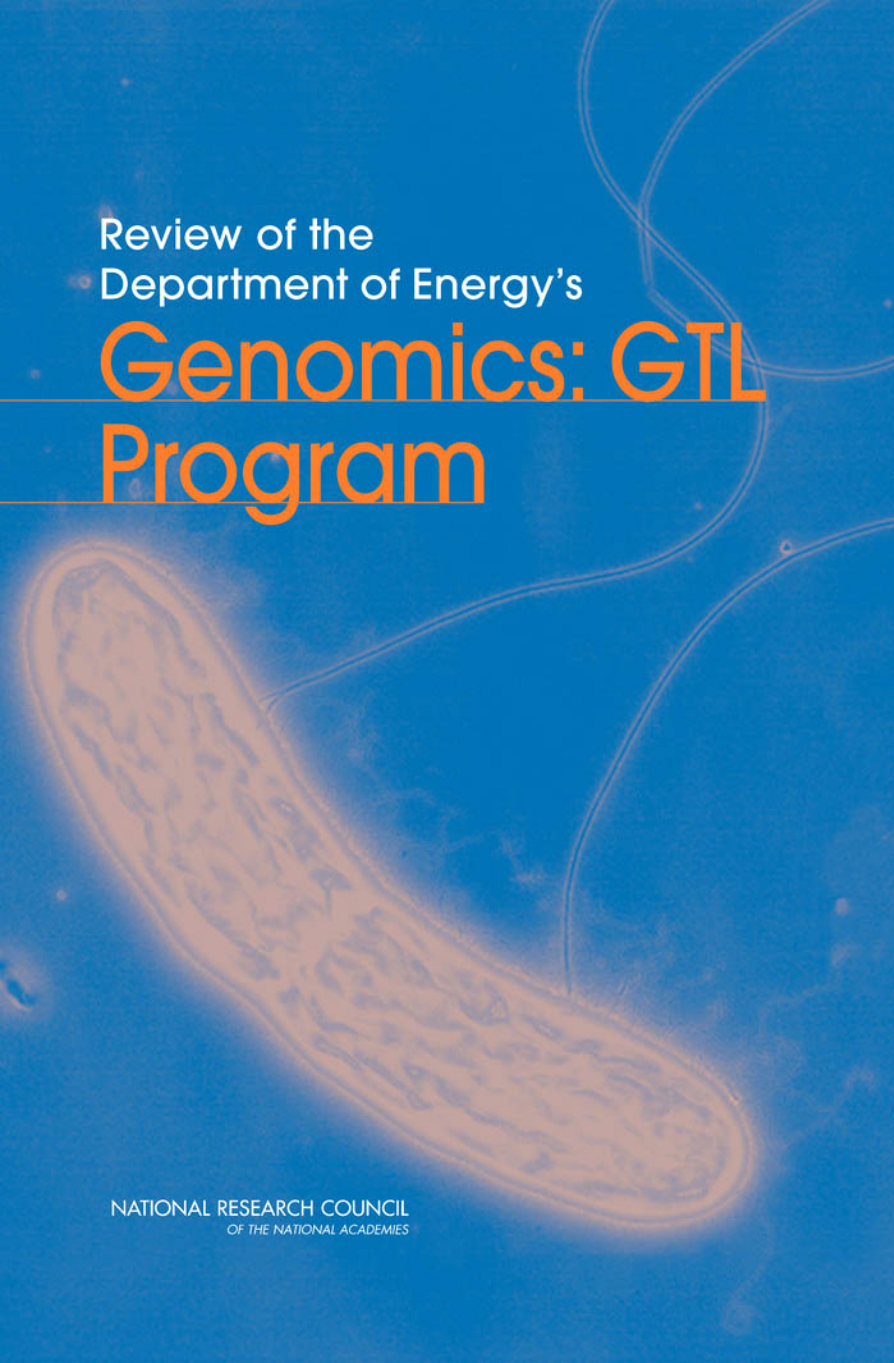

Review of the Department of Energy's Genomics: GTL Program

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

Review of the Department of Energy's Genomics: GTL Program

Committee on Review of the Department of Energy's Genomics: GTL Program

Board on Life Sciences

Division on Earth and Life Studies

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Contract DE-AM01-04PI45013 between the National Academy of Sciences and the U.S. Department of Energy. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Energy, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

International Standard Book Number 0-309-10133-6

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, NW, Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

COVER: Geobacter, a microorganism that can be used for bioremediation of uranium-contaminated groundwater and shows substantial promise for harvesting electricity from otherwise low-value energy sources. Photo by Derek Lovley, University of Massachusetts, Amherst. Design by Michael Dudzik, the National Academies Press.

