential for bioremediation of organic-compound-contaminated sites, isolated from site in Texas.

‡***Deinococcus geothermalis* DSM11300** (bacteria, ~3 Mb): Resists radiation; can bioremediate radioactive mixed waste at temperatures up to 55°C.

****Deinococcus radiodurans* R1** (bacteria, 3 Mb): Survives extremely high levels of radiation; possesses DNA-repair capabilities for radioactive waste cleanup.

†***Desulfitobacterium hafniense* DCB-2** (bacteria, 4.6 Mb): Degrades pollutants such as chlorinated organic compounds that include some pesticides.

‡***Desulfotomaculum reducens* MI-1** (bacteria, 4 Mb): Gram-positive, spore-forming, metabolically versatile sulfate and metal (iron, manganese, uranium, chromium) reducer. Can reduce uranium and nitrate simultaneously; bioremediation.

*****Desulfovibrio desulfuricans* G20** (bacteria, 3.1 Mb): Anaerobic; reduces sulfate, uranium, and toxic metals; corrodes iron piping; "sours" petroleum with hydrogen sulfide.

****Desulfovibrio vulgaris* Hildenborough** (bacteria, 3.2 Mb): High potential for bioremediation through metal and sulfate reduction and sulfate utilization.

†***Desulfuromonas acetoxidans*** (bacteria, 4.1 Mb): Marine microbe; reduces iron; oxidizes acetate to carbon dioxide under anoxic conditions via process coupled to sulfur reduction or iron (III).

†***Ferroplasma acidarmanus* fer1** (archaea, 2 Mb): Lives in most acidic conditions on earth; oxidizes iron; transforms sulfide in metal ores to sulfuric acid, leading to contamination of mining sites.

†***Geobacter metallireducens*** (bacteria, 6.8 Mb): Widespread in freshwater sediments; gains energy by reducing iron, manganese, uranium, and other metals; oxidizes toluene and phenol.

‡***Geobacter sp*. strain FRC-32** (bacteria, ~5 Mb): Iron and uranium reducer, isolated from uranium-contaminated subsurface at U.S. DOE-NABIR Field Research Center; bioremediation.

◊***Geobacter sulfurreducens*** (bacteria, 2.5 Mb): Reduces a variety of metals, including iron and uranium.

‡***Glomus intraradices*** (fungi, ~11 to 12 Mb): Forms spores to establish a functional symbiotic (and pathogenic) relationship with plant roots.

‡***Kineococcus radiotolerans* nov** (bacteria, 4.3 to 4.6 Mb): Highly radioresistant; degrades organic pollutants.

‡***Laccaria bicolor*** (fungi, ~40 Mb): Commonly found mushroom; stimulates root formation, differentiation in various plants.

†***Mesorhizobium* BNC1** (bacteria, 5 Mb): Fixes nitrogen with leguminous plants; agriculturally important.

*****Methylobium petroleophilum* PM1** (bacteria, 4.6 Mb): Degrades diverse hydrocarbons, including MTBE (methyl tertiary butyl ether, a common fuel additive), benzene, toluene, xylene, and phenol; biodegradation.

◊***Methylococcus capsulatus*** (bacteria, 4.6 Mb): Uses methane as single carbon and energy source; generates pollutant-oxidizing enzymes; used commercially to produce biomass and other proteins.

Mycobacteria (5 isolates, est. ~5 Mb each): Fast growing, nonpathogenic; degraders of polycyclic aromatic hydrocarbons (PAH); found in soils.

- ‡***Mycobacterium flavescens***: Isolated from PAH-contaminated site in Indiana.
- ‡***Mycobacterium vanbaalenii***: Isolated from PAH-contaminated site in Texas.
- ‡***Mycobacterium sp.* KMS**: Isolated from remediated superfund site, Libby, Montana.
- ‡***Mycobacterium sp.* JLS**: Isolated from remediated superfund site, Libby, Montana.
- ‡***Mycobacterium sp.* MCS**: Isolated from remediated superfund site, Libby, Montana.

‡***Nectria haematococca* MPVI** (fungi, ~40 Mb): Member of *Fusarium solani* species complex; ubiquitous; degrades lignins, hydrocarbons, plastics, some pesticides; useful in bioremediation.

‡***Nocardioides* strain JS614** (bacteria, ~4.5 Mb): Grows aerobically and efficiently on vinyl chloride (VC) and ethene. If starved of VC for more than 1 day, will not recover for more than 40 days; 300-Kb plasmid containing VC- and ethene-degradation pathways.

†***Novosphingobium aromaticivorans* F199** (bacteria, 3.8 Mb): Degrades aromatic compounds in soil, including toluene, xylene, naphthalene, and fluorine.

Bacteria Involved in Microbial Arsenic Transformation (~2 to 4 Mb each)

- ‡***Bacillus selenitireducens* MLS-10**: Haloalkaliphile, respires toxic selenium, argon, sulfur, nitrates.
- ‡***Bacillus selenitireducens* MLMS-1**: Respires argon, fixes carbon dioxide in apparent absence of RuBisCo.
- ‡***Clostridium sp.* OhILAs**: Strict anaerobe, spore forming; respires argon, nitrates, sulfur, and selenium.
- ‡***Clostridium sp.* MLHE-1**: Oxidizes arsenite, potentially can fix carbon dioxide via Form 1 RuBisCo.

‡***Paracoccus denitrificans*** (bacteria, 3.66 Mb): Bioremediates various pollutants; involved in carbon sequestration and denitrification; may be closely related to evolutionary precursor of mitochondria.

‡***Polaromonas naphthalenivorans sp.* strain nov CJ2** (bacteria, ~6 Mb): Degrades PAHs, naphthalene in situ in contaminated environment; bioremediation.

‡***Beta proteobacterium sp.* JS666** (bacteria, ~4.5 Mb): Only aerobic bacterium reported to grow on *cis*-dichloroethene (cDCE, a common contaminant at DOE sites); yellow, nonmotile; devoid of vacuoles; prefers 20°C but will not grow at 30°C or on vinyl chloride or ethene.

‡***Pseudoalteromonas atlantica*** (bacteria, 3.5 Mb): Marine, gram-negative, motile, biofilm forming, secretes degradative enzymes, polysaccharides that bind metals; bioremediation.

†***Pseudomonas fluorescens* PFO-1** (bacteria, 5.5 Mb): Metabolically diverse; degrades pollutants such as styrene, TNT, and polycyclic aromatic hydrocarbons; useful in applications requiring bacteria release and survival in soil.

****Pseudomonas putida*** (bacteria, 6.1 Mb): High potential for bioremediation by reducing metal and pollutants.

‡***Pseudomonas putida* F1** (bacteria, 6.2 Mb): Grows well on a variety of aromatic hydrocarbons including benzene, toluene, ethylbenzene; bioremediation of organics.

‡***Ralstonia eutropha* JMP-134** (bacteria, 7.24 Mb): Gram negative; degrades chloroaromatic compounds and chemically related pollutants; potential for bioremediation.

†***Ralstonia metallidurans* CH34** (bacteria, 5 Mb): Contains two "mega" plasmids; resistant to wide variety of heavy metals, which accumulate on the cell surface; strong potential for bioremediation of metals.

****Rhodobacter sphaeroides* 2.4.1** (bacteria, 4.4 Mb): Metabolically diverse, grows in wide variety of conditions; photosynthetic, providing fundamental insights into light-driven, renewable-energy production; can detoxify metal oxides, useful in bioremediation.

Metal-Reducing *Shewanella* Bacteria: Affect metals including uranium, technetium, and chromium; important in carbon cycling in anaerobic environments; thrive in redox gradient environments; produce energy by generating weak electrical current; display metabolic diversity, potential for bioremediation.

- ‡***Shewanella amazonensis*** (4.3 Mb): Isolated from sediments in Amazon River delta; active in reduction of iron, manganese, and sulfur compounds; optimal growth at 35°C, with 1% to 3% salt.
- ‡***Shewanella baltica* OS195** (est. ~5 Mb): Second *S. baltica* strain, ~69% DNA homology with **OS155.**
- ‡***Shewanella baltica* OS1155** (3.6 Mb): Isolated from Gotland Deep in central Baltic Sea, predominantly low- and zero-oxygen regions; can use glycogen, cellobiose, and sucrose as sole sources of carbon and energy.
- ‡***Shewanella denitrificans* OS220** (3.1 Mb): Denitrifies vigorously; isolated from Gotland Deep in central Baltic Sea; uses nitrate, nitrite, and sulfite as electron acceptors.
- ‡***Shewanella frigidimarina* NCMB400** (2.1 Mb): Isolated from North Sea off coast of Aberdeen; rich in *c*-type cytochromes, with increased cytochrome synthesis during growth in low- to zero-oxygen conditions when iron is present.
- ****Shewanella oneidensis* MR-1** (bacteria, 4.5 Mb): May degrade organic wastes and reduce or sequester a range of toxic metals.
- ‡***Shewanella putrefaciens* CN-32** (3.22 Mb): Isolated from uranium-bearing subsurface formation in northwestern New Mexico; reduces array of metals and radionuclides, including solid-phase iron and manganese oxides, uranium (VI), technetium (VII), and chromium (VI) with hydrogen, formate, or lactate; has unusual membrane sugars.
- ‡***Shewanella putrefaciens* ML-S2** (est. ~5 Mb): Hypersaline, pH ~10 environment; isolated from Mono Lake, Calif.
- ‡***Shewanella putrefaciens* p200** (3.2 Mb): Isolated from corroding oil pipeline in Canada; among most genetically characterized metal-reducing *Shewanellae;* degrades carbon tetrachloride under low- to zero-oxygen conditions.
- ‡***Shewanella putrefaciens* W3-6-1** (est. ~5 Mb): Marine; forms magnetite at 0°C.
- ‡***Shewanella sp.* ANA-3** (est. ~5 Mb): Fast-growing, unique As(V) respiratory mechanism.
- ‡***Shewanella sp.* MR-4** (est. ~5 Mb): Isolated from 5-M depth (oxic) of Black Sea.
- ‡***Shewanella sp.* MR-7** (est. ~5 Mb): Isolated from 60-M depth (anoxic) of Black Sea.
- ‡***Shewanella sp.* PV-4** (4 to 4.5 Mb): Most diverse of *Shewanellae;* prefers cold temperatures; produces magnetite at 0°C; reduces cobalt at –4°C.

‡***Xanthobacter autotrophicus* Py2** (bacteria, 5 Mb): Ubiquitous, nutritionally versatile; degrades chlorinated hydrocarbons, fixes nitrogen, synthesizes biodegradable plastics; bioremediation.

Cellulose Degradation

†***Clostridium thermocellum*** ATCC27405 (bacteria, ~5 Mb); Degrades cellulose; potentially useful for conversion of biomass (cellulose) to energy.

*****Cytophaga hutchinsonii* ATCC33406** (bacteria, 4 Mb): Very abundant in nature; decomposes cellulose, lacks cellulosomes.

‡***Flavobacterium johnsoniae*** (bacteria, 4.8 Mb): Common in soils and freshwaters; degrades chitin and numerous other macromolecules via direct contact; possible use in biomass conversion.

*****Microbulbifer degradans* 2-40** (bacteria, 6 Mb): Marine microbe; degrades and recycles insoluble complex polysaccharides via protruding membrane structures called hydrolosomes; potential for conversion of complex biomass to energy.

****Phanerochaete chrysosporium*** (eukarya, ~30 Mb): "White rot" fungus; aerobic and degrades both celluloses and lignins; can also degrade polyaromatic hydrocarbons.

‡***Postia placenta* MAS 698** (eukaryote, ~40 Mb): "Brown-rot fungus" degrades cellulose and hemicellulose, secretes oxalic acid, detoxifies certain metals.

‡***Rubrobacter xylanophilus*** (actinobacteria, ~2.6 Mb): Thermophile, highly radioresistant; degrades hemicellulose, xylan.

†***Thermobifida fusca* YX** (bacteria, 3.6 Mb): Major degrader of organic materials.

‡***Trichoderma reesei* RUT-C30, ATCC56765** (fungi, 33 Mb): Efficiently degrades cellulose.

Biotechnology and Applied Microbiology

‡***Acidothermus cellulolyticus* ATCC 43068** (bacteria, ~6 Mb): Thermophile isolated from acid hot spring in Yellowstone; degrades cellulose, source of high-temperature enzymes; biotechnology.

‡***Actinobacillus succinogenes* 130Z (ATCC 55618)** (bacteria, ~2 Mb): From biomass, produces large amounts of succinate; intermediate for production of various chemicals; biotechnology.

****Aquifex aeolicus* VF5** (bacteria extremophile, 1.5 Mb): Potential for identifying high-temperature enzymes.

****Archaeoglobus fulgidus* DSM4304** (archaea extremophile, 2.1 Mb): Potential for identifying high-temperature and high-pressure enzymes; useful in oil industry.

‡***Aspergillus niger*** (fungi, ~32 Mb): Common in soils; model for microbial fermentation and bioproduction of organic acids, enzymes, processing and secretion of proteins; biotechnology.

†***Bifidobacterium longum* DJO10A** (bacteria, 2.1 Mb): Anaerobic, gram-positive prokaryote; key component in promoting healthy human gastrointestinal tract.

†***Brevibacterium linens* BL2** (bacteria, 3 Mb): Applications in industrial production of vitamins, amino acids for fine chemicals, and cheese; survives high salt, carbohydrate starvation, and extended drying conditions.

****Clostridium acetobutylicum*** (bacteria, 4.1 Mb): Produces acetone, butanol, and ethanol; useful for industrial enzymology.

†***Ehrlichia chaffeensis* Sapulpa** (bacteria, 1 Mb): Intracellular, tick-transmitted rickettsia endemic in wild deer populations; causes human monocytic ehrlichiosis.

*****Ehrlichia canis* Jake** (bacteria, 1 Mb): Closely related to *E. chaffeensis*; causes tick-borne disease in dogs (canine monocytic ehrlichiosis).

◊***Gemmata obscuriglobus* UQM 2246** (bacteria, 9 Mb): Planctomycete; widely distributed in freshwater environments; displays a membrane-bound, DNA-containing nucleoid (possibly presaging the nucleus).

****Halobacterium halobium* plasmid** (archaea, 2.3 Mb): Potential for identifying high-salinity enzymes.

‡***Halorhodospira halophila*** (bacteria, ~ 4 Mb): Photosynthetic (fixes carbon dioxide), tolerant of high salt concentrations and high pH; biotechnology.

†***Lactobacillus brevis* ATCC367** (bacteria, 2 Mb): Vital in fermentation of food, feed, and wine.

†***Lactobacillus casei* ATCC334** (bacteria, 2.5 Mb): Used as starter culture in dairy fermentations and for bulk lactic acid production; found in plant, milk, and sourdough environments as well as human intestinal tract, mouth, and vagina.

†***Lactobacillus delbrueckii* bulgaricus ATCCBAA365** (bacteria, 2.3 Mb): Classic example of obligate homofermentative pathway for bulk production of lactic acid.

†***Lactobacillus gasseri* ATCC33323** (bacteria, 1.8 Mb): Naturally inhabits gastrointestinal tract of man and animals. Important for healthy intestinal microflora.

†***Lactococcus lactis cremoris* SK11** (bacteria, 2.3 Mb): Used extensively in food fermentation, especially cheese.

†***Leuconostoc mesenteroides*** (bacteria, 2 Mb): Important role in several industrial and food fermentations.

†***Magnetococcus* MC-1** (bacteria, 4.5 Mb): Requires limited oxygen; reduces iron; produces magnetite, which has many practical commercial uses.

†***Magnetospirillum magnetotacticum* MS-1 ATCC31632** (bacteria, 4.5 Mb): Requires limited oxygen; reduces iron, produces magnetite; possible model for biomineralization and evolutionary responses; may serve as a geomagnetic tracer.

†***Oenococcus oeni* PSU1** (bacteria, 8 Mb): Lactic acid microbe occurring naturally in fruit mashes; used in wineries for fermentation; acid and alcohol tolerant.

†***Pediococcus pentosaceus* ATCC25745** (bacteria, 2 Mb): Gram positive; facultatively anaerobic lactic acid microbe; acid tolerant; used as starter culture in sausage, cucumber, green bean, and soya milk fermentations; ripening agent of cheeses.

‡***Phytophthera ramorum* UCD Pr4** (fungi, 24 to 40 Mb): Pathogen of California oak.

‡***Phytophthora sojae* P6497** (fungi, 62 to 90 Mb): Soybean pathogen.

‡***Psychromonas ingrahamii*** (bacteria, ~ Mb): Grows in Arctic sea ice at –12°C; large, rod shaped; doubles every 10 days; will promote studies of low-temperature enzymes.

*****Pseudomonas syringae* B728a** (bacteria, 5.6 Mb): Pathogenic to a variety of plant species, severely impacting both food and biomass production.

****Pyrobaculum aerophilum*** (archaea extremophile, 2.2 Mb): May identify high-temperature enzymes.

****Pyrococcus furiosus*** (archaea extremophile, 2.1 Mb): May identify high-temperature enzymes.

†***Streptococcus thermophilus* LMD-9** (bacteria, 1.8 Mb): Used as starter in cheese and yogurt fermentations; thermotolerant; noted for exopolysaccharide production.

****Thermotoga maritima* M5B8** (bacteria extremophile, 1.8 Mb): Potential for identifying high-temperature, high-pressure enzymes; metabolizes many simple and complex carbohydrates; possible source of renewable carbon and energy.

Microbial Consortia

‡**Acid mine drainage communities (Iron Mountain, Calif.):** Main site is very acidic (pH <0.5) but geochemically well characterized, with six major species (including *F. acidarmanus*); this site, as well as other nearby sites being sampled, is heavily contaminated with metals; insights into "simple" communities and metal bioremediation.

‡**Active methylotroph community:** Dominant members of Lake Washington, Seattle, one-carbon compound metabolizing bacterial population; carbon cycling, bioremediation.

‡**Anaerobic bioreactor granule samples** (some 200 BACs from Hanford PNNL site): Potential for methane and hydrogen production; exhibit archetypical systems for metabolic-interaction studies among microbes; relatively simple complex microcosm of organic matter's methanogenic degradation in environment.

‡**Boiling thermal pool** (Yellowstone National Park): Characterization of complete communities making up extreme environments; relevance for bioremediation, carbon management.

*****Chlorochromatium aggregatum*** (green sulfur bacteria, plus epibiont, 2 to 10 Mb): Two-component culturable consortium; utilizes hydrogen, sulfur, as electron donors for carbon fixation.

‡**Environmental *Geobacteraceae***: Samples from former uranium mining sites and marine and freshwater sediments.

‡**Microbial population from The Cedars (Calif.)**: Site with pH ~12, low-salt, high-metal concentrations; limited population diversity, high-carbonate deposition; carbon processing.

‡**Obsidian hot spring (Yellowstone)**: Community genomic sampling of microbes from 74°C pool; carbon management, bioremediation.

‡**PAH-degrading mycobacteria:** Mycobacteria from three sites where pollutants (polycyclic aromatic hydrocarbons) are degraded; bioremediation.

‡**Picoplankton BACs** [Hawaii Ocean Time Series (HOTS) site]: Oceanic picoplankton affecting global carbon cycle, energy production, and geochemical and elemental cycling.

‡**Sargasso Sea community**: Catalogue of marine microbial diversity in a low-nutrient environment.

‡**Uncultured microbes in soil environments:** Being sequenced by JGI-Diversa collaboration.

‡**Viruses infecting globally distributed microalgae**: Pathogens of phytoplankton; may regulate phytoplankton populations and therefore carbon-dioxide fixation in oceans.

Technology Development, Pilot Projects

****Borrelia burgdorferi* B31** (bacteria, 1.4 Mb): Human pathogen that causes Lyme disease; one linear chromosome (915 kb) supported by DOE; entire genome published by TIGR.

****Brucella melitensis* 16M** (bacteria, 3.3 Mb): Pathogenic to animals and humans; biothreat agent.

†***Enterococcus faecium*** (bacteria, 2.8 Mb): Pathogenic to many organisms, including humans; tolerates relatively high salt and acid concentrations.

†***Exiguobacterium* 255-15** (bacteria, 3 to 4 Mb) (NASA): Isolated from 2- to 3-million-year-old Siberian permafrost sediment; grows well at –2.5°C; associated with infections in humans.

****Haemophilus somnus* 129PT** (bacteria, ~2.5 Mb): Vaccine strain of *H. somnus,* which causes systemic diseases in cattle; lacks surface-binding protein for immunoglobulins.

****Mycoplasma genitalium* G-37** (bacteria, 580 kb): Human pathogen; serves as model for minimal set of genes sufficient for free living.

****Psychrobacter* 273-4** (bacteria, 2.5 Mb) (NASA): Isolated from 20,000- to 40,000-year-old Siberian permafrost sediment; grows well at –2.5°C; radiation resistant.

†***Streptococcus suis* 1591** (bacteria, 2.2 Mb): Pathogenic to pigs and humans; causes meningitis, especially in tropical regions.

†***Xylella fastidiosa* Dixon** (almond) (bacteria, 2.6 Mb): Pathogenic to economically important plants such as orange and almond trees.

†***Xylella fastidiosa*** Ann1 (oleander) (bacteria, 2.6 Mb): Pathogenic to plants, particularly oleanders.

---

†Draft sequence by the DOE Joint Genome Institute (JGI).

‡New microbes being sequenced by JGI.

◊Sequenced by the The Institute for Genomic Research.

*Completed and published (see www.genomesonline.org).

**Completed, not published (as of July 15, 2005).

APPENDIX G

Appendix H. Programs Complementary to GTL Research

Introduction 266

DOE Office of Science (SC) Programs 267

- **SC Office of Biological and Environmental Research (BER)** 267
- Environmental Remediation Sciences Division (ERSD) 267
- Climate Change Research Division (CCRD) 267
- **SC Office of Basic Energy Sciences (BES)** 267

DOE Office of Fossil Energy (FE) Programs 268

- **Bioprocessing and Biotechnology Research** 268
- **Carbon Sequestration** 268

DOE Office of Energy Efficiency and Renewable Energy (EERE) 268

- **Biomass Program** 268
- **Hydrogen Production** 268

National Science Foundation (NSF) Programs 269

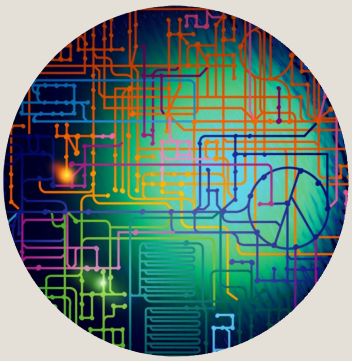
National Institutes of Health (NIH) 270

Department of Agriculture (USDA): Cooperative State Research, Education, and Extension Service (CSREES) 272

Department of Defense (DoD): Defense Advanced Research Projects Agency (DARPA) 272

Other Programs 272

Interagency Cooperation 273



Programs Complementary to GTL Research

Introduction

This section details some governmental research complementary to GTL. Because of the centrality of genomics to the study of all life, GTL can benefit other life sciences programs, and GTL progress can be accelerated by synergies attained through data and resources from others. Although vastly different in focus, scope, and scale of research, the projects listed here are united by their underlying study of DNA and its corresponding technologies. Breakthroughs in one can lead to innovations in another.

DOE already cooperates with other federal agencies on numerous complementary programs. BER coordinates its GTL microbial research with other federal agencies through the National Science and Technology Council's Interagency Working Group on Microbial Genomics. The initial report of this group can be found at www.ostp.gov/html/microbial/start.htm. BER also participates in GTL-related interagency research solicitations that provide additional opportunities to coordinate complementary research programs across agencies. Examples can be viewed on the web (www.nsf.gov/pubs/2004/nsf04607/nsf04607.pdf). Coordination of GTL with those programs may include shared technology development, computing and information tools and standards, databases, and use of resources and facilities.

The information below was taken from web resources and may be incomplete. URLs are provided for further information and exploration. Updates to the online version of this list are encouraged (contact: millsmd@ornl.gov).

DOE Office of Science (SC) Programs

SC Office of Biological and Environmental Research (BER)

www.sc.doe.gov/ober/ober_top.html

BER supports basic biological and environmental research relevant to DOE missions. Biological discoveries are needed to clean and protect the environment, offer new energy alternatives, and understand the impacts of energy use on climate change. BER consists of four divisions: Climate Change Research Division, Environmental Remediation Sciences Division, Life Sciences Division, and Medical Sciences Division. GTL is a Life Sciences Division program jointly supported by BER and the Advanced Scientific Computing Research program in DOE's Office of Science. GTL is complementary to other BER research programs.

Environmental Remediation Sciences Division (ERSD)

www.sc.doe.gov/ober/ersd_top.html

See sidebar, Environmental Remediation Sciences Division Activities Complementary to GTL, p. 223.

- **Natural and Accelerated Bioremediation Research (NABIR) (www.lbl.gov/nabir/)**
  NABIR concentrates on field- and laboratory-based studies of natural microbial communities and their interactions with heavy-metal and radionuclide contaminants. NABIR focuses on understanding and enhancing natural microbial activities that remove contaminants from groundwater and transform them into chemical forms that pose less risk to humans and the environment. Interactions between GTL and NABIR will be important for providing the scientific foundation to develop more effective microbe-based remediation strategies.

Climate Change Research Division (CCRD)

www.sc.doe.gov/ober/ccrd_top.html

CCRD fosters research on understanding the basic chemical, physical, and biological processes of the earth's atmosphere, land, and oceans and how these processes may be affected by energy production and use, primarily the emission of carbon dioxide from fossil-fuel combustion. CCRD modeling aims to quantify sources and sinks of greenhouse gases, especially carbon dioxide; accurately predict and assess the potential consequences of climate change; and evaluate the benefits and costs of alternative response options. GTL-related programs include:

- **Program for Ecosystem Research (PER; http://per.ornl.gov)**
  PER supports research that aims to understand and predict impacts of energy-related environmental changes on the processes and component organisms of terrestrial ecosystems.
- **Ocean Science (OS; www.sc.doe.gov/ober/ccrd/oceans.html)**
  OS focuses on understanding ocean-atmosphere carbon exchange and evaluating ocean-based carbon-sequestration strategies.
- **Terrestrial Carbon Processes (TCP; www.sc.doe.gov/ober/ccrd/tcp.html)**
  TCP's goal is to understand terrestrial carbon cycling and evaluate the potential of long-term carbon sequestration in terrestrial environments.
- **Carbon Sequestration Research (http://cdiac2.esd.ornl.gov)**
  BER's carbon-sequestration research program includes both CCRD and Life Sciences Division research that could lead to strategies to improve the use of trees within the genus *Populus* (poplar) or other trees for long-term sequestration of meaningful amounts of atmospheric carbon in terrestrial ecosystems. Research also emphasizes strategies to use the poplar and microbial genomic sequences to enhance partitioning of carbon into quantitatively important recalcitrant components of trees or soil organic matter that could lead to enhanced carbon sequestration.

SC Office of Basic Energy Sciences (BES)

www.sc.doe.gov/bes/bes.html

BES supports research that provides a scientific foundation for developing new and improved energy technologies and for understanding and mitigating the environmental impacts of energy use. GTL-related programs:

- **Energy Biosciences (EB) (www.sc.doe.gov/bes/eb/ebhome.html)**
  Energy Biosciences is part of the Chemical Sciences, Geosciences, and Biosciences Division

at BES. This program supports basic research to understand the processes of plants and microorganisms that could be used to develop future energy-related biotechnologies. EB emphasizes understanding biological principles rather than optimizing biological processes. Research topics include mechanistic studies of photosynthetic solar-energy capture; mechanisms and regulation of carbon fixation and carbon or energy storage; regulation of plant growth and development; and examination of metabolic pathways relevant to the production of useful chemicals and fuels.

DOE Office of Fossil Energy (FE) Programs

www.fossil.energy.gov

FE supports research and development that address technological challenges of the nation's energy and environmental initiatives. GTL scientific insights could spur R&D in the following FE programs:

Bioprocessing and Biotechnology Research

www.fossil.energy.gov/programs/powersystems/advresearch/advresearch-bioprocessing.html

Bioprocessing and biotechnology activities are part of FE's Advanced Research Programs for Coal and Natural Gas Power Systems. Research is directed toward using biology to develop applications for generating clean, efficient electric power and producing clean fuels from coal. Some research topics include biomodification of coal to reduce mercury emissions; bioremediation of waste streams from power plants; use of microbial toxins to reduce fouling of cooling water intake and discharge systems; investigation of marine microalgae for carbon dioxide biofixation potential; and use of biological systems to produce hydrogen from coal and coal-waste products.

Carbon Sequestration

www.fossil.energy.gov/programs/sequestration/

Carbon Sequestration Core R&D is developing technologies that can capture and permanently store greenhouse gases. GTL will be a scientific foundation for technology development in the following areas:

- **Ocean Sequestration Research (www.fossil.energy.gov/programs/sequestration/ocean/)**
- **Carbon Capture Research (www.fossil.energy.gov/programs/sequestration/capture/)**
- **Terrestrial Sequestration Research (www.fossil.energy.gov/programs/sequestration/terrestrial/)**
- **Novel Carbon Sequestration Concepts (www.fossil.energy.gov/programs/sequestration/novelconcepts/)**

DOE Office of Energy Efficiency and Renewable Energy (EERE)

www.eere.energy.gov

See sidebar, DOE Activities Complementary to GTL Research, p. 203.

Biomass Program

www.eere.energy.gov/biomass

The Biomass Program supports the research and development of advanced technologies that transform biomass into biofuels, biopower, and high-value bioproducts. GTL will play an important role in providing a better understanding of current microbial processes and discovering new microbial capabilities relevant to the Sugar Platform and Products research areas.

Hydrogen Production

www.eere.energy.gov/hydrogenandfuelcells/hydrogen_production.html

Hydrogen Production, within EERE's Hydrogen, Fuel Cells, and Infrastructure Technologies Program, aims to research and develop low-cost, highly efficient hydrogen-production technologies from diverse domestic sources. GTL science could benefit two Hydrogen Production research areas: (1) Biological and Biomass-Based Production, for improving efficiencies of anaerobic fermentation systems; and (2) Photolytic Hydrogen, for photobiological production of hydrogen by green algae.

National Science Foundation (NSF) Programs

www.nsf.gov

- **Biochemical Engineering and Biotechnology (BEB; www.nsf.gov/funding/pgm_summ.jsp?pims_id=13368)**
  BEB supports basic engineering research that aims to understand and achieve quantitative assessments of biomolecular processes (in vivo, in vitro, and ex vivo) that can be used to develop practical biotechnological applications. BEB projects cover a wide range of biotechnological research areas: Fermentation, enzyme studies, recombinant DNA technology, bioprocess control and optimization, metabolic-pathway engineering, cell culturing, tissue engineering, food processing, and relevant information-technology development.
- **Biocomplexity in the Environment (BE): Integrated Research and Education in Environmental Systems (www.nsf.gov/funding/pgm_summ.jsp?pims_id=5532)**
  BE promotes new approaches to studying the dynamic nature of biological systems and their impact on physical and chemical processes of the environment. All environments (including natural ecosystems and agricultural and urban lands) and organisms from microbes to humans fall within the BE framework.
- **Biological Databases and Informatics (BD&I) Program (www.nsf.gov/funding/pgm_summ.jsp?pims_id=5444&org=BIO)**
  BD&I supports new approaches to the management, analysis, and dissemination of biological knowledge that will benefit the scientific community and the general public. BD&I will explore theoretical research on data structures; develop new types of databases with architectures better suited to the complexity of biology; and design easy-to-use interfaces and tools for data analysis and use.
- **Biological Oceanography Program (www.nsf.gov/funding/pgm_summ.jsp?pims_id=11696)**
  This program supports the study of marine organisms and their interactions with each other and with elements in their environment. Subfields in this program include ecosystem and biogeochemical processes; community and population ecology; behavioral, reproductive, and life-history ecology; physiological and chemical ecology; and evolutionary ecology.
- **Biomolecular Systems Cluster (www.nsf.gov/funding/pgm_summ.jsp?pims_id=12771)**
  The Biomolecular Systems Cluster supports research to develop technologies and computational and experimental approaches for the study of biomolecular complexes, mechanistic studies of biomolecular activity, and characterization of higher-order biochemical processes by which organisms acquire and use energy.
- **Catalysis and Biocatalysis (www.nsf.gov/funding/pgm_summ.jsp?pims_id=13360)**
  This program fosters fundamental and applied research in the following areas: Kinetics and mechanisms of chemical reactions important to the production of fuels, chemicals, and specialized materials; characterization of chemical reactions at or near solid surfaces; electrocatalytic processes with industrial or commercial importance; green chemistry or use of biorenewable resources; kinetic modeling and theory of biocatalysis; reactive deposition and processing for thin-film materials; and the use of chemical reaction or transport knowledge to design or control chemical reactors.
- **Cellular Systems Cluster (www.nsf.gov/funding/pgm_summ.jsp?pims_id=12772)**
  This program focuses on the structure, function, and regulation of plant, animal, and microbial cells and their interactions with the environment and with one another. Microbial Observatories (MO) and Microbial Interactions and Processes (MIP) are included in this cluster. MO's goal is to establish a network of sites for observing and understanding microbial diversity in different habitats over long time periods. MIP supports shorter-term, smaller-scale microbial-diversity research that is not site based.
- **Ecological Biology Cluster (EBC; www.nsf.gov/funding/pgm_summ.jsp?pims_id=12823)**
  EBC supports experimental, observational, theoretical, and modeling studies on the structure and function of complex biological associations in natural and managed ecological systems. This program includes the National Center for Ecological Analysis and Synthesis, which analyzes ecological information, tests ecological theories,

examines sociological issues relevant to ecology, supports education and outreach, and informs science policy and management decisions.

- **Ecosystem Science Cluster (ESC) (www.nsf.gov/funding/pgm_summ.jsp?pims_id=12822)**
  ESC supports investigations of terrestrial, aquatic, and wetland ecosystems. Projects that use new or existing quantitative or conceptual models to synthesize and integrate knowledge are encouraged. ESC research includes Ecosystem Studies, which concentrate on whole-system processes and relationships in ecosystems, spanning a wide range of spatial and temporal scales; and Long-Term Ecological Research (LTER), which involves studies at a network of more than a dozen field sites.
- **Environmental Engineering and Technology (www.nsf.gov/funding/pgm_summ.jsp?pims_id=13370)**
  The Environmental Engineering and Technology program supports research on the use of innovative biological, chemical, and physical processes to remediate polluted land, water, and air resources and the development of principles for pollution avoidance.
- **Frontiers in Integrative Biological Research (FIBR; www.nsf.gov/funding/pgm_summ.jsp?pims_id=6188)**
  FIBR supports large interdisciplinary research to seek answers to important, understudied, nondisease-related biological questions. By encouraging research that creatively applies science concepts and strategies with research tools that span a broad range of disciplinary and intellectual boundaries, FIBR supports collaborative projects that may not fit readily into existing programs.
- **Genes and Genome Systems Cluster (GGSC; www.nsf.gov/funding/pgm_summ.jsp?pims_id=12780)**
  GGSC supports research on the genetic mechanisms and genome organization, expression, and regulation of all organisms (prokaryote, eukaryote, phage, and virus).
- **Geobiology and Environmental Geochemistry (GEG; www.nsf.gov/funding/pgm_summ.jsp?pims_id=13410)**
  GEG fosters research on (1) biological factors in geophysical and geochemical processes; (2) rates and mechanisms of inorganic and organic geochemical processes; (3) natural and anthropogenic impacts on biogeochemical cycles; (4) geochemical phenomena, widely ranging spatially from planetary and regional to mineral surface and supramolecular; and (5) development of tools, methods, and models for low-temperature geochemistry and geobiological research. GEG encourages the use of new bioanalytical tools to study terrestrial environments.
- **Instrument Development for Biological Research (IDBR; www.nsf.gov/funding/pgm_summ.jsp?pims_id=9187)**
  IDBR supports research that will develop new and improved instrumentation, software for operating instrumentation, and data-analysis methods to advance the study of biological systems at any level. Proof-of-concept development for entirely novel instrumentation is encouraged.
- **Mathematical Biology (www.nsf.gov/funding/pgm_summ.jsp?pims_id=5690)**
  This program supports research on mathematics important to the biological sciences that does not involve statistics or probability. NSF programs in statistics and probability may include research specific to other areas of science and engineering.
- **Research in Biogeosciences 2005 (BioGeo; www.nsf.gov/funding/pgm_summ.jsp?pims_id=5508)**
  BioGeo fosters research that explores the interactions of microbes with earth materials (including minerals, rocks, hydrates, soils, and dust). Research should elucidate past and present roles of microbial communities in earth processes, microbial strategies for deriving energy and nutrients, and how these strategies alter earth materials and the environment.