Printed in the United States of America.

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON REVIEW OF THE DEPARTMENT OF ENERGY'S GENOMICS: GTL PROGRAM

JENNIE HUNTER-CEVERA (*Chair*), University of Maryland Biotechnology Institute, Rockville
CHARLES R. CANTOR, Sequenom, Inc., San Diego, California
WAH CHIU, Baylor College of Medicine, Houston, Texas
DOUGLAS R. COOK, University of California, Davis
ERIC W. KALER, University of Delaware, Newark
THOMAS KALIL, University of California, Berkeley
DAVID T. KINGSBURY, Gordon and Betty Moore Foundation, San Francisco, California
CLAUDIA NEUHAUSER, University of Minnesota, St. Paul
GREGORY A. PETSKO, Brandeis University, Waltham, Massachusetts
MARIAM STICKLEN, Michigan State University, East Lansing
LARRY P. WALKER, Cornell University, Ithaca, New York
JANET WESTPHELING, University of Georgia, Athens

Staff

EVONNE P. Y. TANG, Study Director
FRANCES E. SHARPLES, Director, Board on Life Sciences
ANN H. REID, Program Officer
JOSEPH C. LARSEN, Postdoctoral Research Associate
SETH STRONGIN, Senior Program Assistant (through September 2005)
ANNE F. JURKOWSKI, Program Assistant (since September 2005)
NORMAN GROSSBLATT, Senior Editor

BOARD ON LIFE SCIENCES

COREY S. GOODMAN *(Chair)*, Renovis, Inc., South San Francisco, California
ANN M. ARVIN, Stanford University School of Medicine, Stanford, California
JEFFREY L. BENNETZEN, University of Georgia, Athens
RUTH BERKELMAN, Emory University, Atlanta, Georgia
DEBORAH BLUM, University of Wisconsin, Madison
R. ALTA CHARO, University of Wisconsin, Madison
DENNIS CHOI, Merck Research Laboratories, West Point, Pennsylvania
JEFFREY L. DANGL, University of North Carolina, Chapel Hill
PAUL R. EHRLICH, Stanford University, Stanford, California
JAMES M. GENTILE, Research Corporation, Tucson, Arizona
JO HANDELSMAN, University of Wisconsin, Madison
ED HARLOW, Harvard Medical School, Boston, Massachusetts
DAVID HILLIS, University of Texas, Austin, Texas
KENNETH H. KELLER, University of Minnesota, Minneapolis
RANDALL MURCH, Virginia Polytechnic Institute and State University, Alexandria
GREGORY A. PETSKO, Brandeis University, Waltham, Massachusetts
STUART L. PIMM, Duke University, Durham, North Carolina
JAMES TIEDJE, Michigan State University, East Lansing
KEITH YAMAMOTO, University of California, San Francisco

Staff

FRANCES E. SHARPLES, Director
KERRY A. BRENNER, Senior Program Officer
ADAM P. FAGEN, Program Officer
DENISE GROSSHANS, Financial Associate
TOVA JACOBOVITS, Program Assiant
ANNE F. JURKOWSKI, Program Assistant
ANN H. REID, Program Officer
MARILEE K. SHELTON-DAVENPORT, Senior Program Officer
EVONNE P. Y. TANG, Senior Program Officer
ROBERT T. YUAN, Senior Program Officer

Preface

The Department of Energy (DOE) is one of the most complex of the U.S. government science agencies. It has led the nation in advancing many fields of science from physics and biology to large-scale computation. It was DOE that started the Human Genome Project and the first subsurface microbiology program and provided unparalleled capacity to the scientific community to define macromolecular structures in its synchrotrons. DOE has other user facilities, such as the Joint Genome Institute, which enables many researchers to benefit from the data acquired through whole genome sequences of both macroorganisms and microorganisms. Today, DOE maintains 17 national laboratories across the United States equipped with the latest technologies and housing some of the brightest minds in the country.