National Institutes of Health (NIH)

www.nih.gov

- **Bioinformatics and Computational Biology (BCB) Roadmap Initiatives (www.nihroadmap.nih.gov/bioinformatics/)**
  Four National Centers for Biomedical Computing (NCBC) established in 2004 are the key programmatic initiatives of the NIH BCB Roadmap. These centers aim to develop and implement the core of a universal computing infrastructure urgently needed to speed progress in biomedical research.

The centers will create innovative software programs and other tools that will enable the biomedical community to integrate, analyze, model, simulate, and share data on human health and disease.

- **Biomedical Information Science and Technology Initiative (BISTI; www.bisti.nih.gov)**
  Launched in 2000, BISTI's goal is to make optimal use of computer science and technology to address problems in biology and medicine. A BISTI consortium serves as the focus of biomedical computing issues at NIH and facilitates implementation of BISTI recommendations. The consortium is composed of representatives from NIH centers and institutes and other federal agencies concerned with bioinformatics and computational applications. The consortium's mission is to make optimal use of computer science and technology to address problems in biology and medicine by fostering new basic understandings, collaborations, and initiatives between the disciplines of computational and biomedical sciences.
- **Complex Biological Systems Initiative (www.nigms.nih.gov/funding/complex_summary.html)**
  This National Institute for General Medical Sciences (NIGMS) program promotes quantitative, interdisciplinary approaches to problems of biomedical significance, particularly those that involve the complex, interactive behavior of many components. Three classes of initiatives are supported: Interdisciplinary research to attract investigators trained in mathematically based disciplines to the study of biomedical problems; mechanisms to train biomedical scientists in quantitative approaches and to acquaint nonbiologists with biological problems; and interdisciplinary training for scientists at the pre- and postdoctoral levels.
- **National Institute of Allergy and Infectious Diseases (NIAID) Microbial Sequencing Centers (www.niaid.nih.gov/dmid/genomes/mscs/overview.htm)**
  NIAID's Microbial Sequencing Centers (MSCs) sequence microorganisms and invertebrate vectors of disease that are considered agents of bioterrorism or responsible for emerging and reemerging diseases.
- **National Technology Centers for Networks and Pathways (www.nihroadmap.nih.gov/buildingblocks/)**
  As part of the NIH Roadmap for Medical Research, two National Technology Centers for Networks and Pathways were established and several more are planned. The primary goal of these centers is to develop new technologies to study the dynamics of molecular interactions within cells. Such capabilities are crucial for expanding the identification of biological pathways and developing treatments for diseases involving such pathways. The awards are administered by the National Center for Research Resources, an NIH component that supports primary research to create and develop critical resources, models, and technologies.
- **Protein Structure Initiative (PSI; www.nigms.nih.gov/psi/)**
  PSI is a 10-year project funded largely by NIGMS to determine the three-dimensional (3D) shapes of a wide range of proteins. These structures are expected to shed light on protein function in many life processes and could lead to the development of new medicines. The long-range goal of PSI is to make 3D atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences. The first half of this project—a pilot phase that started in 2000—has centered on developing new tools and processes that enable researchers to determine quickly, cheaply, and reliably the shapes of many proteins found in nature. PSI projects are in federal, university, and industry laboratories.
- **Systems Biology Initiative (SBI; www.nigms.nih.gov/funding/systems.html)**
  This NIGMS program supports systems biology research for areas central to its mission of supporting basic biomedical research and developing new computational approaches to biomedical complexity. SBI's goal is to establish national centers for systems biology that develop pioneering research, training, and outreach programs focused on quantitative, systems-level analysis of biomedically important phenomena within the NIGMS mission. High priority is given to projects that integrate multi-investigator, multidisciplinary approaches with a high degree of interplay between computational and experimental approaches. Innovation is critical for design of

both research projects and infrastructure with the mission of serving communities beyond participating investigators, institutions, and collaborators.

Department of Agriculture (USDA): Cooperative State Research, Education, and Extension Service (CSREES)

www.csrees.usda.gov/about/about.html

- **Agricultural Plant Biochemistry (www.csrees.usda.gov/fo/fundview.cfm?fonum=1115)**
  The aim is to characterize the biochemical processes and pathways in the cell and the genes and proteins involved in them.
- **Biobased Products and Bioenergy Production Research Program (www.csrees.usda.gov/fo/fundview.cfm?fonum=1073)**
  This program supports the tripling of U.S. use of biobased products by 2010 and more research on biomass processing and conversion.
- **Biology of Plant-Microbe Associations (www.csrees.usda.gov/fo/fundview.cfm?fonum=1120)**
  This program supports fundamental and mission-linked research on interactions among plants and their associated microbes, including fungi and fungal-like microbes, bacteria, viruses, viroids, and mycoplasma-like organisms.

Department of Defense (DoD): Defense Advanced Research Projects Agency (DARPA)

www.darpa.mil/

- **BioCOMP Program (www.darpa.mil/ipto/programs/biocomp/)**
  BioCOMP develops a computational framework to enable construction of sophisticated models of intracellular processes that can be used to predict and control the behavior of living cells. In addition, BioCOMP generates new computational paradigms and engineering applications that use biomolecules as information-processing, sensing, or structural components.
- **Biological Input/Output Systems Program (BIOS; www.darpa.mil/dso/thrust/biosci/bios.htm)**
  BIOS will develop robust technologies for designing DNA-encoded "plug-and-play" modules that will enable use of organisms (e.g., plants, microbes, lower eukaryotes) as remote sentinels for reporting the presence of chemical or biological analytes.
- **BioSPICE (https://users.biospice.org/)**
  BioSPICE is a set of open-source software tools that can be used by biological researchers to model the processes of living cells. It is being used to study several different biological systems: Bacterial systems to investigate such phenomena as sporulation, chemotaxis, and bacterial metabolism; viral systems to understand Lambda-phage, HIV-1, and host-pathogen interactions; eukaryotic systems to model cell cycles, cellular differentiation, immunological function, and cell signaling; and synthetic systems such as minimal cells. Some mathematical models developed using BioSPICE include pathway and interaction networks, models of gene expression, and probabilistic modeling for sequence analysis. BioSPICE is the product of a collaboration involving DARPA, NSF, academic institutions, and other federal agencies.
- **Defense Against Chemical, Biological, Radiological Weapons Program (DACBRW; www.darpa.mil/ spo/programs/cbr.htm)**
  This program seeks to protect building inhabitants from an indoor release of chemical or biological agents and from radiological attack. DACBRW includes research into sensing of bioaerosols and triangulation identification for genetic evaluation of risks (biosensors).

Other Programs

- **HUPO Proteomics Standards Initiative (http://psidev.sourceforge.net/)**
- **National Aeronautics and Space Administration (NASA) Ames Genome Research Facility (www.phenomorph.arc.nasa.gov)**
  Research at the facility includes the Nanopore Project and functional genomics.
- **NASA Fundamental Space Biology Program (FSB; www.fundamentalbiology.arc.nasa.gov)**
  FSB has increased emphasis on cell and molecular biology and developmental biology, as well as on the growing disciplines of evolutionary biology and genomics. Part of the program's purpose is to

increase visibility and funding for molecular biology research.

- **National Institute of Standards and Technology (NIST) Biotechnology Division (www.cstl.nist.gov/biotech/)**
  The mission of the NIST Biotechnology program is to advance the commercialization of biotechnology by developing the scientific and engineering technical base, reliable measurements, standards, data, and models to enable U.S. industry to quickly and economically produce biochemical products with appropriate quality control. The division is organized into four groups: DNA Technologies; Bioprocess Measurements; Structural Biology; and Cell and Tissue Measurements.
- **National Oceanic and Atmospheric Administration (NOAA) Office of Global Programs (OGP) Climate and Global Change Program (www.ogp.noaa.gov)**
  OGP assists NOAA by sponsoring scientific research aimed at understanding climate variability and predictability. Through studies in these areas, researchers coordinate activities that jointly contribute to improved predictions and assessments of climate change over a continuum of time scales from season to season, year to year, and throughout a decade and beyond.
- **United Nations Educational, Scientific and Cultural Organization (UNESCO) Microbial Resources Centres (MIRCEN; www.portal.unesco.org/sc_nat/ev.php?URLID=2491&URL_DO=DO_TOPIC)**
  MIRCEN comprises 34 academic and research institutes in developed and developing countries involved in a global collaborative effort to harness microbiological research and biotechnological applications for the benefit of humankind. The global MIRCEN network's research and training activities aim to (1) provide a global infrastructure incorporating national, regional, and international cooperating laboratories geared to the management, distribution, and use of the microbial gene pool; (2) reinforce use of the rhizobial gene pool in developing countries with an agrarian base; (3) foster development of new inexpensive technologies native to specific regions; (4) promote economic and environmental applications of microbiology; and (5) serve as the network's focal centers for training.
- **U.S. Geological Survey (USGS) Biological Resource Division (http://biology.usgs.gov)**
  The USGS Biological Resource Division works with others to provide the scientific understanding and technologies needed to support the sound management and conservation of U.S. biological resources. USGS is committed to data and information sharing and has established the National Biological Information Infrastructure, a network of distributed databases and information sources on biological resources.

Interagency Cooperation

For more information on the following programs, see the web site for current and archived solicitations after each entry.

- **Environmental Molecular Science Institutes (EMSI; www.nsf.gov/funding/pgm_summ.jsp?pims_id=5294)**
  EMSI is a partnership between NSF and DOE for collaborative, interdisciplinary research to attain a fundamental, molecular-level understanding of natural and anthropogenic processes in the environment. An institute typically supports a group of six or more investigators from academic institutions, nonprofit organizations, industry, or national laboratories with complementary research interests.
- **Interagency Microbial Genome Sequencing Program (USDA with NSF) (www.csrees.usda.gov/fo/fundview.cfm?fonum=1108)**
  This program supports high-throughput sequencing of the genomes of a wide range of microorganisms (including viruses, bacteria, archaea, fungi, oomycetes, protists, and agriculturally important nematodes).
- **Interagency Modeling and Analysis Group (No web site available; see last call for proposals at www.nsf.gov/pubs/2004/nsf04607/nsf04607.pdf)**
  This group is a collaboration among NSF, NIH, NASA, and DOE to encourage the integrative systems engineering approach to multiscale modeling, combining theoretical and computational approaches. This collaboration aims to formulate and validate novel computational and statistical methods and relationships for spanning multiple scales, broaden and expand currently established levels of modeling expertise and

multiscale modeling activities, and produce models of practical utility to the community at large. The group also plans to form a consortium of investigators for information exchange on critical issues including model intraoperability and evaluation and open-source software sharing.

- **Joint DMS/BIO/NIGMS Initiative to Support Research in the Area of Mathematical Biology (www.nsf.gov/funding/pgm_summ.jsp?pims_id=5300)**
  This initiative supports research on mathematical and statistical problems related to the biological sciences, including conferences, educational research experiences, postdoctoral research fellowships, and acquisition of computational equipment. It involves the NSF Directorate for Mathematical and Physical Sciences' Division of Mathematical Sciences, the NSF Directorate for Biological Sciences, and NIH NGMIS.

- **Mathematical Sciences: Innovations at the Interface of the Sciences and Engineering (www.nsf.gov/funding/pgm_summ.jsp?pims_id=9673)**
  The Mathematical Sciences Priority Area (MSPA) cuts across multiple NSF directorates and supports the integration of mathematical and statistical research with a wide range of science disciplines. Initially, MPSA interdisciplinary projects are focusing on mathematical challenges associated with handling large data sets, managing and modeling uncertainty, and modeling complex nonlinear systems.

- **Metabolic Engineering Working Group (MEWG; www.metabolicengineering.gov)**
  MEWG is a collaboration among eight agencies and departments to provide research funding and agency in-kind support (e.g., equipment, lab space, and materials) to gain a better understanding of metabolic pathways and metabolic engineering in living systems. Conceptual and technical approaches necessary to understand the integration and control of genetic, catalytic, and transport processes will be valuable as fundamental research and also will provide the underpinning for many applications of immediate value. Participating institutions include the Department of Commerce, Environmental Protection Agency, DoD, DOE, NASA, NIH, NIGMS, NSF, and USDA.

- **Microbial Genome Sequencing Program (www.nsf.gov/funding/pgm_summ.jsp?pims_id=5688)**
  This program is a collaboration between NSF and USDA CSREES. It supports high-throughput genome sequencing of microorganisms (e.g., viruses, bacteria, archaea, fungi, oomycetes, protists, and agriculturally important nematodes) that have fundamental biological interest or relevance to such national priorities as productivity and sustainability of agricultural and natural resources and food-supply safety and quality.

- **Nanoscale Science and Engineering (www.nsf.gov/funding/pgm_summ.jsp?pims_id=7169)**
  This program brings together engineering and various scientific disciplines to advance the emerging field of nanotechnology. Research areas include nanoscale investigations of biosystems; environmental processes; nanostructures and devices; multiscale multiphenomena theory, modeling, and simulation; and manufacturing processes; as well as research on societal and educational implications of nanoscale research and technology development.

- **National Science and Technology Council Interagency Working Group on Microbial Genomics (www.ostp.gov/html/microbial/aboutus.htm)**
  This working group consists of representatives from all federal agencies that support or conduct microbial research.

References

Note: Additional citations are included in the sidebars in this document.

Abraham, S. 2004. "The Bush Administration's Approach to Climate Change," *Science* **305**, 616–17.

Aebersold, R., and J. D. Watts. 2002. "The Need for National Centers for Proteomics," *Nat. Biotechnol.* **20**, 65.

Annual Energy Outlook 2005 with Projections to 2025, DOE/EIA-0383. 2005. Energy Information Administration, U.S. Department of Energy (www.eia.doe.gov/oiaf/aeo).

Appella, E., and C. W. Anderson, eds. Accepted for publication in *FEBS. J.,* Fall 2005. Protein Interaction Minireview Series derived from 15th Conference on Methods in Protein Structure Analysis (MPSA2004).

- Altschul, S. F., et al. "Protein Database Searches Using Compositionally Adjusted Substitution Matrices."
- Bertone, P., and M. Snyder. "Advances in Functional Protein Array Technology."
- Bowers, P. M., et al. "Utilizing Logical Relationships in Genomic Data to Decipher Cellular Processes."
- Dunker, A. K., et al. "Flexible Nets: The Roles of Intrinsic Disorder in Protein Interaction Networks."
- Field, S. "High-Throughput Two Hybrid Analysis: The Promise and the Peril."
- Houtman, J. C. D., M. Barda-Saad, and L. E. Samelson. "Examining Multiprotein Signaling Complexes from All Angles: Use of Complementary Techniques to Characterize Complex Formation at the Adapter Protein LAT."
- Noble, W. S., et al. "Identifying Remote Protein Homologs by Network Propagation."
- Ramachandran, N., et al. "Emerging Tools for Real-Time Label-Free Detection of Interactions on Functional Protein Microarrays."

Armbrust, E. V., et al. 2004. "The Genome of the Diatom *Thalassiosira pseudonana:* Ecology, Evolution, and Metabolism," *Science* **306**, 79–86.

Belkin, S. 2003. "Microbial Whole-Cell Sensing Systems of Environmental Pollutants," *Curr. Opin. Microbiol.* **6**, 206–12.

Ben-Ari, E. T. 2002. "Microbiology and Geology: Solid Marriage Made on Earth," *ASM News* **68**(1), 13–17.

Beyenal, H., C. C. Davis, and Z. Lewandowski. 2004. "An Improved Severinghaus-Type Carbon Dioxide Microelectrode for Use in Biofilms," *Sens. Actuators B Chem.* **97**, 202–10.

Beyenal, H., et al. 2004. "Uranium Immobilization by Sulfate-Reducing Biofilms," *Environ. Sci. Technol.* **38**(7), 2067–74.

Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply. 2005. U.S. Department of Agriculture and U.S. Department of Energy.

Bioremediation of Metals and Radionuclides: What It is and How It Works, 2nd ed., 2003. Natural and Accelerated Bioremediation Research Program, U.S. Department of Energy (www.lbl.gov/NABIR).

Bioventing Performance and Cost Results from Multiple Air Force Test Sites. 1996. Prepared by Parsons Engineering Science for the Air Force Center for Environmental Excellence, Brooks Air Force Base, Texas.

Brady, S. F., and J. Clardy. 2000. "Long-Chain N-Acyl Amino Acid Antibiotics Isolated from Heterologously Expressed Environmental DNA," *J. Am. Chem. Soc.* **122**(51), 12903–4.