Some of the most pressing scientific and societal challenges that DOE has to deal with are finding alternative bioenergy sources, bioremediation of mixed wastes (radionuclide-contaminated organics), and enhancing carbon sequestration potentially to decrease the rate of global warming. All three of these fields of multidisciplinary research require understanding of complex biological systems starting at the organismal level and working down to the molecular level and vice versa. Recognizing that one cannot put the cart before the horse, especially if one is to design experiments at the bench to fit the reactor (applications in the field), DOE developed a forward-thinking program called Genomics: GTL (formerly Genomes to Life). The immediate goal of Genomics: GTL is to understand biological systems well enough to predict their behavior accurately with mechanistic computational models; the long-term goal of the program is to be able to develop microorganisms with capabilities for producing alternative energy sources, clean-

ing up the Cold War legacy of contaminated waste sites, and elucidating carbon cycling and sequestration.

The National Research Council (NRC) was asked to put together an ad hoc committee to review the design of the Genomics: GTL program and its infrastructure plan. We heard from various academic, industry, and national-laboratory scientists both funded and not funded by the program or involved with GTL, and we visited DOE headquarters in Germantown, Maryland, to meet and hear the perspective of the program's managers. The committee met twice to discuss our findings and writing tasks and held numerous phone conferences over the 5-month period during which this report took shape.

Our task was not an easy one by any standard, given the complexity of the science being evaluated and the need to unravel the Genomics: GTL components in order to understand better the planned future user facilities. The work of the committee was in some ways analogous to the Genomics: GTL program itself—examining the inner workings of a complex system so as to be able to design its functions for optimal output.

I thank the committee members and the NRC staff for being so giving of their time and talent. It was an honor to serve as chair of such a distinguished group of scientists. We were on a very tight schedule to finish the report and had much to review and digest in a very short time. The committee's discussions were stimulating and forward-thinking as to the role that DOE's Genomics: GTL program could play in our country's future in taking fundamental basic research all the way to translational research in the field, generating both important curiosity-driven discoveries and applied solutions for the most pressing challenges in bioremediation, bioenergy, and carbon sequestration.

We came from many different backgrounds with diverse perspectives based on personal experiences, and yet we came together as a team focused on one mission, that of making a difference and moving DOE's Genomics: GTL program to new heights based on the strong foundation it has established to date. Although the Genomics: GTL program had a "roadmap," the committee paved its own road, bumps included, to reach a final consensus that the Genomics: GTL program is critical to the success of DOE's mission and future as a leader in systems biology. Starting from this premise, the committee came up with a set of recommendations aimed at ensuring that this program would establish DOE as a world leader in microbial systems biology. The opportunity that Genomics: GTL offers scientists to advance both fundamental and applied knowledge in not only the mission focus areas of DOE but many other scientific endeavors worldwide is tremendous.

Jennie Hunter-Cevera
Chair, Committee on Review of the
Department of Energy's
Genomics: GTL Program

Acknowledgments

This report is a product of the cooperation and contributions of many people. The committee would like to thank all the speakers who attended the first meeting of the committee, on September 25-26, 2005, and others who provided information and input.

This report has been reviewed in draft form by persons chosen for their diverse perspectives and technical expertise in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards of objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following for their review of this report:

Nitin Baliga, Institute for Systems Biology, Seattle, Washington
Peter Bottomley, Oregon State University, Corvallis
Steven P. Briggs, University of California, San Diego
Jeffrey L. Dangl, University of North Carolina, Chapel Hill
Edward Dougherty, Texas A&M University, College Station
Jed Fuhrman, University of Southern California, Los Angeles
David Galas, Battelle, Columbus, Ohio
Robert Haselkorn, University of Chicago, Chicago, Illinois
Kenneth Keegstra, Michigan State University, East Lansing

Karl Sanford, Genencor International, Palo Alto, California
John Wooley, University of California, San Diego

Although the reviewers listed above have provided constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Dr. Christopher R. Somerville. Appointed by the National Research Council, Dr. Somerville was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the author committee and the institution.

Contents

Summary

The U.S. Department of Energy (DOE) is charged with promoting scientific and technological innovation in support of its overarching mission to advance the national, economic, and energy security of the United States (DOE, 2005a). Recognizing the potential of microorganisms to offer new energy alternatives and remediate environmental contamination, DOE initiated the Genomes to Life program, now called Genomics: GTL, in 2000. The Genomics: GTL program aims "to understand biological systems well enough to predict their behavior accurately with mechanistic computational models" so that such knowledge can be used to engineer systems for bioenergy production and environmental remediation and to understand carbon cycling and sequestration. Since the inception of the program, it has awarded funding for over 75 research projects to interdisciplinary teams of investigators. In the same period, DOE has been convening workshops to devise a plan for four user facilities for protein production and characterization, characterization and imaging of biomolecular machines, proteomic analysis of microorganisms, and modeling of microbial community cellular systems to facilitate genomic research relevant to its mission. The goals of the program and the plans for building the four user facilities are published in the 2005 Genomics: GTL roadmap (DOE, 2005b). The purpose of the present report, by the National Research Council Committee on Review of the Department of Energy's Genomics: GTL Program, is to evaluate the design of the program and its infrastructure plan.

The committee comprises experts in comparative and structural genomics, computational biology, protein crystallography, molecular imaging, cellular systems, biophysics, biological and agricultural engineering, and economics and

science policy (see Appendix A). The committee was charged to address three specific questions that are highlighted below. In response to its charge, the committee reviewed the Genomics: GTL roadmap and heard from representatives of DOE, scientists who are involved in the program, scientists who conduct research relevant to DOE's mission goals but are not involved in the program, and representatives of other federal agencies that have genomics programs (see Appendix B). This summary is structured as a response to the questions in the charge and includes the committee's recommendations.

Question 1: Is the Genomics: GTL program, as currently designed, scientifically and technically well tailored to the challenges faced by the DOE in energy technology development and environmental remediation?
Answer 1: Yes, the use of systems and synthetic biology approaches in the Genomics: GTL program to address some of the most pressing issues in microbial genomics relevant to DOE's mission in energy security, environmental remediation, and carbon cycling and sequestration is not only appropriate but necessary. The study of individual components only does not provide knowledge on systems integration at the level of pathways, organisms, and microbial consortia—for example, on the effects of introducing new metabolites or new or engineered organisms to a community or on organism or community responses. Systems biology research is needed to develop models for predicting the behavior of complex biological systems, to engineer microorganisms for bioremediation and energy-related needs, and to understand carbon cycling.

Current and planned research of the Genomics: GTL program promises to provide the predictive understanding of microorganisms needed to

- Develop affordable and reliable carbon-neutral energy alternatives from plants and microorganisms.
- Develop biological solutions to the many recalcitrant problems of legacy wastes.
- Increase understanding of the role of microbial communities in global carbon cycling to enable the development of carbon-sequestration techniques for addressing climate change.

The committee endorses DOE's use of a systems approach to achieve its mission goals through Genomics: GTL and supports its plan to enlarge funding of the program to $200 million per year for basic research. The committee suggests

that plant biology research be included in the Genomics: GTL program where appropriate because plants represent a major pathway to the production of bioenergy, play an important role in carbon sequestration and global nutrient cycles, and are potential sources of bioremediation. The committee's suggestion is consistent with the Energy Basic and Applied Sciences Act of 2005, which calls for an emphasis on both plants and microorganisms in the program. Systems biology research on plants and microorganisms is not likely to be conducted on a large scale without DOE's visionary thinking. (See Appendix D for a list of microbial genomics programs in other federal agencies.) Because the productive applications of new technologies to advance science will be hampered by the lack of appropriate tools, the committee finds that the concept of infrastructure for research and technology development offers a logical and even necessary pathway for achieving DOE's research goals.

Recommendation 1: The committee recommends that DOE and the nation give high priority to genomics research aimed at achieving DOE's mission goals.

Question 2: Does the proposed Genomics: GTL research and facility investment strategy leverage DOE scientific and technical expertise in the most cost-effective, efficient, and scientifically optimal manner? Specifically, does the business model (i.e. number, scope, scale, order, and user operations plan) for the proposed Genomics: GTL facilities follow directly from the science case—should one exist—for systems biology at DOE? Are there alternate models for some or all of the proposed effort that could more efficiently deliver the same scientific output?
Answer 2: The committee enthusiastically endorses the goals of the Genomics: GTL program and acknowledges the need for infrastructure, but it believes that DOE's current plan for building four independent facilities for protein production, molecular imaging, proteome analysis, and systems biology sequentially may not be the most cost-effective, efficient, and scientifically optimal way to provide this infrastructure. As an alternative, the committee suggests the construction of up to four institute-like facilities, each of which integrates the capabilities of all four of the originally planned facility types and focuses on one or two of DOE's mission goals.