Brady, S. F., C. J. Chao, and J. Clardy. 2002. "New Natural Product Families from an Environmental DNA (cDNA) Gene Cluster," *J. Am. Chem. Soc.* **124**, 9968–69.

Buckley, M. R. 2004a. *The Global Genome Question: Microbes as the Key to Understanding Evolution and Ecology,* American Society for Microbiology.

Buckley, M. R. 2004b. *Systems Microbiology: Beyond Microbial Genomics,* American Society for Microbiology.

Carbon Sequestration Research and Development. 1999. Office of Science Office of Fossil Energy, U.S. Department of Energy.

Check, B. 2002. "AAM Sees Bright Prospects for Microbial Ecology Research in the Genomic Era," *ASM News* **68**(9), 427–31.

Climate Change Science Program (CCSP). 2003. *Strategic Plan for the U.S. Climate Change Science Program,* U.S. Climate Change Science Program and Subcommittee on Global Change Research (www.climatescience.og/Library/stratplan2003/final/default.htm).

Climate Change Technology Program (CCTP) (www.climatetechnology.gov)

Closure Planning Guidance. 2004. Office of Environmental Management, U.S. Department of Energy (http://web.em.doe.gov/program.html).

Critical Choices: Science, Energy, and Security: Final Report of the Secretary of Energy Advisory Board's Task Force on the Future of Science Programs at the Department of Energy. 2003. Secretary of Energy Advisory Board, U.S. Department of Energy (www.seab.energy.gov/publications/FSPFinalDraft.pdf).

Croal, L. R., et al. 2004. "The Genetics of Geochemistry," *Annu. Rev. Genet.* **38**, 175–202.

Demain, A., et al. 2005. "Cellulase, Clostridia, and Ethanol," *Microbiol. Mol. Biol. Rev*. **69**, 124–54.

The Department of Energy Strategic Plan, DOE/ME-0030. 2003. Office of Program Analysis and Evaluation, Office of Management, Budget, and Evaluation, U.S. Department of Energy (strategicplan.doe.gov).

Doney, S. C., et al. 2004. "From Genes to Ecosystems: The Ocean's New Frontier," *Front. Ecol. Environ.* **2**(9), 457–66.

Dunker, A. K., et al. 1998. "Protein Disorder and the Evolution of Molecular Recognition: Theory, Predictions, and Observation," *Pac. Symp. Biocomp.* **3**, 473–84.

Edmonds, J. A., et al. 2003. "The Potential Role of Biotechnology in Addressing the Long-term Problem of Climate Change in the Context of Global Energy and Economic Systems," *Greenhouse Gas Control Technologies: Proceedings of the Sixth International Conference on Greenhouse Gas Control Technologies, 1–4 October 2002, Kyoto, Japan*, ed. J. Gale and Y. Kaya (Pergamon, Amsterdam, Netherlands), 1427–32.

Edmonds J., et al. 2004. "Stabilization of CO_2 in a B2 World: Insights on the Roles of Carbon Capture and Disposal, Hydrogen, and Transportation Technologies," *Energy Econ.* **26**(4), 517–37.

Elowitz, M. B., et al. 2002. "Stochastic Gene Expression in a Single Cell," *Science* **297**, 1183–86.

Ellis, D. I., et al. 2004. "From Genomes to Systems: A Report on the 2nd Conference of the Consortium for Post-Genome Science (CPGS) 'Genomes to Systems,' Manchester, U.K., 1–3 September 2004," *Genome Biol.* **5**, 354.

Endo,Y., and T. Sawasaki. 2003. "High-Throughput, Genome-Scale Protein Production Method Based on the Wheat Germ Cell-Free Expression System," *Biotechnol. Adv.* **21**(8), 695–713.

Falciatore, A., and C. Bowler. 2002. "Revealing the Molecular Secrets of Marine Diatoms," *Annu. Rev. Plant Biol.* **53**, 109–30.

Falkowski, P., et al. 2000. "The Global Carbon Cycle: A Test of Our Knowledge of Earth as a System," *Science* **290**, 291–96.

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

Field, C. B., et al. 1998. "Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components," *Science* **281**, 237–81.

Finelli, A., et al. 2003. "Use of In-Biofilm Expression Technology to Identify Genes Involved in *Pseudomonas aeruginosa* Biofilm Development," *J. Bacteriol.* **185**, 2700–2710.

Frazier, M. E., et al. 2003a. "Realizing the Potential of the Genome Revolution: The Genomes to Life Program," *Science* **300**, 290–93.

Frazier, M. E., et al. 2003b. "Stepping Up the Pace of Discovery: The Genomes to Life Program," *Proc. IEEE Comput. Soc. Bioinform. Conf. (CSB '03).*

Fredrickson, J. K., and D. L. Balkwill. 2005. "Geomicrobial Processes and Biodiversity in the Deep Terrestrial Subsurface," *Geomicrobiol. J.*, in press.

Fuhrman, J. 2003. "Genome Sequences from the Sea," *Nature* **424**, 1001–2.

Gaietta, G., et al. 2002. "Multicolor and Electron Microscopic Imaging of Connexin Trafficking," *Science* **296**(5567), 503–7.

Ghirardi, M. L., et al. 2000. "Microalgae: A Green Source of Renewable H_2," *Trends Biotechnol*. **18**, 506–11.

Gold, T. 1992. "The Deep, Hot Biosphere," *Proc. Natl. Acad. Sci.* **89**, 6045–49.

Greene, N., et al. 2004. *Growing Energy: How Biofuels Can Help End America's Oil Dependence*, Natural Resources Defense Council, New York.

Handelsman et al. 1998. "Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products," *Chem. Biol.* **5**, R245–49.

Herrera, S. June 2004. "Industrial Biotechnology—A Chance at Redemption," *Nat. Biotechnol.* **22**(6), 671–75.

Hess, W. R. 2004. "Genome Analysis of Marine Photosynthetic Microbes and Their Global Role," C*urr. Opin. Biotechnol.* **15**(3), 191–98.

Homegrown for the Homeland: Ethanol Industry Outlook 2005. 2005. Renewable Fuels Association (www.ethanolrfa.org/outlook2005.html).

Houghton, J. T., et al., eds. 1995. *Climate Change 1995: The Science of Climate Change: Contribution of Working Group I to the Second Assessment of the Intergovernmental Panel on Climate Change,* Cambridge University Press, U.K.

The Hydrogen Economy: Opportunities, Costs, Barriers, and R&D Needs. 2004. National Research Council and National Academy of Engineering, National Academies Press, Washington, D.C.

International Energy Outlook, DOE/EIA-0484. 2004. Energy Information Administration, U. S. Department of Energy (www.eia.doe.gov/oiaf/ieo).

Johnston, C. A., et al. 2004. "Carbon Cycling in Soil," *Front. Ecol. Environ.* **2**(10), 522–28.

Kawasaki, T., et al. 2003. "Efficient Synthesis of a Disulfide-Containing Protein Through a Batch Cell-Free System from Wheat Germ," *Eur. J. Biochem.* **270**(23), 4780–86.

Keller, M., and K. Zengler. Feb. 2004. "Tapping into Microbial Diversity," *Nat. Rev. Microbiol.* **2**, 141–50.

Kigawa, T., et al. 1999. "Cell-Free Production and Stable-Isotope Labeling of Milligram Quantities of Proteins," *FEBS Lett.* **442**, 15.

King, G. M., et al. 2001. *Global Environmental Change: Microbial Contributions, Microbial Solutions,* American Society for Microbiology.

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

Klaper, R., and M. A. Thomas. 2004. "At the Crossroads of Genomics and Ecology: The Promise of a Canary on a Chip,"*BioScience* **54**(5), 403–12.

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

Levin, D. B., L. Pitt, and M. Love. 2004. "Biohydrogen Production: Prospects and Limitations to Practical Application," *Int. J. Hydrogen Energy* **29**, 173–85.

Linking Legacies: Connecting the Cold War Nuclear Weapons Production Processes to Their Environmental Consequences. 1997. Office of Environmental Management, U.S. Department of Energy.

Lipton, M. S., et al. 2002. "Global Analysis of the *Deinococcus radiodurans* Proteome by Using Accurate Mass Tags,"*Proc. Natl. Acad. Sci*. **99**(17), 11049–54.

Littlehales, C. 2004. "Industrial Biotech Takes Center Stage at World Congress in 2004: Summary Article on World Congress 2004," Biotechnology Industry Organization (bio.org/worldcongress/media/20040426.asp).

Luengo, J. M., et al. 2003. "Bioplastics from Microorganisms," *Curr. Opin. Microbiol*. **6**, 251–60.

Madamwar, D., N. Garg, and V. Shah. 2000. "Cyanobacterial Hydrogen Production," *World J. Microbiol. Biotechnol*. **16**, 757–67.

Madsen, E. L. 2005. "Identifying Microorganisms Responsible for Ecologically Significant Biogeochemical Processes," *Nat. Rev. Microbiol.* **3**, 439–46.

Majdalani, N., C. K. Vanderpool, and S. Gottesman. 2005. "Bacterial Small RNA Regulators," *Crit. Rev. Biochem Mol. Biol.* **40**, 93–113.

Mann, C. 2004. "Ethanol from Biomass," Memorandum to the National Commission on Energy Policy, Appendix IV.4e of NCEP report, *Ending the Energy Stalemate: A Bipartisan Strategy to Meet America's Energy Challenges*.

Marburger, J. H., III, and J. B. Bolten. 2004. "Memorandum for the Heads of Executive Departments and Agencies," M-04-23, Executive Office of the President, Washington, D.C.

Martinez, D., et al. 2004. "Genome Sequence of the Lignocellulose Degrading fungus *Phanerochaete chrysosporium strain RP78," Nat. Biotechnol.* **22**(6), 695–700.

Melis, A., and T. Happe. 2001. "Hydrogen Production. Green Algae as a Source of Energy," *Plant Physiol*. **127**, 740–48.

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

Moller, G. M., et al. 1998. "In Situ Gene Expression in Mixed-Culture Biofilms: Evidence of Metabolic Interactions Between Community Members," *Appl. Environ. Microbiol.* **64**, 721–32.

Nakicenovic, N., et al. 2000. *Special Report on Emissions Scenarios,* ed. N. Nakicenovic and R. Swart, Cambridge University Press, New York (www.grida.no/climate/ipcc/emission/index.htm).

Nass, S. J., and B. Stillman. 2003. *Large-Scale Biomedical Science: Exploring Strategies in Future Research,* National Academies Press, Washington, D.C.

Nath, K., and D. Das. 2004. "Biohydrogen Production as a Potential Energy Resource: Present State-of-Art," *J. Sci .Ind. Res.* **63**, 729–38.

National Biodiesel Board (www.biodiesel.org)

National Hydrogen Energy Roadmap. 2002. Office of Energy Efficiency and Renewable Energy, U.S. Department of Energy (www.eere.energy.gov/hydrogenandfuelcells/pdfs/national_h2_roadmap.pdf).

Nealson, K. H. 2005. "Hydrogen and Energy Flow as 'Sensed' by Molecular Genetics," *Proc. Natl. Acad. Sci.* **102**(11), 3889–90.

Nicholson, J. K., et al. 2002. "Metabonomics: A Platform for Studying Drug Toxicity and Gene Function," *Nat. Rev. Drug Discov.* **1**, 153–61.

O'Toole, G. A. 2003. "To Build a Biofilm," *J. Bacteriol.* **185**(9), 2687–89.

Pacala, S., and R. Socolow. 2004. "Stabilization Wedges: Solving the Climate Problem for the Next 50 Years with Current Technologies," *Science* **305**, 968–72.

Patrinos, A. 2005. "Biotechnology Reenergized: The Goals and Promise of Genomes to Life Program Have Energy and Environmental Applications," *The Scientist* **19**, 20.

Pennisi, E. 2003. "Tracing Life's Circuitry," *Science* **302**, 1646–49.

Prince, R. C., and H. S. Kheshgi. 2005. "The Photobiological Production of Hydrogen: Potential Efficiency and Effectiveness as a Renewable Fuel," *Crit. Rev. Microbiol.* **31**, 19–31.

Reliable, Affordable, and Environmentally Sound Energy for America's Future: Report of the National Energy Policy Development Group. 2001. National Energy Policy Development Group, Washington, D.C. (www.whitehouse.gov/energy/).

Report on the Imaging Workshop for the Genomes to Life Program, Charlotte, North Carolina, April 16–18, 2002, DOE/SC-0066. 2002. Prepared by the Office of Advanced Scientific and Computing Research and Office of Biological and Environmental Research, Office of Science, U.S. Department of Energy (www.doegenomestolife.org/technology/imaging/workshop2002/).

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

Relman, D. A., and E. Strauss. 2000. *Microbial Genomes: Blueprints for Life,* American Society for Microbiology.

Roberts, R. J. 2004. "Identifying Protein Function—A Call for Community Action," *PLoS Biology* **2**(2), 1.

Roberts, R. J., et al. 2004. *An Experimental Approach to Genome Annotation,* American Society for Microbiology.

Romero, P., et al. 1998. "Thousands of Proteins Likely to Have Long Disordered Regions," *Pac. Symp. Biocomp*. **3**, 437–48.

Rosenberg, N. J., F. B. Metting, and R. C. Izaurralde, eds. 2004. *Applications of Biotechnology to Mitigation of Greenhouse Warming: Proceedings of the St. Michaels II Workshop April 2003,* Battelle Press, Columbus, Ohio.

Rosenberg, N. J., R. C. Izaurralde, and E. L. Malone, eds. 1999. *Carbon Sequestration in Soils: Science, Monitoring, and Beyond. Proceedings of the St. Michaels Workshop December 1998,* Battelle Press, Columbus, Ohio.

Saha, B. C. 2004. "Lignocellulose Biodegradation and Applications in Biotechnology," pp. 2–34 in *Lignocellulose Biodegradation*, American Chemical Society, Washington, D.C.

Sawasaki, T., et al. 2002. "A Cell-Free Protein Synthesis System for High-Throughput Proteomics," *Proc. Natl. Acad. Sci. USA.* **99**(23), 14652–57.

Schaechter, M., R. Kolter, and M. Buckley. 2004. *Microbiology in the 21st Century: Where are We and Where Are We Going?* American Society for Microbiology.

Schaechter, M., R. Kolter, and S. Maloy. 2005. "Microbiology Happens," *ASM News* **71**(2), 54–55.

Schloss, P. D., and J. Handelsman. 2003. "Biotechnological Prospects from Metagenomics," *Curr. Opin. Biotechnol.* **14**, 303–10.

Scott, M. J., et al. 1998. "Research Investment Pays Off: Subsurface Barrier Technology Results in Cost Savings," *Soil and Groundwater Cleanup* (Oct. 6–13, 1998).

Smith, H. O., et al. 2003. "Generating a Synthetic Genome by Whole Genome Assembly: ΦX174 Bacteriophage from Synthetic Oligonucleotides," *Proc. Natl. Acad. Sci.* **100**(26), 15440–45.

Smith, S. J., et al. 2004. *Near-Term US Biomass Potential: Economics, Land-Use, and Research Opportunities,* PNWD-3285, Battelle Memorial Institute, Joint Global Change Research Institute, Baltimore, Md.

Socolow, R. H. July 2005. "Can We Bury Global Warming?" *Sci. Am.* **293**(1), 49–55.

Spear, J. R., et al. 2005. "Hydrogen and Bioenergetics in the Yellowstone Geothermal Ecosystem," *Proc. Natl. Acad. Sci.* **102**(7), 2555–60.

Stahl, D. A., and J. M. Tiedje. 2002. *Microbial Ecology and Genomics: A Crossroads of Opportunity,* American Society for Microbiology.

Staley, J. T., et al. 1997. *The Microbial World: Foundation of the Biosphere,* American Society for Microbiology.

Stein, J. L., et al. 1996. "Characterization of Uncultivated Prokaryotes: Isolation and Analysis of a 40-Kilobase-Pair Genome Fragment from a Planktonic Marine Archaeon," *J. Bacteriol.* **178**, 591–99.

Tamagnini, P., et al. 2002. "Hydrogenase and Hydrogen Metabolism of Cyanobacteria," *Microbiol. Mol. Biol. Rev.* **66**, 1–20.

Toner, B., et al. 2005. "Spatially Resolved Characterizion of Biogenic Manganese Oxide Production Within a Bacterial Biofilm," *Appl. Envir. Microbiol.* **71**(3), 1300–1310.

Tringe, S. G., et al. 2005. "Comparative Metagenomics of Microbial Communities," *Science* **308**, 554–57.

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

U.S. Congress. 2000. Biomass Research and Development Act of 2000, H.R. 2559.

Venter, J. C., et al. 2004. "Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science* **304**, 66–74.

Vision for Bioenergy and Biobased Products in the United States. 2002. Biomass Research and Development Technical Advisory Committee (www.bioproducts-bioenergy.gov/pdfs/BioVision_03_Web.pdf).

Winzer, K., K. R. Hardie, and P. Williams. 2002. "Bacterial Cell-to-Cell Communication: Sorry, Can't Talk Now—Gone to Lunch!" *Curr. Opin. Microbiol.* **5**, 216–22.

Wolfaardt, G. M., et al. 1994. "Multicellular Organization in a Degradative Biofilm Community," *Appl. Environ. Microbiol.* **60**, 434–46.

Zengler et al., 2002. "Cultivating the Uncultured," *Proc. Natl. Acad. Sci. USA* **99**(24), 15681–86.

Glossary

1D gel (one-dimensional gel electrophoresis): See *electrophoresis*.

2D gel (two-dimensional electrophoresis or 2DE): See *electrophoresis*.