Under the current implementation plan, DOE is to construct the facility of different types consecutively, each taking 6 years for design and construction. It would therefore take at least 24 years for the Genomics: GTL program to reach

full capabilities. In the fast-changing era of genome-enabled science, DOE cannot afford to spend that many years in pulling together the tools needed to make progress if it wants to remain a leader in cutting-edge science.

Irrespective of the timeline, building four single-purpose facilities may not be the most effective way to meet the multiscale and multidisciplinary challenges of systems biology. The properties of complex systems can best be studied by interdisciplinary teams of scientists who have access to state-of-the-art techniques and computational capacity. This collaborative environment will also facilitate removal of cultural barriers between experimentalists and theoreticians. The needs of the studies would be met not only by scaling up protein production, molecular imaging, proteome analysis, and modeling of cellular systems but also by integrating them so that a specific problem in systems biology (as related to the Genomics: GTL objectives), rather than only a technique, is the focus of each facility. Having four independent user facilities that merely provide services to researchers who have different research foci will reduce DOE to the role of enabler rather than leader. In addition, because the economies of scale that made large genome projects successful have not yet been obtained in work with proteins or in most aspects of systems and synthetic biology, the large-scale facilities envisioned by DOE may not be as cost-effective as expected.

The committee strongly encourages DOE to rethink its user-facility construction plans and to consider the creation of up to four integrated facilities. Each facility will combine the capabilities of the original planned facility types in a vertically integrated manner so that it can tackle all aspects of a problem or small set of problems in parallel and potentially achieve goals more quickly. The first vertically integrated facility would focus on one or two of DOE's mission goals, such as bioenergy. It would operate as an intellectual center concerned with how to provide affordable energy alternatives and become a leading institute in bioenergy. Technologies would be developed with a greater emphasis on the needs of bioenergy research than on the objective of serving the scientific community as a whole. The synergy between researchers and technology developers would also speed technology outcomes and keeps them at the cutting edge. Experience gained from the first facility would be used to refine approaches and improve planning and execution of the second facility, which would have a different theme (Box S-1). The committee feels that the revised strategy would greatly improve the cost effectiveness and efficiency of DOE's investment while optimizing the achievement of useful scientific results.

Recommendation 2: DOE should revise its plans for creating four single-purpose technology-driven facilities in sequence. Instead, DOE should create up to four institute-like facilities that each contain all the capabilities of the original planned facility types—protein production, molecular imaging, whole-proteome analysis, and systems biology—in a vertically integrated manner. Each facility should focus on one or two of the DOE mission objec-

BOX S-1
Merits of vertically integrated facilities for Genomics: GTL program

• Vertically integrated facilities would establish the Genomics: GTL program in a leadership position to launch a world-class, comprehensive, integrated research and training program in systems and synthetic biology. They would create a paradigm shift in biological research that will integrate data from a broad spectrum of spatial and temporal scales to advance understanding of biological phenomena to be able to predict or alter capabilities for optimal performance under field conditions. The facility would provide an intellectual and physical environment for both multidisciplinary teams and individual-based research.

• The research programs of the facilities would be built on overarching biological themes relevant to the DOE missions in energy production, environmental remediation, and carbon sequestration. They would involve diverse disciplines, including genomics, genetics, physiology, biochemistry, structural and computational biology, nanoscience, and engineering. The facilities would provide an intellectual and physical environment for both multidisciplinary teams and individuals pursuing research in relevant missions.

• New technologies would be developed in the facilities on the basis of well-justified scientific problems. The technologies will be aimed at particular ends rather than being ends themselves.

• The successful development of the integrated facilities would attract investigators around the country to use them. The resulting scientific discoveries and technology development can be expected to benefit not only a subset of biologists but a broad spectrum of scientists and engineers in different disciplines. Because of the diverse disciplines of the investigators, the integrated facilities are likely to have complex organization charts. That will leverage the experience of DOE to administer this new research enterprise.

• Modeling plays a central role in studying and understanding complexity. New computational approaches and tools would be developed in the facility to promote synergy between modeling and experimentation at both bench and field level.

• The first facility constitutes a pilot to validate the hybrid systems approach and to identify roadblocks to be addressed in the later facilities. All facilities can be designed so that they will not be outdated by the rapid pace of scientific discovery and technological development.