16S rRNA: RNA molecule (about 1500 nucleotides long) that combines with proteins to form the small subunit of the ribosome in prokaryotes. The gene for 16S rRNA is well studied, highly conserved, and present in all prokaryotic organisms, so variations in this gene's sequence can be used to determine relatedness (phylogenetic linkages) among prokaryotes. In eukaryotes, the small subunit of the ribosome contains an RNA molecule called 18S rRNA (about 1900 nucleotides long), which is analyzed to determine phylogenetic relationships.

ab initio: Type of method for predicting protein structure using first principles of physics and chemistry rather than comparisons with known homologous structures. Also used to describe gene-prediction methods based on analyzing the composition of raw genome sequence rather than using comparisons with other sequence data.

accurate mass and time tag (AMT): Peptide for which the liquid chromatography elution time and mass has been measured so accurately that it can be identified uniquely among all possible peptides predicted from a genome. In proteome mass spectrometry (MS), the proteome is digested enzymatically to produce protein fragments or peptides before MS analysis. A database of AMT mass data can be used to identify peptides in other samples analyzed by MS.

activator: Regulatory protein that binds the operator site and enhances transcription of genes in an operon. See also *repressor*.

adenine (A): Nitrogenous base, one member of the base pair AT (adenine-thymine) in DNA.

ADP (adenosine diphosphate): Molecule consisting of a nitrogenous base adenine linked to a ribose with two phosphate groups. Cells use solar energy from photosynthesis or the energy from oxidation of chemical compounds to synthesize ATP (adenosine triphosphate) by adding a phosphate group to ADP. Relative concentrations of ADP vs ATP control electron-transfer processes in cells.

affinity chromatography: Method for isolating or purifying a target molecule from a mixture by applying the mixture to a column containing immobilized ligands known specifically to bind the target. After nontargeted molecules in the mixture have passed through the column, it is washed with a solution that releases the target molecule from the immobilized ligand, allowing that molecule to be collected in a purified form.

affinity reagent: Antibody, peptide, nucleic acid, or other small molecule that specifically binds a target molecule of interest. Affinity reagents can be used to identify, track, capture, and influence the activity of larger proteins and molecular complexes in living systems.

agonist: Molecule that enhances the activity of another molecule.

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

alkylation: Modification of a molecule by adding a methyl ($-CH_3$), ethyl ($-CH_2CH_3$), or other alkyl group.

amensalism: Relationship in which one species inhibits the survival of another.

amino acid: Organic compound containing an amino group ($-NH_2$) on one end and a carboxyl group ($-COOH$) on the other. Any of 20 different amino acids are linked together in a linear fashion to form peptides or proteins. The sequence of amino acids in a protein, and hence protein function, are determined by the nucleotide sequence of genes.

analyte: Chemical substance to be experimentally measured.

ancillary pathway: Secondary biochemical pathway that supports a primary pathway of interest.

angstrom (C): One-tenth of a nanometer (10^{-10} meter).

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

anoxic: Without oxygen.

antagonist: Molecule that interferes with the action of another.

anthropogenic: Resulting from human activity.

antibody: Protein molecule synthesized as part of the immune response in vertebrate animals. When an animal's cells recognize a substance as a foreign invader (the antigen), they produce an antibody capable of binding the invading molecule. Antibodies specific to a protein of interest can be synthesized and used as reagents to detect or isolate the protein.

aptamer: Short DNA segment that can fold into a structural shape that specifically binds to another target molecule not a nucleic acid (e.g., proteins or small molecules such as NADH or ATP).

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

assimilation: Process of taking up essential elements (e.g., carbon, nitrogen, phosphorous) and converting them to biologically useful forms.

atomic force microscopy (AFM): Type of scanning probe microscopy that generates images with molecular and atomic detail by moving a probe over the surface of a biological structure. Any change in the probe's vertical position as it follows the structure's contour is detected by deflection of a laser beam pointed at the probe's tip.

atomic resolution: Level of resolution for a molecular structure that involves identifying the specific position of every atom in 3D space. Nuclear magnetic resonance spectroscopy and X-ray crystallography are used to determine molecular structures with atomic resolution.

ATP (adenosine triphosphate): Important energy carrier of all living cells. An ATP molecule consists of a nitrogenous base (adenine) linked to a ribose with three phosphate groups. Cleavage of ATP's terminal phosphate group yields ADP (adenosine diphosphate), inorganic phosphate, and the energy used to power cellular processes. Relative concentrations of ATP vs ADP control electron-transfer processes in cells.

attomole: Unit quantifying the amount of a chemical substance; equal to 10^{-18} mole where a mole represents 6.022 x 10^{23} items (e.g., molecules, atoms).

Bacteria: One of the three domains of life (along with Archaea and Eukarya) distinguished through DNA sequence analysis. Also a general term (sing., bacterium) referring to prokaryotic organisms that do not belong to the Archaea domain.

bacteriorhodopsin: Transmembrane protein that acts as a light-driven proton pump involved in ATP synthesis. Bacteriorhodopsins have been found in microorganism known to tolerate high levels of salinity and resemble light-sensitive rhodopsin proteins in the retinas of animals.

base pair (bp): Pair of weakly bonded nitrogenous bases (either adenine and thymine or guanine and cytosine) that hold together the two complementary DNA strands of a double helix.

biochemical characterization: Use of a variety of techniques to determine a protein's mechanism of action or biochemical function [e.g., its affinity for substrates and inhibitors, how it chemically modifies a substrate, how cofactors determine its mechanism of action, and how quickly it catalyzes a reaction (kinetics)].

biodiesel: Renewable alternative to petroleum diesel fuel; synthesized from the lipids of soybeans and plant materials, animal fats, and other biological sources.

biodiversity: Range of species living together in a particular environment.

bioethanol: Ethanol derived from biomass.

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

biofuel: Liquid, solid, or gaseous fuel derived from renewable biomass. Biological materials can be used to produce such fuels as biodiesel, ethanol, methanol, methane, and hydrogen.

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

biohydrogen: Molecular hydrogen (H_2) gas generated from biological processes.

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

biological pump: Collection of biological ocean processes that regulate the uptake, storage, and release of carbon.

biomarker: Chemical substance that can be used to detect the presence of a particular organism or biochemical activity in the environment.

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

biomineralization: Process in which living organisms transform a substance into a mineral.

biomolecule: Molecule synthesized by living systems, including nucleic acids (DNA, RNA), proteins, lipids, carbohydrates, and metabolites.

biophotolysis: Biological process observed in green algae and cyanobacteria that can generate hydrogen from the photosynthetic splitting of water.

biophysical characterization: Use of a variety of analytical techniques to determine a molecular machine's composition and structure.

biophysics: Application of physical principles to the study of biological structures and processes.

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

bioremediation: Use of biological organisms such as plants or microbes to degrade or chemically transform hazardous substances that have been released into the environment.

biosensor: Device that uses biological material (e.g., microorganisms, oligonucleotides, enzymes, antibodies) to detect other biological molecules or chemicals.

biosphere: Portion of earth and its atmosphere that supports life.

biotin: Water-soluble B vitamin that can be used to label macromolecules via a chemical reaction known as biotinylation. Molecules labeled with biotin are mixed with avidin proteins (from egg whites) labeled with a fluorescent compound or other reporter molecule. Avidin binds biotin very tightly, thus labeling the biotinylated molecule.

BLAST: Computer program that identifies homologous (similar) genes in different organisms by comparing all sequence data available in public databases.

calorimetry: Measurement of heat released or taken up in a chemical reaction or physical process to derive thermodynamic data (e.g., dissociation constant, free-energy change, enthalpy change, entropy change) for molecular interactions. Two kinds of calorimetry important to biomolecular studies are isothermal titration calorimetry (ITC), which can be used to detect the number of binding sites on an enzyme; and differential scanning calorimetry (DSC), which monitors the energetics of conformational changes in proteins.

capillary electrophoresis (CE): Rapid, high-resolution electrophoresis technique that separates molecules by applying an electric current to a narrow tube (<1 mm in diameter) filled with a liquid or gel.

carbohydrate: Organic compound containing carbon, hydrogen, and oxygen; most simple carbohydrates or sugars contain three to seven carbons with a chemical composition represented by the general formula $(CH_2O)_n$. Many sugars can be linked together as linear or branched chains known as polysaccharides. Carbohydrates play key roles in a variety of cell functions including energy storage, structural support, and chemical modification of proteins and lipids.

carbon cycle: The global flow of carbon from one reservoir (carbon sink) to another. Each carbon exchange among reservoirs is mediated by a variety of physical, biogeochemical, and human activities.

carbon dioxide (CO_2): Gas that is an important part of the global carbon cycle. CO_2 is emitted from a variety of processes (e.g., cellular respiration, biomass decomposition, fossil-fuel use) and taken up primarily by the photosynthesis of plants and microorganisms. CO_2 is a greenhouse gas that absorbs infrared radiation and traps heat in the earth's atmosphere.

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

carbon flux: Rate of carbon movement as it flows from one carbon reservoir to another in the global carbon cycle, usually expressed in gigatons of carbon per year (GtC/yr).

carbon free: Describes an energy source that releases no carbon during its production and use.

carbon neutral: Describes an energy source that introduces no additional carbon to the global carbon cycle. For example, carbon dioxide released from the consumption of biofuels is recaptured by photosynthesis, which generates additional biomass.

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

carbon sink: Region of the earth that takes up carbon as it moves through the carbon cycle. Four main carbon sinks include the atmosphere, terrestrial environments, oceans, and sediments.

catalyst: Substance that speeds up a chemical reaction without being altered by the reaction.

CD: See *circular dichroism*.

cell-based expression system: Protein-production technique in which genes encoding proteins needed for analysis are introduced into living host cells (e.g., *E. coli*, yeast). The host's cellular machinery synthesizes the proteins, which then can be harvested from the cells and analyzed.

cell-free expression system: Protein-production technique that uses cell lysates (typically from *E. coli* or wheat germ) containing the molecular machinery needed to synthesize proteins. DNA segments encoding the proteins of interest are added to the cell lysate mixture, and the proteins are synthesized in vitro.

cell lysate: Inner contents of a cell released by rupturing the cell membrane.

cellulase: Enzyme involved in the conversion of cellulose to simple glucose molecules. Different types of cellulases work together as a cooperative system to carry out cellulose breakdown. The three main classes of cellulases are endoglucanases, exoglucanases, and cellobiases.

cellulolytic: Having the ability to hydrolyze or break down cellulose into carbohydrate subunits.

cellulose: Large, complex polysaccharide that provides structural support to plant cell walls and is synthesized by some bacteria. Each cellulose molecule is a linear chain of thousands of glucose subunits. Cellulose is the most abundant form of carbon in the biosphere.

centrifugation: Spinning of cells, proteins, or other particles in a centrifuge to separate them from the solution in which they are suspended. The centrifugal force from spinning causes the cells or proteins to form a pellet at the bottom of the sample tube. The pellet then can be separated from the solution.

chaperone: Type of protein that ensures proper folding of other proteins into functional, 3D structures in cells; also called chaperonins.

chemostat: Apparatus for the continuous cultivation of bacteria. Chemostats keep bacterial cultures in an optimal growth state by continually adding media and removing old cells.

chromatography: Method for separating mixtures of chemical compounds. In one form, liquid chromatography, a mixture is dissolved in a solvent and applied to or passed through an adsorbent solid material. Chemical compounds migrate through the solid material at different rates, thus separating the mixture's components. Other types include affinity, size-exclusion, gas, and high-performance liquid chromatography.

chromophore: Light-absorbing pigment that gives color to a molecule.

chromosome: Self-replicating molecular structure that contains an organism's genome. In most prokaryotes, the entire genome is packaged into a single chromosome consisting of a circular DNA molecule. Eukaryotic genomes are packaged into several different chromosomes, each consisting of a linear DNA molecule wrapped around proteins.

circular dichroism (CD): Spectroscopy technique that provides structural information about molecules such as proteins and peptides. Elements of asymmetry in proteins produce characteristic CD signals in the far UV region (190 to 250 nm) of the electromagnetic spectrum that can be used to determine how much of a protein is made up of alpha-helices, beta-sheets, or random coils. CD signals from the near-UV spectral region (250 to 350 nm) can be used to determine if a protein is folded into a well-defined structure or if protein-protein interactions or changes in environmental conditions cause conformational changes in a protein's tertiary structure.

climate model: Mathematical model used to understand, simulate, and predict climate trends by quantitatively analyzing interactions between the earth and its atmosphere.

clone: Exact copy of biological material such as a DNA segment (e.g., gene or other region), whole cell, or complete organism. Gene clones inserted into cloning vectors are used to produce proteins for laboratory analysis.

cloning: Technique used to produce multiple, exact copies of a single gene or other segment of DNA to obtain enough material for further study.

cloning vector: Self-replicating DNA molecule originating from a virus, a plasmid, or the cell of a higher organism into which a DNA fragment of interest is inserted. Vectors transfer DNA into host cells, where it can be reproduced in large quantities. Examples are plasmids, cosmids, and yeast artificial chromosomes; vectors often are recombinant molecules containing DNA sequences from several sources.

cluster computing: Linking of many smaller, less expensive computers to obtain the throughput and computing power of a larger, more expensive machine; redundancy in the cluster provides greater protection from system failure. See also *grid computing.*

codon: Set of three consecutive nucleotides in mRNA that specify a particular amino acid in the protein synthesized during translation; a codon also may signal the beginning or end of the message to be translated (i.e., start codon, stop codon). See also *genetic code.*

codon bias: Preference for the use of certain codons by different organisms. Codon bias presents a problem for heterologous expression in which a gene rich in one type of codon is inserted into a host cell that rarely uses that codon, so there may not be enough of the corresponding tRNA to synthesize the protein.

cofactor: Small, nonprotein substance required for enzyme activity.

coimmunoprecipitation: Technique that uses antibodies to detect interacting proteins. An antibody that specifically binds a target protein is added to a cell lysate. The antibody forms a complex with its target and any protein or molecule bound to the target. Then an antibody-binding protein immobilized on a tiny bead is added and used to pull the antibody-protein complex out of solution.

colicin: Protein, secreted by certain strains of bacteria, that kills but does not lyse other strains.

colony: Cluster of cells originating from a single cell and growing together on a solid medium.

commensalism: Relationship in which only one party obtains some advantage.

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

comparative genomics: Field of study that compares DNA sequences of genes and genomes from different organisms to predict functions of newly discovered genes and gain insights into phylogenetic relationships among organisms.

competition: Relationship between two populations in which each is adversely affected by the other.

complementary sequence: Nucleic acid base sequence that can form a double-stranded structure with another DNA fragment by following base-pairing rules (A pairs with T and C with G). The complementary sequence to GTAC, for example, is CATG.

confocal microscopy: Type of microscopy that focuses a beam of light onto a fluorescently labeled specimen. As the laser scans the specimen within a narrow plane (<1 μm thick), light emitted from the excited fluorescent dye passes through a pinhole in a screen before reaching the light detector. The pinhole helps generate higher-resolution images by preventing out-of-focus light rays from reaching the detector and blurring the image. A computer digitizes these optical sections and develops a 3D representation of the specimen. Confocal microscopy is useful for viewing organisms that live at different depths within a biofilm. Also known as confocal scanning laser microscopy (CSLM) or laser scanning confocal microscopy (LSCM).

contaminant fate and transport model (transport model, fate model): Computer model that uses experimental data and known properties of subsurface constituents such as minerals to simulate groundwater conditions and predict how contaminants will move through and be chemically transformed by physical, chemical, and biological factors.

contaminant plume: Zone of contamination in soil, sediments, water, or air that originated from a point source.

cross-linker: Chemical group that forms a crosswise covalent connection between two parallel chains of a molecular complex.

cryoelectron microscopy (cryoEM): Type of electron microscopy (EM) that involves freezing samples to allow generation of high-resolution, 3D images of biological structures in their native, hydrated forms. Samples are

dipped in liquid ethane and chilled with liquid nitrogen in the electron microscope; samples are not stained or dried out, thus eliminating distortion associated with other EM techniques. CryoEM visualizes molecular complexes too large for nuclear magnetic resonance spectroscopy and X-ray crystallography (techniques that yield structural data with atomic resolution) and too small for conventional EM. CryoEM data are detailed enough to be used for molecular modeling.

crystallization: Formation of crystals (solid structures with highly ordered, three-dimensional, regularly repeated arrangements of atoms, ions, or molecules).

culturable: Cells capable of being grown on or in prepared media in the laboratory.

culture: Process of growing cells in the laboratory; the mass of cells produced during cultivation.

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

cysteine (cys): One of the amino acids linked together to form proteins. Cysteine is unique among all amino acids and important to protein structure because it contains a sulfhydryl group (-SH), which can form a disulfide bond with another cysteine.

cytochrome: Any of a family of iron-containing proteins that can serve as electron acceptors or donors in the electron-transfer reactions of cells.

cytoplasm: Liquid matrix, enclosed by the cell membrane, in which all inner contents of a cell are suspended.

cytosine (C): Nitrogenous base, one of the base pair GC (guanine and cytosine) in DNA.

Dalton (Da): Unit of molecular mass equal to 1/12 the mass of a ^{12}C atom and typically used in the life sciences to describe the mass of large biomolecules.