• The vertically integrated facilities lend themselves to a staged investment with expandable bases, flexibility to shift directions without losing prior investments, leverage, and open-source positioning.

tives and develop short-term, medium-term, and long-term goals to chart a course for the program. Short-term milestones should be used as a metric for independent evaluation.

The committee believes that selection of appropriate contractors, timing, and location will be key factors in the success of the Genomics: GTL program.

Contractors should be selected through an open and all-inclusive competitive process that provides adequate opportunities for universities and industry to partner with DOE and its national laboratories. Criteria for selection of contractors to implement the Genomics: GTL facilities should include innovations of the project plan relevant to DOE missions, management organization, educational outreach, technology dissemination, intellectual property management plan, proximity to a concentration of high-caliber participating scientists in diverse disciplines, and possibly provision of matching funds by the applicant institutions.

The committee encourages DOE to consider cost sharing by applicant institutions for design and construction because timely establishment of the proposed facilities is crucial in the fast-moving field of systems biology. For example, allowing the successful applicant to fund new construction or renovation of an existing facility upfront could greatly speed up the process. As federal funds become available through the appropriations process, DOE could then "lease-purchase" the facility to eventually acquire it from the private-sector partner. To expedite the operation of the facility, DOE could also use vacant space in buildings in localities that were once targeted or are being considered for major development by the biotechnology industry. Reuse of existing space would not only reduce costs but also improve access to the facilities for academic and industry scientists.

The committee feels that it would be a mistake to create new user facilities behind the fences of some of the remote existing DOE laboratories (for example, Hanford and Los Alamos) that are not close to major centers of biotechnology research. The Genomics: GTL program will not succeed in achieving DOE's mission goals unless it is embedded in a culture of strong basic biology and innovative biotechnology. To ensure the program's success, DOE should consider locating the facilities close to universities or federal or private research institutions that have established centers of excellence in biology and biotechnology. An open-access policy will also encourage the best scientists to conduct their research at the Genomics: GTL facilities.

Recommendation 3: DOE should consider locating user facilities on private land off DOE reservations to allow an open-access policy and close to research institutions that have established programs or centers of excellence in biosciences and biotechnology. The locations for the user facilities should be selected in an open and all-inclusive competitive process that provides adequate opportunities for universities and industry to partner with DOE and its national laboratories.

In addition to facilitating research and developing enabling technologies—bioinformatics, computational biology, mathematical modeling, protein production, molecular imaging, and proteome analysis—the Genomics: GTL program and its facilities can serve as training grounds for the next generation of scientists.

This is an especially important aspect of a program that promises to deliver a high degree of interdisciplinary cooperation. The National Science Foundation's (NSF) Integrative Graduate Education and Research Traineeship program is a model that DOE could consider adapting to its own purposes. Several agencies have programs in microbial genomics that complement research at DOE. DOE should also be strongly encouraged to partner with and leverage the programs of the other federal agencies with common interests in microbial biology (NSF), bioremediation (the Environmental Protection Agency), biofuels (the U.S. Department of Agriculture), and genomics (the National Institutes of Health) (DOE 2005d).

Recommendation 4: DOE should consider partnering with universities and other federal agencies to develop programs that use Genomics: GTL institute-like facilities as training grounds for the next generation scientists.

Question 3: In an era of flat or declining budgets, which aspects of the proposed Genomics: GTL program are the most meritorious? Which appear to have the highest ratio of scientific benefit to cost?
Answer 3: The Genomics: GTL program's research has resulted in and promises to deliver many more scientific advancements that contribute to the achievement of DOE mission goals. The committee proposes the initiation of facilities that integrate the four capabilities of protein production, molecular imaging, proteome analysis, and modeling and analysis of cellular systems because they are all necessary for achieving a predictive understanding of microbial systems through systems biology. Therefore, the question of which facility with a distinct capability is the most meritorious is irrelevant. If DOE were to set up vertically integrated facilities, it would have to select which of its mission foci should be the targets of the first integrated facility. The committee suggests that bioenergy be its first choice given the pressing concern of energy security.

The committee believes that there are compelling reasons for DOE to give its highest priority to creation of a "bioenergy institute." First, the U.S. (and global) economy is increasingly vulnerable to oil shocks caused by political unrest, terrorism, or natural disasters. A recent analysis conducted by Securing America's Future Energy and the National Commission on Energy Policy concluded that even "small incidents" that reduced global oil supply by 4 percent would cause oil prices to increase dramatically to more than $161 per barrel. Second, as several National Research Council reports have concluded, we need to act now if we are to have any chance of stabilizing greenhouse-gas emissions; this is espe-

cially true given the 100-year residence time of carbon dioxide in the atmosphere and the 30- to 50-year lifetime of capital stock in the energy industry. Because reduction of carbon dioxide in the atmosphere and carbon cycling are related to use of bioenergy, the bionergy institute should have a secondary focus on carbon sequestration. Finally, recent advances in biology (for example, rapid sequencing, directed evolution, and whole-genome synthesis) may enable us to design biological systems capable of generating affordable, carbon-free energy.

CONCLUSION

The committee finds that the systems biology focus of the Genomics: GTL program is essential to achieving DOE's mission goals and other goals and objectives in biology in general and in sustainable development. The current research program is excellent, and the need for infrastructure for protein production, molecular imaging, proteome analysis and modeling and analysis of cellular systems is well justified. If the committee's recommendation on reconfiguring the plans to construct vertically integrated facilities that operate as institutes were followed, they could have an especially high ratio of scientific benefit to cost because the need for new technology will be directly tied to the biology goals of the program.

1

Role of Genomics in Advancing Science

INTRODUCTION

The Genomics: GTL program of the U.S. Department of Energy (DOE) is a fundamental research program to achieve a predictive understanding of microbial systems through systems biology. The goal is to build models of organisms and communities to predict their behavior under different environmental conditions on the basis of their genomes. The program has been funding microbial genomics projects relevant to DOE mission goals since 2002. DOE plans to expand the program and build infrastructure for it. On the basis of the Energy Basic and Applied Sciences Act of 2005, DOE asked the National Research Council to convene an ad hoc committee to review the plans for the Genomics: GTL program, specifically the facilities plans.

Charge to the Committee

The committee was asked to address the following questions:

1. Is the Genomics: GTL program, as currently designed, scientifically and technically well tailored to the challenges faced by the DOE in energy technology and development and environmental remediation?
2. Does the proposed Genomics: GTL research and facility investment strategy leverage DOE scientific and technical expertise in the most cost-effective, efficient, and scientifically optimal manner? Specifically, does the business model (i.e., number, scope, scale, order, and user operation plan) for the proposed

Genomics: GTL facilities follow directly from the science case—should one exist—for systems biology at DOE? Are there alternate models for some of the proposed effort that could more efficiently deliver the same scientific output?

3. In an era of flat or declining budgets, which aspects of the proposed Genomics: GTL program are the most meritorious? Which appear to have the highest ratio of scientific benefit to cost?

This report was prepared by the committee in response to that charge. To provide background information, the committee gives a brief introduction on genomics and the scientific advances that genomics has brought and describes DOE's role in genomics research and its Genomics: GTL program in Chapter 1. In Chapter 2, the committee examines the role that the Genomics: GTL program could play in achieving DOE's mission goals. The committee reviews the design of the program and its infrastructure plan in the last chapter.

SCIENTIFIC ADVANCES BROUGHT BY GENOMICS

Genomics is the study of the structure, content, and evolution of genomes and the analysis of the expression and function of genes and proteins at the level of the whole cell or organism (Gibson and Muse, 2002). Genomics has many subfields—including functional genomics, structural genomics, proteomics, and metagenomics—and it makes use of bioinformatics and other computational tools to study the global properties of genomes. Such genomic tools as high-throughput DNA sequencing, microarrays, and the polymerase chain reaction have revolutionized biomedical science. The first full genome sequence of a free-living organism, *Haemophilus influenzae,* was determined 10 years ago (Fleischmann et al., 1995). The process was expensive and took years to accomplish, but completion of the sequence established several important principles. It showed that the so-called shotgun assembly technique was workable and effective in sequencing whole genomes. And it became clear that our understanding of the genetic information in a microorganism was much less than expected—a lesson still true 10 years later, when as much as 30 percent of the open reading frames of new microbial genomes are found to have unknown function.

Genome sequencing was quickly applied to microorganisms with larger and more complex genomes, including the yeasts *Saccharomyces cerivisiae* and *Schizosaccharomyces pombe*, and then to a series of model organisms, including the nematode, fruit fly, mustard, and mouse. With each new organism came a greater understanding of the organization and function of genomes and the identification of new genes and metabolic pathways. With the completion of the draft human genome sequence in 2003, the basis for rapidly understanding much of the genome information through comparative genomics was in place.