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

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

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

decrystallization: Breakdown of a solid, crystalline structure. To produce ethanol from cellulose, biomass must be pretreated with chemicals or steam to decrystallize or disrupt the highly ordered crystalline structure of cellulose and make the cellulose fibers more accessible to degradation by enzymes.

denaturation: Disruption of the native structures of proteins and nucleic acids that can be caused by increases in temperature, changes in pH, or exposure to certain chemicals. Proteins unfold and collapse into random coils, which results in loss of function; denaturation of DNA causes the two strands of the double helix to separate.

desorption: Removal of a substance that has permeated or attached to the surface of another substance; the opposite of absorption or adsorption.

detection limit: Lowest number or concentration of a particular kind of atom or molecule that can be detected by an analytical instrument or technique.

detergent: Chemical substance that contains both water-soluble (hydrophilic) and water-insoluble (hydrophobic) portions and can be used to solubilize proteins.

deuterium: Heavy isotope of hydrogen in which the nucleus contains one proton and one neutron. Also called heavy hydrogen, it is given the symbol ^{2}H or D. The most common form of hydrogen has one proton but no neutron in its nucleus.

dinoflagellate: Any of a group of eukaryotic microorganisms containing both plant-like and animal-like species that lives in marine and freshwater environments. These unicellular microorganisms use a pair of dissimilar cellular appendages called flagella for motility.

diatom: Type of microscopic, photosynthetic algae known for its intricately designed, silica-containing shell. Thousands of diatom species are known; most are unicellular, but some form colonies. Diatoms are responsible for a large portion of photosynthetic carbon assimilation in marine and freshwater environments.

direct CO_2 injection: Carbon-sequestration technique in which carbon dioxide is injected directly into the ocean depths.

directed evolution: Laboratory process used on isolated molecules or microbes to cause mutations and identify subsequent adaptations to novel environments.

directed mutagenesis: Alteration of DNA at a specific site and its subsequent reinsertion into an organism to study any effects of the change.

discovery-driven science: Research paradigm focused on creating resources and infrastructure to facilitate and advance hypothesis-based research. The Human Genome Project is an example of discovery-driven science undertaken to provide the scientific community with resources (sequence data, computational tools, technologies) to enable the pursuit of new hypothesis-based investigations. See also *hypothesis-driven science*.

disulfide bond: Structurally important covalent bond in protein complexes that can form between cysteine residues within the same or different polypeptide chains.

DNA (deoxyribonucleic acid): Molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases adenine (A), guanine (G), cytosine (C), and thymine (T). In nature, base pairs form only between A and T and between G and C; thus, the base sequence of a single strand can be deduced from that of its partner.

DNA sequence: Relative order of base pairs in a DNA fragment, gene, chromosome, or entire genome.

docking: See *molecular docking*.

domain: Discrete portion of a protein with its own function; the combination of domains in a single protein determines its overall function. Alternately, "domain" may refer to one of the three main categories of living organisms (Archaea, Bacteria, and Eukarya), whose distinctions are based on DNA sequence analysis.

dynamic range: Range of concentrations that an instrument is capable of measuring.

ecogenomics: Approach for determining the genetic potential, structure, and functional capabilities of a natural microbial community by sequencing and analyzing DNA samples isolated from the environment.

ecophysiology: Study of the physiological functions of organisms as they pertain to their ecology or interactions with each other and their environment.

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

electron acceptor: Substance that gains electrons from another substance in an oxidation-reduction reaction.

electron donor: Substance that loses electrons to another substance in an oxidation-reduction reaction.

electron microscopy (EM): Technique that uses electrons instead of light to obtain images of organelles or other structural components within cells. Imaging with electrons usually requires that a sample is analyzed in a vacuum, so living specimens cannot be visualized directly with EM. In an electron microscope, magnets and electrically charged surfaces are used to direct electrons toward a sample. As electrons pass through or are reflected by the sample, they are detected by a screen or camera that generates an image.

electron-transport chain: Series of membrane-bound proteins that receive electrons released from the oxidation of organic and inorganic compounds and mediate a sequence of electron-transfer reactions involved in the synthesis of ATP.

electrophoresis: Method of separating large molecules (such as DNA fragments or proteins) in a sample. An electric current is passed through a medium containing the sample; each molecule travels through the medium at a different rate, depending on its electrical charge, shape and size. Agarose and acrylamide gels are commonly used media for electrophoresis of proteins and nucleic acids. In one-dimensional (1D) gel electrophoresis, proteins and nucleic acids are separated in one direction on a gel, primarily by size. Two-dimensional (2D) gel electrophoresis is used in proteome analyses to separate complex protein mixtures using two separation planes (e.g., vertically down a gel by net charge and horizontally by molecular mass). Each unique protein mixture produces a characteristic pattern or fingerprint of protein separation on a 2D gel.

electrospray ionization (ESI): Method used to charge analytes as they transition from a liquid to a gaseous state. Analytes dissolved in a volatile liquid solvent are passed through a fine needle. A high voltage is applied to the analytes as they exit the needle, forming a fine mist of charged analytes (ions). Once the droplets of liquid solvent have evaporated from the ions, the ions are transported by a neutral carrier gas into the mass analyzer of a mass spectrometer.

ELSI: Ethical, legal, and social implications or issues relevant to new scientific-research initiatives.

endogenous: Originating from within a cell or organism.

environmental remediation: Removal from or immobilization of hazardous substances in a contaminated environment.

enzyme: Protein that acts as a catalyst, speeding the rate of a biochemical reaction but not altering its direction or nature.

epitope: Specific site on a protein to which an antibody will bind.

***Escherichia coli*:** Common bacterium that has been studied intensively by geneticists because of its small genome size, normal lack of pathogenicity, and ease of growth in the laboratory.

ethanol (CH_3CH_2OH): Simple alcohol containing only two carbon atoms. Ethanol is a product of the enzymatic breakdown of carbohydrates during microbial fermentation. Ethanol is combustible and can be used as a transportation fuel or fuel additive to improve gasoline combustion and reduce carbon monoxide emissions.

Eukarya: One of the three domains of life (along with Archaea and Bacteria) distinguished through DNA sequence analysis. Eukarya include animals, plants, fungi, and a variety of single-celled organisms that may have plant-, animal-, or fungi-like characteristics.

eukaryote: Cell or organism with membrane-bound, structurally discrete nucleus and other well-developed subcellular compartments. See also *prokaryote.*

exogenous DNA: DNA originating outside an organism that is introduced into the organism.

expression vector: Cloning vector engineered with regulatory signals in its DNA that enhance the transcription and translation (protein synthesis) of the gene clone inserted into it.

extracellular: Outside the cell.

extremophile: Type of microorganism that can survive extremes in temperature, salinity, pressure, and other environmental conditions detrimental to most forms of life.

exudate: See *root exudate*.

fatty acid: Long-chain carboxylic acid, typically containing 4 to 24 carbons, liberated from the hydrolysis of fats and oils.

fermentation: Metabolic pathway that breaks down organic compounds to generate cellular energy in the absence of oxygen. The production of ethanol using yeast is a fermentation pathway.

fermenter, fermentor: Large growth chamber (containing liters of liquid media) in which optimal conditions are maintained for the production of some desired product (e.g., ethanol) from microbial fermentation processes.

FISH: See *fluorescence in situ hybridization*.

flow cytometry: Analysis of biological material by detection of light-absorbing or fluorescing properties of cells or subcellular components (e.g., chromosomes) passing in a narrow stream through a laser beam. An absorbance or fluorescence profile of the sample is produced. Automated sorting devices, used to fractionate samples, separate successive droplets of the analyzed stream into different fractions depending on the fluorescence emitted by each droplet.

fluorescence: Ability of a substance to emit light at one wavelength after it has been activated by absorption of light at a shorter (higher-energy) wavelength.

fluorescence in situ hybridization (FISH): Technique that uses fluorescent probes targeted to 16S rRNA to identify and locate different populations in a microbial community. In this technique, fluorescently labeled oligonucleotide probes (containing about 20 nucleotides) are designed that will hybridize with a unique sequence in the 16S rRNA of a microbial population of interest. Cells from a culture or environmental sample are treated so they will be permeable to the fluorescent probe and then immobilized on a microscope slide. After the probes have been applied and allowed to hybridize with rRNA in the cells, the sample can be imaged using confocal microscopy. Any cells that have hybridized with the florescent probes can be identified and localized within a microbial community.

fluorescence (Förster) resonance energy transfer (FRET): Fluorescence-labeling technique that uses two different fluorescent dye molecules (fluorophores) to identify interacting pairs of proteins in vivo. Two proteins are engineered genetically so one protein is tagged with a donor fluorophore and a second protein is tagged with an acceptor fluorophore. When the two flourophores are within a few nanometers of each other, light energy emitted from the donor excites the acceptor, which emits light at another wavelength. The acceptor cannot emit light without being close to the donor, therefore, any detection of acceptor fluorescence indicates where the two proteins are interacting inside a cell. FRET emissions can be detected using confocal microscopy.

fluorophore: Group of atoms capable of fluorescence. Fluorophores can be used to label and track proteins and other molecules in vivo.

Fourier transform ion cyclotron resonance (FTICR): Type of mass spectrometry with higher resolution and mass accuracy than other MS techniques. FTICR can be used to analyze the mass of large ions generated by electrospray ionization (ESI) or matrix-assisted laser desorption ionization (MALDI). FTICR uses electrical and magnetic fields to trap ions in a chamber. As the ions circulate inside the chamber, they generate an electrical signal that is received by a detector. A mathematical function (the Fourier transform) is used to convert the detected signal into a mass-to-charge ratio for each ion.

FRET: See *fluorescence (Förster) resonance energy transfer*.

fuel cell: Device that converts the chemical energy of a fuel (e.g., hydrogen) into electricity without combusting the fuel.

fusion protein: Protein formed by genetically fusing or combining a gene encoding a target protein of interest with a gene encoding a protein or portion of protein that adds a desired functionality to the target (e.g., the ability to fluoresce or bind a small molecule on an affinity column). The fused genes then can be used as a template for synthesizing the "fusion protein" (a target protein engineered to have some additional, desired functionality).

fusion tag: Short peptide, protein domain, or entire protein that can be fused to a target protein of interest to create a fusion protein. The fusion tag generally possesses a special functionality or biochemical property that can be added to the target protein. A fusion tag can be removed from the target protein by the enzymatic cleavage of a linker region that connects the tag to the target.

gas chromatography (GC): Analytical technique used to separate the chemical components of a mixture. A sample is vaporized and carried by a stream of inert gas through a separation column (millimeters in diameter and meters in length). The column contains a solid or liquid material through which the chemical components migrate at different rates. As each separated component exits the column, it generates a signal that can be used to determine the amount and identity of each chemical. A gas chromatograph can be interfaced with a mass spectrometer.

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

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

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

gene prediction: Computer prediction identifying possible genes based on how well an unknown stretch of DNA sequence matches known gene sequences.

gene product: Biochemical material, either RNA or protein, resulting from expression of a gene. The amount of gene product is used to measure a gene's level of activity.

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

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

genetic engineering: Alteration of the genetic material of cells or organisms to enable them to make new substances or perform new functions.

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

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

genomic plasticity: Alterable nature of prokaryotic genomes that enables the fluid exchange of DNA from one microorganism to another and allows prokaryotes to adapt their genomes rapidly so they can survive changes in environmental conditions. See also *lateral (horizontal) gene transfer.*

genomics: The study of genes and their function.

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

geochemistry: The study of the chemical components that make up the earth's crust and the reactions and processes that influence the formation and cycling of those components.

gigabyte: Unit of computer storage equal to one billion (10^9) bytes.

gigaton (Gt): One billion metric tons; a metric ton is a unit of mass equal to 1000 kg (about 2200 lb).

global assay: Any of a variety of techniques that examine comprehensive sets of biomolecules (e.g., proteins, mRNA molecules, metabolites) present in a cell under certain conditions.

glucose: A six-carbon sugar with the chemical formula $C_6H_{12}O_6$. Glucose is a widely used carbon and energy source in biology and an important product of photosynthesis.

glycosyl hydrolase: Group of enzymes capable of hydrolyzing (breaking) the glycosidic bond that links a carbohydrate to another molecule. A cellulase is a type of glycosysl hydrolase that breaks the bond between glucose subunits in cellulose.

green fluorescent protein (GFP): Protein, originally isolated from jellyfish, containing a fluorophore molecule that emits green light when it absorbs UV light. GFP can be used to fluorescently label a target molecule and track it in vivo.

greenhouse gas (GHG): Heat-trapping gas such as carbon dioxide, methane, nitrous oxide, or dimethyl sulfide released into the atmosphere as a result of human activities (primarily fossil-fuel combustion) and natural processes (e.g., cellular respiration, biomass decomposition, volcanic activity).

grid computing: Large-scale computer processing tasks carried out by high-speed connections among many smaller computer systems housed at different locations and administered separately. The aggregation of unused computing resources can be applied to solving complex computational problems that otherwise would require more expensive supercomputers.

guanine (G): Nitrogenous base, one member of the base pair GC (guanine and cytosine) in DNA.

hemicellulose: Any of several polysaccharides (e.g., xylans, mannans, and galactans) that cross-link and surround cellulose fibers in plant cell walls. Hemicellulose molecules have less complicated structures than those of cellulose and are broken down more easily into their simple sugar subunits.

heterocyst: Specialized cell formed by some species of filamentous cyanobacteria when nitrogen sources in the environment have been depleted. A heterocyst has a thick, layered cell wall that minimizes the flow of oxygen into its interior; nitrogen-fixation reactions, which are sensitive to oxygen, take place within the heterocyst.

heterologous host: Host cell into which an expression vector is inserted and used to express large amounts of a protein naturally synthesized in a different species.

homology: Similarity in DNA or protein sequences among individuals of the same species or among different species.

homologous host: Host cell into which an expression vector is inserted and used to express large amounts of a protein derived from the host cell's genome. The expression vector may encode an altered form of the protein (e.g., a protein that has been mutated or labeled with a fluorophore).

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

humus: Long-lived mixture of organic compounds derived from the microbial decomposition of plant and animal matter in soils.

hybridization: Process of joining two complementary strands of DNA or one each of DNA and RNA to form a double-stranded molecule.

hybridoma: Cell line resulting from the fusion of a cancer cell with a lymphocyte (a cell that produces antibodies); hybridomas are used for the continuous production of antibodies.

hydrogenase: Enzyme capable of one or both of the following activities: (1) reduce (add electrons to) protons to generate molecular hydrogen (H_2) and (2) oxidize H_2 to generate protons and electrons (the electrons are used to reduce other molecules). Enzymes that generate H_2 are called evolving hydrogenases; those that oxidize hydrogen are called uptake hydrogenases; and those capable of catalyzing both types of reactions are called bidirectional hydrogenases.

hydrology: Study of the properties and movement of water in the atmosphere and the earth's lakes, streams, and groundwater. The study of marine waters is part of oceanography.

hydrolysis: Type of chemical reaction that uses water to cleave chemical bonds and break a large molecule into smaller components.

hypothesis-driven science: Approach in which experimental methods are used to test the validity of a hypothesis and answer specific scientific questions. See also *discovery-driven science*.

inducer: Substrate that enhances gene transcription by preventing a repressor from inhibiting the expression of a gene involved in the substrate's metabolism.

in silico: Using computers to simulate and investigate natural processes.

in situ: In a natural environment.

in vitro: "In a test tube" or outside a living organism.

in vivo: Within a living organism.

infrared (IR) spectroscopy: Technique used to characterize the structures of organic molecules. Infrared radiation has lower energy and longer wavelength (between 800 nm and 1 mm) than visible light. An infrared spectrometer measures a sample's transmission of infrared radiation. The covalent bonds within a molecule absorb infrared radiation at characteristic wavelengths. Molecules thus absorb infrared radiation in a unique pattern that can be used as a "fingerprint" for identifying that molecule.

interaction: Binding together of two or more molecules to carry out a specific cellular function.

interaction network: Diagram that shows numerous molecular interactions of a cell. Each point or node on the diagram represents a molecule (typically a protein), and each line connecting two nodes indicates that two molecules are capable of interacting.

interactome: Molecular interactions of a cell, typically used to describe all protein-protein interactions or those between proteins and other molecules.

ion: Atom or group of atoms that carry an electrical charge. Ions with a positive charge are called cations; ions with a negative charge are called anions.

ion suppression: In mass spectrometry, inhibition of ion formation of an analyte caused by the presence of less-volatile compounds. Analytes lost due to ion suppression never reach the detector, which results in artificially low readings for certain analytes.

iron fertilization: Delivery of iron-containing micronutrients to ocean regions to enhance the growth of phytoplankton that use carbon dioxide from the atmosphere to build biomass.

isoform: Any of a group of functionally similar proteins that vary slightly in amino acid sequence.

isomer: Molecule that has the same chemical formula as another but differs in how the atoms are bonded together or structurally arranged.

isotope: Atom that has the same number of protons as another atom but a different number of neutrons and hence atomic mass. For example, ^{13}C is an isotope of carbon that has one more neutron than the most common isotope of carbon, ^{12}C.

isotope-coded affinity tag (ICAT): Reagent used to label proteins analyzed by mass spectrometry. Each ICAT reagent has three parts: (1) a chemical group that reacts with a protein, (2) a linker chain that is synthesized in both light versions (e.g., containing hydrogen atoms) and heavy versions (e.g., containing isotopes such as deuterium atoms), and (3) an affinity tag (e.g., biotin). If two protein samples (e.g., from the same types of cells grown under different conditions) are each labeled with a different ICAT reagent and mixed together, the relative abundance of proteins in each sample can be determined by MS analysis.

kinase: Enzyme that catalyzes phosphorylation reactions (transfer of a phosphoryl group between ATP and another molecule). Phosphorylation reactions often have important roles in turning certain cellular processes on and off.

kinetics: Field of study that deals with determining the rates of biological, chemical, and physical processes (e.g., how quickly reactants are converted into products) under various conditions.

knockout: Deactivation of specific genes in an organism's genome; used in the laboratory to study gene function.

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

lab on a chip: Device consisting of a silicon (sometimes glass) chip chemically etched and fitted with tiny tubes and compartments (microns in size) through which materials flow. Advantages of experimentation at such a small scale are faster analysis times and significant reduction in required sample size. Also known as a MEMS (microelectromechanical system) device.

labeling: Incorporation of traceable chemical group (e.g., containing an isotope or a fluorescent dye) into a protein or other biomolecule of interest so it can be tracked or quantified during experimental analysis. See *tag*.

laboratory information management system (LIMS): Computer system used by laboratories to track samples; automate data capture from laboratory instruments; and facilitate the storage, presentation, and sharing of data among collaborating researchers.

laser confocal microscopy: See *confocal microscopy*.

lateral gene transfer: See *horizontal gene transfer*.

ligand: Any small molecule that binds a larger molecule.

ligation: Process of joining molecules or molecular fragments via covalent-bond formation.

lignin: Complex, insoluble polymer whose structure, while not well understood, gives strength and rigidity to cellulose fibers in the cell walls of woody plants. Lignin makes up a significant portion of the mass of dry wood and, after cellulose, is the second most abundant form of organic carbon in the biosphere.

ligninase: Type of enzyme capable of breaking down the complex polymeric structure of lignin into aromatic acid subunits called phenylpropanoids. Ligninases are known to be secreted by certain species of white rot fungi.

LIMS: See *laboratory information management system*.

lipid: Diverse class of biomolecules that are insoluble or minimally soluble in water. Fatty acids are key components of many complex lipid molecules that can include sugars and amino acids. Lipids take on many important cellular roles; they are the primary components of biological membranes, provide long-term storage of cellular energy, and carry electrons between membrane-embedded molecular complexes in electron-transport chains.

lysate: See *cell lysate.*

lyse: To rupture a cell and cause it to release its inner contents.

macromolecule: Large molecule (typically with a mass greater than several thousand Daltons) such as a protein, carbohydrate, or nucleic acid.

mass analyzer: Component of a mass spectrometer that uses electrical and magnetic fields to separate ionized molecules by their mass-to-charge ratios. Different mass analyzers include quadropole, time-of-flight, sector, ion trap, and FTICR.

mass spectrum (pl., spectra): Data output from a mass spectrometer consisting of a viewgraph that appears as a series of sharp peaks with each peak representing a particular ion fragment. The placement of each peak on the X axis corresponds to an ion's mass-to-charge ratio, and the height of each peak represents the relative abundance of each ion.

mass spectrometer: Instrument that ionizes molecules and then separates the resulting ions by mass and charge. A mass spectrometer consists of three basic components that operate in a vacuum: Ion source, which imparts a charge on each sample molecule; mass analyzer, which uses electrical or magnetic fields to separate each ionized molecule by its mass-to-charge ratio; and detector, which detects each separated ion and amplifies its electronic signal. The electronic signal from the detector is sent to a computer, which generates a mass spectrum for each component in the sample.

mass spectrometry (MS): Analytical technique that uses a mass spectrometer to determine mass-to-charge ratios of ions formed from the molecules in a mixture. The resulting data are used to identify each chemical component in the mixture. In proteomics analyses, MS techniques can be used to determine the mass, amino acid sequence, and post-translational modification for each protein in a sample.

massively parallel processing (MPP): Type of high-performance computing that involves running multiple processors in parallel to execute a single program.

mass-to-charge ratio (m/z): Dimensionless value measured by a mass spectrometer for each ion in a sample and determined by dividing the ion's mass (m) by its charge number (z). For example, a molecule with an atomic mass of 180 mass units and a net charge of +1 would have a mass-to-charge ratio of 180/1 or 180. Another molecule with a mass of 360 and a net charge of +2 also would have a mass-to-charge ratio of 180 or 360/2.

matrix-assisted laser desorption ionization (MALDI): Ionization method used in the MS analysis of proteins and other large biomolecules. Sample biomolecules (e.g., proteins or DNA fragments) are embedded within a solid, crystalline matrix of organic molecules. A laser beam is directed at the sample, and, as the crystals absorb the laser energy, they protect the more fragile biomolecules from destruction. The excited matrix molecules are vaporized and converted to ions that can be carried into the mass analyzer of a mass spectrometer.

megabase (Mb): Unit of length for DNA fragments, equal to 1 million nucleotides.

membrane: Semipermeable biological barrier consisting of lipids, proteins, and small amounts of carbohydrate. Membranes control the flow of chemical substances (e.g., nutrients, protons, ions, and wastes) in and out of cells or cellular compartments. They also serve as structural supports for systems of membrane-embedded proteins that mediate important biological processes such as photosynthesis and cellular respiration.

MEMS: Microelectromechanical system. See *lab on a chip*.

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

metabolic flux analysis (MFA): Method for measuring all the metabolic fluxes of an organism's central metabolism; ^{13}C-labeled substrate is taken up by an organism, and the distribution of ^{13}C throughout the metabolic network enables the quantification of labeled metabolite pools.

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

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

metabolome: All metabolites present in a cell at a given time.

metabolomics: Type of global molecular analysis that involves identifying and quantifying the metabolome.

metadata: Data that describe specific characteristics and usage aspects (e.g., what data is about, when and how data was created, who can access it, and available formats) of raw data generated from different analyses.

metagenome: Collective genomic DNA isolated from a community of organisms living in a particular environment.

metalloprotein: Protein that incorporates one or more metals into its molecular structure by binding individual metal ions [e.g., iron (Fe^{2+} or Fe^{3+}), zinc (Zn^{2+}), or magnesium (Mg^{2+})] or nonprotein organic compounds containing metals. Metalloproteins are important components of electron transport chains.

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

microbial genetics: The study of genes, gene function, and the transmission and regulation of genetic information in prokaryotic microorganisms.

microbial strain: See *strain*.

microgram (μg): Unit of mass equal to one-millionth (10^{-6}) of a gram or one-thousandth of a milligram.

micrometer (μm): Unit of length equal to one-millionth (10^{-6}) of a meter or one-thousandth of a millimeter.

micron: See *micrometer*.

microniche: 1. Specialized set of environmental conditions (e.g., pH, nutrient availability, electron-acceptor availability) that enables the survival of certain populations within a microbial community. 2. Function expressed by a microorganism or group of microorganisms living within a small portion of a community.

microorganism: Any unicellular prokaryotic or eukaryotic organism, sometimes called a microbe.

model organism: Organism studied widely by a community of researchers. Biological understanding obtained from model organism research is used to provide insights into the biological mechanisms of other organisms. Microbial model microorganisms include the bacteria *Escherichia coli* and *Bacillus subtilis*, the yeast *Saccharomyces cerevisiae*, and the green alga *Chlamydomonas reinhardtii*.

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

moiety: Portion of a molecule that carries out a particular function or gives a molecule a particular chemical characteristic.

mole: Quantity equal to 6.022×10^{23} items (e.g., molecules, atoms).

molecular docking: Binding of a molecule (e.g., ligand or protein) to a specific site on another protein to form a three-dimensional complex.

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

molecular tag: See *fusion tag*.

monoculture: Batch of microbial cells belonging to a single microbial strain or species grown in a laboratory.

motif: A sequence motif is a characteristic sequence pattern observed in different proteins or nucleic acids and typically associated with a particular function such as molecular binding. A structural motif is a recurring three-dimensional arrangement of structural elements observed in different proteins.

mutagenesis: Any process that alters an organism's genetic material (DNA or RNA sequence).

mutation: Permanent change in DNA sequence. See also *polymorphism*.

mutualism: Relationship in which both parties benefit.

NADH: Reduced form of nicotinamide adenine dinucleotide (NAD^+), a molecular carrier of high-energy electrons in living cells. NADH is formed when NAD^+ accepts a pair of electrons released from oxidation reactions. NADH then transfers its electrons to other molecules in the cell. NADH is an important source of electrons for the electron-transport pathway that generates ATP during cellular respiration.

nanometer (nm): Unit of length equal to one-billionth (10^{-9}) of a meter or one-millionth (10^{-6}) of a millimeter.

nanowire: Conductive, extracellular appendages that some bacteria can grow under certain environmental conditions and use to transfer electrons to metals. Nanowires have been observed in *Shewanella* and *Geobacter* species.

niche: 1. Set of environmental conditions required for the survival of a particular organism or group of organisms. 2. The functional role taken on by an organism in a particular ecosystem.

nitrogenase: Enzyme that catalyzes the conversion of atmospheric nitrogen (N_2) to nitrate in nitrogen-fixing bacteria.

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

NMR: See *nuclear magnetic resonance (NMR) spectroscopy*.

Northern blot: Gel-based laboratory procedure that locates mRNA sequences complementary to a piece of DNA used as a probe.

nuclear magnetic resonance (NMR) spectroscopy: Non-destructive technique that uses magnetic fields and radio-frequency (rf) pulses to analyze the structures of metabolites, proteins, or other molecules in solution. The magnetic nuclei of certain atoms (e.g., ^{1}H , ^{13}C, ^{31}P) can absorb rf energy of a particular frequency at different magnetic-field strengths. A detector within the NMR spectrometer monitors the absorbance of rf energy associated with different magnetic environments, and this information can be used to determine the position of each nuclei within a molecular structure.

nucleic acid: Large molecule composed of nucleotide subunits. See also *DNA* and *RNA*.

nucleotide: Chemical subunit of DNA or RNA consisting of a nitrogenous base (adenine, guanine, thymine, or cytosine in DNA; adenine, guanine, uracil, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA).

oligo: See *oligonucleotide*.

oligonucleotide: Short segment of 25 or fewer nucleotides that can hybridize with complementary sequence in a DNA sample.

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

open reading frame (ORF): Sequence of DNA or RNA located between the start-code sequence (initiation codon) and the stop-code sequence (termination codon).

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

ORF: See *open reading frame*.

oxidation: Loss of one or more electrons from a chemical substance.

oxidative stress: In aerobic organisms, the cellular damage caused by reactive species of oxygen (e.g., free radicals and peroxides) that are by-products of the metabolism of oxygen.

oxygenic: Producing oxygen.

parasitism: Relationship in which one organism (parasite) obtains the resources it needs for survival from another organism (host) on which or within which it lives.

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

peptide: Two or more amino acids joined by a bond called a peptide bond.

petabyte: Unit of computer storage representing one quadrillion or 10^{15} bytes (equal to 1000 terabytes).

petaflop: Measure of computer speed representing one quadrillion or 10^{15} floating-point operations per second.

petascale: Level of high-performance computing capable of petaflop processing and the management of enormous petabyte data sets.

pH: Scale used to specify acidity or alkalinity. The hydrogen ion (H^+) concentration of a sample determines its pH ($pH = -\log_{10} [H^+]$); the higher the H^+ concentration, the lower the pH. A solution with a pH value of 7 is neutral; less than 7 is acidic; and greater than 7 is alkaline or basic.

phage: Virus for which the natural host is a bacterial cell.

phage display: Method used to detect interactions between peptides or proteins and other molecules. The gene for one of many random peptide or protein variants is fused to a gene encoding a coat protein expressed on the surface of bacteriophage. Libraries of phages, each displaying a different peptide on its surface, are created and applied to a target ligand immobilized on a solid support. Phage-displaying peptides that bind the ligand are purified and used to infect *E. coli*. DNA sequencing of the infected *E. coli* is used to identify the peptides that bound the target ligand.

phenotype: Physical characteristics of an organism.

phosphorylation: Type of chemical modification that adds a phosphate group (PO_4^{3-}) to a molecule. Phosphorylation is an important type of post-translational modification involved in the regulation of protein activity.

photolysis (photolytic): Use of light energy to break a chemical bond, such as cleavage of hydrogen-oxygen bonds in water to produce oxygen and hydrogen ions.

photon: Fundamental unit (quantum) of electromagnetic energy (e.g., light) that has no mass or electric charge.

photosynthesis: Process by which plants, algae, and certain types of prokaryotic organisms capture light energy and use it to drive the transfer of electrons from inorganic donors [e.g., water, thiosulfate (H_2S)] to carbon dioxide to produce energy-rich carbohydrates.

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

phylogenetic tree: Branching, hierarchical diagram that organizes species or other taxonomic units based on evolutionary relationships.

phylogeny: Evolutionary history that traces the development of a species or taxonomic group over time.

physiology: Study of the functions of living organisms and the factors that influence those functions.

phytoplankton: Microscopic photosynthetic organisms (e.g., algae, cyanobacteria, dinoflagellates) found in the surface layers of marine and freshwater environments.

plasmid: In prokaryotes, a circular DNA segment distinct from chromosomal DNA that autonomously replicates and can be transferred from one organism to another. Plasmids can be engineered and used to introduce modified or foreign genetic material into host organisms.

polar molecule: Molecule having an uneven distribution of electrons. A polar molecule has a region of partial positive charge and a region of partial negative charge and usually dissolves in other polar substances (e.g., water).

polymerase chain reaction (PCR): Rapid technique for generating millions or billions of copies of any piece of DNA. PCR also can be used to detect the existence of a particular sequence in a DNA sample.

polymerase (DNA or RNA): Enzyme that catalyzes the synthesis of nucleic acids on preexisting nucleic acid templates, assembling RNA from ribonucleotides or DNA from deoxyribonucleotides.

polymorphism: Common variation in a gene's DNA sequence observed among individuals of the same species.

polysemy: Diversity of meanings, as when the same term represents multiple concepts. Polysemy is one of many issues addressed when defining data standards. See also *synonymy*.

population: Collection of organisms of the same species living together in a given area. A microbial community comprises several different populations.

post-transcriptional regulation: Process that controls gene expression in cells by influencing the conversion of an mRNA transcript into protein.

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

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

primary structure: Linear sequence of amino acids in a protein.

probe: Molecule used to isolate or detect the presence of certain biomolecules in a sample. A probe molecule may be a segment of DNA of known sequence that can hybridize complementary sequences in a genome or an antibody that specifically binds some protein of interest. A probe may be labeled with fluorescent groups or radioactive isotopes to facilitate isolation and detection.

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

promoter: DNA site to which RNA polymerase will bind and initiate transcription.

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

protein chip: Glass slide onto which an ordered array of proteins has been chemically attached. Protein chips can be used to identify and measure the abundance of proteins in a sample, detect protein-protein interactions, or screen thousands of proteins simultaneously for a particular biochemical function.

protein complex: Aggregate structure consisting of multiple protein molecules.

protein digestion: See *proteolysis*.

protein family: Category of related proteins similar in structure and function. Members of a protein family have highly conserved sequence with greater than 50% sequence identity.

protein folding: Process that structurally arranges a linear polypeptide chain to form a three-dimensional, biologically active form of a protein in cells. Predicting a protein's functional 3D structure from its sequence is difficult due to the many possible different interactions that can occur between atoms in the same protein molecule.

protein interaction network: See *interaction network*.

proteolysis: Breakdown of a large protein into shorter polypeptide chains by the hydrolysis of peptide bonds.

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

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

pulldown: Isolation of a protein or molecular complex from a mixture of molecules.

pull-down bait: Molecule (e.g., tagged protein or antibody) that binds a protein or molecular complex of interest and facilitates its isolation from a mixture of molecules.

QA/QC: See *quality assurance* and *quality control*.

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

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

quantum dots: Inorganic nanocrystals that can be used as fluorescent tags in a variety of live-cell imaging techniques. When a quantum dot is excited by the absorption of light at one wavelength, it emits a narrow spectrum of light at other wavelengths depending on its size and shape. Multiple quantum dots of different shapes can be used simultaneously to label different components in a sample. In contrast, organic dyes (fluorescent molecules naturally synthesized in fireflies and jellyfish) are shorter lived and tend to emit light over broad, overlapping ranges of wavelengths, thus preventing the use of more than two or three organic dyes at once.

quaternary structure: Three-dimensional molecular complex consisting of two or more folded polypeptide chains.

quorum sensing: Mechanism by which bacteria communicate and coordinate activity by sensing the concentrations of signaling molecules they release into the environment.

recombinant: Type of molecule or organism created in the laboratory by combining DNA from two or more sources.

reductant: See *electron donor.*

reduction: Electron-transfer reaction in which a substance gains one or more electrons.

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

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

regulatory map: See *gene regulatory network*.

regulatory region or sequence: Segment of DNA sequence to which a regulatory protein binds to control the expression of a gene or operon.

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

repressor: Regulatory protein that binds the operator site and inhibits transcription of genes in an operon. See also *activator*.

resolution: 1. The smallest distance between two points that can be distinguished by a microscope as separate objects; the smaller the distance, the higher the resolution of the microscope. 2. The ability of a separation technique (e.g., electrophoresis, chromatography) to separate two similarly sized components in a sample. 3. In mass spectrometry, the ability of an instrument to separate ions that differ only slightly in their mass-to-charge ratios.

resolving power: See *resolution*.

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

rhizosphere: Zone immediately surrounding the root of a plant.

rhodopsin: Light-sensitive pigment protein found in the retinas of animals. It shares structural similarities to the bacteriorhodopsins found in prokaryotes.

ribosome: Molecular machine, composed of specialized RNA and proteins, which binds mRNA and uses mRNA sequence as a template for protein synthesis.

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

root exudate: Chemical substance released from the root of a plant.

scanning electron microscopy (SEM): Type of electron microscopy in which a focused beam of electrons is scanned back and forth across the surface of a specimen, which is dehydrated and coated with a thin layer of a metal such as gold. The beam of primary electrons knocks off secondary electrons from the sample surface. The emitted secondary electrons generate signals that are amplified and used to build a 3D representation of the specimen.

scanning near-field optical microscopy (SNOM): Type of scanning probe microscopy in which a metal-coated optical fiber tip, positioned nanometers above a specimen, beams laser light onto the specimen's surface. An optical microscope detects the optical response of the laser light as it interacts with the sample. Passing light through a tiny aperture (25 to 100 nm in diameter) at the probe tip onto a specimen at such a close proximity produces an optical response that can be used to construct images with a resolution of about 50 to 80 nm, much higher than that of typical optical microscopes.

scanning probe microscopy (SPM): Any of several imaging techniques that involve sweeping a probe attached to a flexible cantilever across the surface of a specimen. Interactions between the probe and the specimen surface produce signals that can be used to generate an image of the specimen. Different types of scanning probe microscopy include atomic force microscopy, scanning tunneling microscopy, and scanning near-field optical microscopy.

scanning transmission electron microscopy (STEM): Type of electron microscopy that can be used to determine the mass and generate images of large biomolecular struc-

tures (e.g., proteins, DNA). A focused beam of electrons is scanned across the specimen, and a series of detectors within the instrument collect electrons that are transmitted through or scattered by the specimen. Signals from the detector may be used for compositional analysis of the molecular structure.

scanning tunneling microscopy (STM): Type of scanning probe microscopy that passes a sharp, conductive probe (consisting of a single atom at its tip) slightly above the surface of an electrically conductive speciman. A weak current of electrons, the "tunneling current," flows across the tiny gap between the tip of the needle and the specimen surface. The amount of current detected is related to the distance separating the tip and the specimen surface, and this information can be used to generate a 3D representation of a specimen's topography with atomic resolution.

secondary structure: Arrangement of a polypeptide chain into regions of recurring structural elements (e.g., alpha helices, beta sheets, turns) caused by hydrogen bonding among amino acids in the chain. Nucleotides of a single-stranded RNA molecule also interact to form secondary structures (e.g., the looped cloverleaf structure seen in tRNA).

sensitivity: Signal produced for a given amount of an analyte using an instrument or analytical technique.

sequence assembly: Arranging sequenced DNA fragments in their correct chromosomal positions.

sequestration: See *carbon sequestration*.

shotgun sequencing: Common approach to sequencing microbial genomes that involves breaking the genome into random fragments, which are cloned into vectors and sequenced. Computational analysis is used to compare all DNA sequence reads from random fragments and assemble the entire genome by aligning overlapping sequences.

siderophore: Chemical compound, secreted by certain species of microorganisms, that binds and solubilizes iron from the environment and facilitates the transport of iron into cells.

signal-transduction pathway: Series of biochemical reactions that receive extracellular chemical signals. These signals are transmitted and amplified within the cell and ultimately used to stimulate or repress a certain type of molecular activity (e.g., gene expression).

simulation: Combination of multiple models into a meaningful representation of a whole system that can be used to predict how the system will behave under various conditions. Simulations can be used to run in silico experiments to gain first insights, form hypotheses, and predict outcomes before conducting more expensive physical experiments.

small-angle neutron scattering (SANS): Type of molecular structural analysis carried out at facilities that have access to a neutron source. Neutrons are beamed at a sample in solution (no crystallization required). By measuring the angles at which neutrons are scattered, nanometer-scale information about the shape and structure of a molecule can be obtained.

small-angle X-ray scattering (SAXS): Type of molecular structural analysis in which X rays are beamed at a sample in solution (no crystallization required). By measuring the scattering pattern of the X rays after they have interacted with the sample, nanometer-scale information about the shape and structure of a molecule can be obtained.

small RNA molecule (sRNA): Functional RNA molecule, typically 350 nucleotides or fewer in length, that does not code for protein. sRNAs are known to regulate transcription, translation, and protein activity and can take on catalytic or structural functions as components of protein-RNA machines.

solubility pump: System of physical processes [e.g., changes in water temperature, ocean circulation, and gradient of carbon dioxide (CO_2) spanning the ocean depth] that influence the ocean's uptake of CO_2 from the atmosphere. In combination with ocean circulation, the solubility pump results in net CO_2 emissions at the equator and net CO_2 drawdown at high latitudes.

species: Taxonomic group of closely related organisms sharing structural and physiological features that distinguish them from individuals belonging to other species. In organisms capable of sexual reproduction, individuals of the same species can interbreed and generate fertile offspring. For microorganisms, a species is a collection of closely related strains.

spectromicroscopy: Combination of microscopy and spectroscopy techniques.

sporulation: Process by which certain species of bacteria produce differentiated cells called endospores. Under conditions unfavorable for growth (e.g., low nutrient availability, loss of hydration), endospores are formed and persist in a dormant state until favorable growth conditions return, causing the endospore to germinate and give rise to cells capable of normal growth and reproduction.

steady state: Growth state in which the concentration of bacterial cells is in equilibrium with the concentration of nutrients or substrates (i.e., the concentrations remain constant over time).

stochastic: Relating to a series of random events.

stoichiometry: Ratio of molecules in a structural complex.

strain: Representative of a species that differs genetically from others of the same species but not enough to be considered a new species. A strain of a microorganism often is created by genetically manipulating it to have some desired characteristic or phenotype.

structural genomics: The effort to determine the 3D structures of large numbers of proteins using both experimental techniques and computer simulation.

substrate: Substance transformed by enzymatic activity.

symbiosis: See *mutualism*.

synchrotron: Large machine that uses electric fields to accelerate charged particles to provide a continuous source of different types of electromagnetic radiation (e.g., infrared, ultraviolet, X ray) that can be used for a variety of applications, including the determination of molecular structure.

synonymy: Different terms having the same meaning. Synonymy is one of many issues addressed when defining data standards. See also *polysemy*.

synthetic biology: Field of study that aims to build novel biological systems designed to carry out particular functions by combining different biological "parts" or molecular assemblies.

syntrophy: Relationship in which two (or more) microbial populations metabolically interact to degrade a substance that one cannot metabolize alone.

systems biology: Use of global molecular analyses (e.g., measurements of all genes and proteins expressed in a cell at a particular time) and advanced computational methods to study how networks of interacting biological components determine the properties and activities of living systems.

systems microbiology: Systems biology approach that focuses on understanding and modeling microorganisms at molecular, cellular, and community levels.

tag: Molecule, chemical group, or amino acid sequence added to a protein of interest so it can be isolated or distinguished from other proteins in a mixture.

tandem mass spectrometry (MS/MS): Coupling of two mass spectrometers with a chamber known as a collision cell. The first mass spectrometer is used to separate and identify all ions in a sample. Selected ions are broken into smaller pieces in the collision cell before they enter the second instrument, which produces a mass spectrum for the pieces of selected ions. Analysis of the resulting mass spectrum provides structural information for each of the selected ions.

taxonomy: Hierarchical classification system for naming and grouping organisms based on evolutionary relationships.

terabyte: Unit of computer storage representing one trillion or 10^{12} bytes.

teraflop: Measure of a computer's speed representing one trillion floating-point operations per second.

terahertz (THz): Unit of frequency equivalent to 10^{12} hertz (10^{12} cycles per second).

tertiary structure: Arrangement of a polypeptide's folded secondary structural elements (e.g., alpha helices, beta sheets) into a three-dimensional structure.

terminal electron-accepting process (TEAP): Biochemical process in which electrons released from the oxidation of organic or inorganic compounds ultimately are transferred to a molecule or atom at the end of the electron-transport chain consisting of a series of intermediary electron-transfer reactions.

thylakoids: Membranes that contain the light-absorbing pigments, photosystems, and other proteins needed for photosynthesis. Thylakoids extend throughout the cytoplasm of photosynthetic prokaryotes. In eukaryotic plants and algae, thylakoids are housed in a special organelle called a chloroplast.

thymine (T): Nitrogenous base, one member of the base pair AT (adenine-thymine) in DNA.

transcript: RNA molecule (messenger RNA or mRNA) generated from a gene's DNA sequence during transcription.

transcription: Synthesis of an RNA copy of a gene's DNA sequence; the first step in gene expression. See also *translation.*

transcription factor: Protein that binds to regulatory regions in the genome and helps control gene expression.

transcriptome: All RNA transcripts present in a cell at a given time.

transcriptomics: Global analysis of expression levels of all RNA transcripts present in a cell at a given time.

transformation: Process by which genetic material carried by an individual cell is altered by incorporation of exogenous DNA into its genome.

translation: Process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids. See also *transcription*.

transmembrane: Term used to describe a protein embedded within a membrane that spans the entire thickness of that membrane from its external surface to its internal surface.

transmission electron microscopy (TEM): Type of electron microscopy used to image the internal structure of specimens sliced into thin sections. A focused beam of electrons passes through the specimen and onto a fluorescent screen to generate a two-dimensional image. Less-dense portions of the sample transmit more electrons and are represented by brighter regions in the image; darker regions indicate the sample's denser portions. Stains can be used to enhance contrast between light and dark regions of an image.

transporter: Protein that transports a molecule from one location to another; in most cases, transporters are membrane proteins that control the movement of molecules in and out of cells.

ultrastructure: Cellular structure too small to be visualized with light microscopy; must be examined using higher-resolution imaging techniques such as electron microscopy.

ultraviolet (UV): Form of electromagnetic radiation having a wavelength roughly in the range from 100 to 400 nm. On the electromagnetic spectrum, it is found between the violet region of the visible light spectrum and X rays.

uracil: Nitrogenous base found in RNA but not DNA; uracil is capable of forming a base pair with adenine.

UV-CD (ultraviolet-circular dichroism): See *circular dichroism*.

vertical gene transfer: Inheritance or passing of genetic material from one generation to another. See also *horizontal gene transfer*.

virus: Noncellular biological entity that can replicate only by infecting a host cell and using its reproductive capabilities.

Western blot: Method for immunologically detecting the presence of a protein in a sample. Proteins are separated by electrophoresis and then transferred to a special paper. Labeled antibodies specific to the protein of interest are used to reveal the position of the immobilized target proteins.

wide-angle X-ray scattering (WAXS): Technique for studying how ligand binding causes changes in protein structure. A beam of X rays is directed at a solution containing the protein and its ligand. The X-ray scattering pattern is used to produce structural information for the ligand-protein complex. These data can be compared with structural information for the ligand-free protein to identify changes in protein structure. Although small-angle X-ray scattering is a similar technique, it is unable to detect small changes in protein structure. Since crystallization is not required, structural information is more quickly obtained with WAXS than with X-ray crystallography.

wild type: Form of an organism that occurs most frequently in nature.

wiring diagram: Visual representation of all components and connections that make up various cellular networks including signaling, regulatory, and metabolic networks.

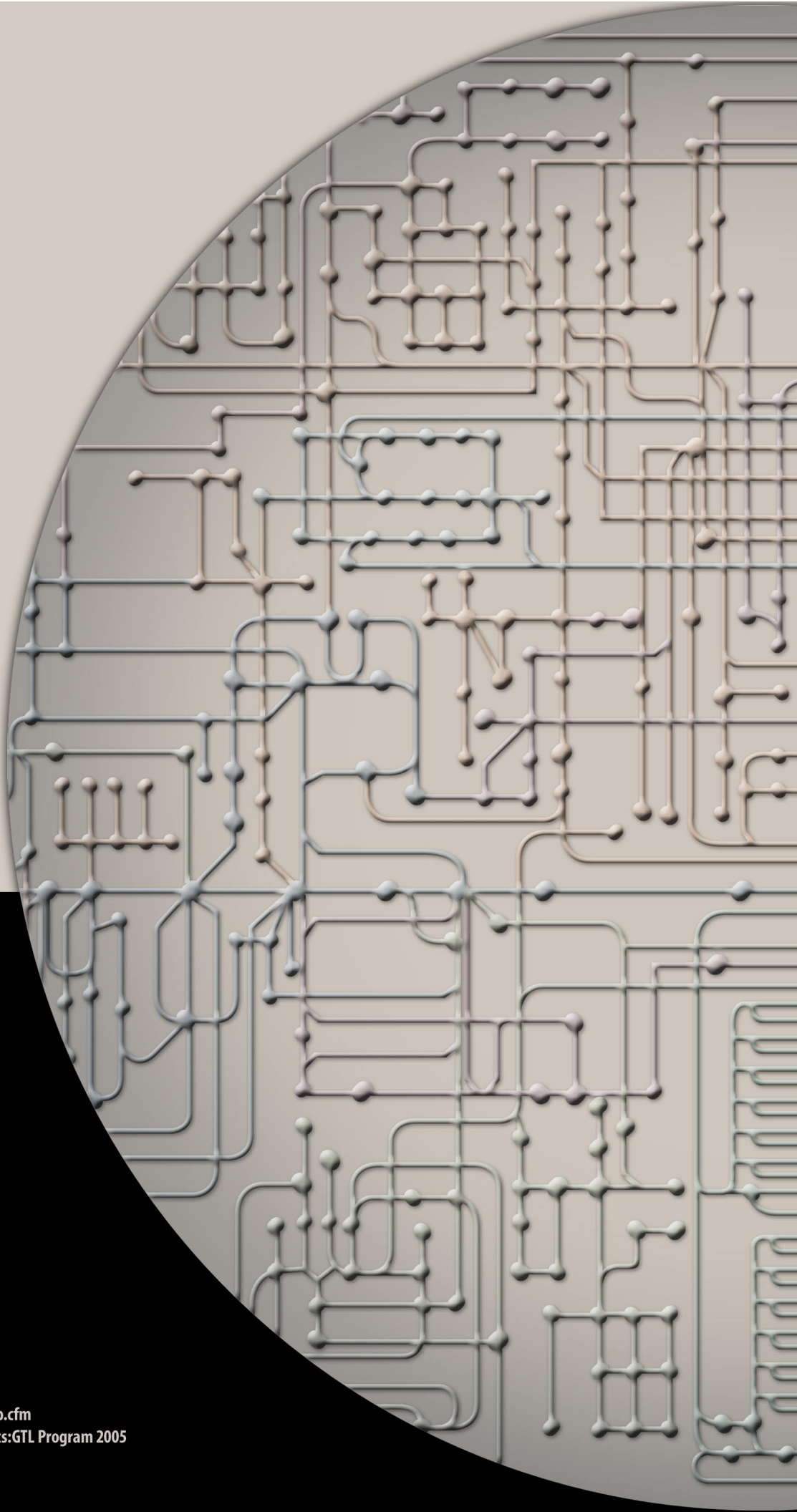
X-ray crystallography: Technique used to obtain structural information for a substance (e.g., protein, molecular complex) that has been crystallized. A beam of X rays is focused on the crystals, and the scattering pattern of the X rays is used to create 3D representations of the crystal with atomic resolution.

Acronym List

2DE two-dimensional electrophoresis
3DE three-dimensional electrophoresis
ADP adenosine diphosphate
AFM atomic force microscopy
AI artificial intelligence
AMT accurate mass and time
ASCI Accelerated Strategic Computing Initiative (DOE National Nuclear Security)
ATP adenosine triphosphate
BER Biological and Environmental Research, DOE
BLAST Basic Local Alignment Search Tool
CARS coherent anti-Stokes Raman spectroscopy
CCSP Climate Change Science Program
CCTP Climate Change Technology Program
CPU central processing unit
CryoEM cryoelectron microscopy
CSLM confocal scanning laser microscopy
CSP Community Sequence Program, DOE Joint Genome Institute
Da dalton
DNA deoxyribonucleic acid
DOE U.S. Department of Energy
ELSI ethical, legal, and social issues
EPR electron paramagnetic resonance
ERSD Environmental Remediation Sciences Division, DOE
ESI electrospray ionization
EXAFS extended X-ray absorption fine structure
FIE/L fluorescence emission/lifetime
FISH fluorescence in situ hybridization
FlAsH fluorescein arsenical hairpin
FLIM fluorescence lifetime imaging
FRET fluorescence resonance energy transfer
FT-IR Fourier transform infrared spectroscopy
FTICR Fourier transform ion cyclotron resonance
FWHM full width at half maximum
GC gas chromatography
GHG greenhouse gas
Gt gigaton
GTL Genomics:GTL
GW gigawatt
H/D hydrogen-deuterium ratio
HGP Human Genome Program
HT high throughput
IA integrated assessment
IB-PCR intact biofilm polymerase chain reaction
ICAT isotope-coded affinity tag
ICP inductively coupled plasma
IPCC Intergovernmental Panel on Climate Change
IR infrared
kDa kilodalton
LC liquid chromatography
LIMS laboratory information management system
MALDI matrix-assisted laser desorption ionization
MEMS microelectromechanical systems
MFA metabolic flux analysis
MGP Microbial Genome Program
MPP massively parallel processing
MRI magnetic resonance imaging
mRNA messenger RNA
MS mass spectrometry
MS/MS tandem mass spectrometry
MW megawatt
NABIR Natural and Accelerated Bioremediation Research
NAD, NADH nicotinamide adenine dinucleotide, a carrier of electrons produced in biological oxidations
NAE National Academy of Engineering
NIH National Institutes of Health
NLO nonlinear optics
NMR nuclear magnetic resonance
NSOM (also SNOM) scanning near-field optical microscopy
OASCR Office of Advanced Scientific Computing Research, DOE
OMB Office of Management and Budget
OPH organophosphorus hydrolase enzyme
ORF open reading frame
OSTP Office of Science and Technology Policy, White House
PCR polymerase chain reaction
PET positron emission tomography
PSI photosystem I
PSII photosystem II
PV photovoltaic
QA quality assurance
QC quality control
Q-TOF quadrupole time-of-flight
R&D research and development
RC reaction center
RNA ribonucleic acid
rRNA ribosomal RNA
SANS small-angle neutron scattering
SAXS small-angle X-ray scattering
SC Office of Science, DOE
SEC size exclusion chromatography
SEM scanning electron microscopy
SHM second harmonic microscopy
SPM scanning probe microscopy
SPR surface plasmon resonance
sRNA small RNA
STEM scanning transmission electron microscopy
STM scanning tunneling microscopy
TEAP terminal electron accepting process
TEM tunneling or transmission electron microscopy
TIR total internal reflection (type of microscopy using visible light)
TMSE theory, modeling, simulation, and experimentation
UV ultraviolet
UV-vis ultraviolet visible
UV-CD ultraviolet-circular dichroism
WAXS wide-angle X-ray scattering

Decimal Units Covered in this Roadmap

10^{+18}	exa
10^{+15}	peta
10^{+12}	tera
10^{+9}	giga
10^{+6}	mega
10^{+3}	kilo
10^{0}	
10^{-3}	milli
10^{-6}	micro
10^{-9}	nano
10^{-12}	pico
10^{-15}	femto
10^{-18}	atto



Comments: http://public.ornl.gov/hgmis/gltroadmap.cfm

U. S. Department of Energy Office of Science Genomics:GTL Program 2005