The sequencing of the human genome has provided detailed genetic information about specific genes and pathways in humans and has opened vast possi-

bilities for new therapies. For example, understanding of genetic changes associated with colon cancer has provided a specific basis for new cancer therapies and has been used to guide development of new drugs to treat resistant cases (Mount and Pandey, 2005), and cancer cells that are resistant to treatment can be classified on the basis of a specific gene sequence. Continuing work on the genomics of microbial species is also contributing to the improvement of human health. Scientists at Chiron Corporation, for example, used information from the sequencing of the bacterium *Neisseria meningitis* group B as the basis of a vaccine against this microorganism (Pizza et al., 2000). And current efforts to develop a vaccine for malaria, supported by the Bill and Melinda Gates Foundation, are based on interpreting genetic information on the malarial parasite (Gates Foundation, 2005).

As experience with sequencing has grown, its cost has fallen from $10 per base pair in 1990 (DOE, 2000), when it would have cost more than $30 billion to sequence the 3 billion base pairs of the human genome, to $0.001 per base pair in 2005, when the same sequence could be obtained at 1x coverage for about $3 million. The decrease in cost can be represented as a linear log curve and suggests a sequencing version of Moore's law of computing power. In this analogy, just as the complexity of an integrated circuit doubles about every 18 months, the cost of sequencing a base pair of DNA decreases by a factor of 10 roughly every 4 years. If that rate is sustained, sequencing the genome of an individual human for less than $1,000 may be possible within the next 15 years.

The time required to obtain a gene sequence is also falling rapidly. In 1989, Andre Goffeau set up a consortium to sequence the 12.5-million-base-pair genome of the budding yeast *Saccharomyces cerevisiae*. The successful effort involved 74 laboratories and took 7 years (Goffeau et al., 1996). Today, only 10 years later, the complete genome of a new strain of *Saccharomyces* can be sequenced by a single facility in less than a week, and smaller bacterial genomes can be sequenced in less than a day. In fact, the U.S. Department of Energy (DOE) Joint Genome Institute (JGI) is sequencing at a rate of more than 3 billion base pairs of DNA each month—the equivalent of 1x coverage of the human genome.

Other technologies are also revolutionizing genomic research. Microarray technology (also known as gene chips) allows the transcription level of most of the genes in an organism to be examined in a single experiment. A gene-chip experiment on budding yeast identified a previously uncharacterized gene, YDR533c, as being upregulated when the microorganism went into a quiescent state because of an accumulation of misfolded proteins (Trotter et al., 2002). The human homolog of that gene, DJ-1, was immediately identified in the human genome and was later shown to be a mutated autosomal recessive gene that affects early-onset Parkinson disease (Bonifati et al., 2003). (Parkinson disease is a protein-misfolding disorder that affects neurons, which are quiescent cells in the human body.)

The development of vast amounts of data about genomes and genetic potential defined a new approach in biomedical science of *discovery science* in contrast with the traditional *hypothesis-driven* approach. Discovery science aims to develop data resources with no specific vision as to the scientific questions to be approached. The idea is that vast data stores—when properly collected, annotated, and stored in accessible databases—are available for intense data mining by members of the scientific community who have specific hypothesis-driven questions. The various genome projects are considered discovery science, and this has proved to be a powerful scientific tool. Recently, the same approach has been extended to other "-omics" projects, the most notable being the proteomics projects that aim to define the entire protein library of a genome, including protein-protein interactions and posttranslational protein modifications. Likewise, the definition of all the metabolic pathways of a cell and their regulation (metabolomics) has begun to be an active research approach. The collection of massive data stores in -omics projects is one step in a complex "systems biology" approach to science. But although genome sequencing has proved to be a highly effective tool for gaining biological understanding, the other -omics tools have been less immediately productive thus far, because of the biological complexity of cells. Therefore, the complexity of biological systems beyond the information content of DNA—for example, proteins, metabolites, and molecular interactions, many of which are manifest only under specific developmental or environmental conditions—is not well understood.

To quote David Galas in a commentary in *Science* (Galas, 2001):

> As simple as it sounds, to know that there are no other unknown genetic components that can provide alternative explanations of experimental results is a fundamental shift of perspective. This shift is beginning to transform our approach to science, enabling researchers to face the challenge of identifying all the molecular components of the cell, as well as understanding how they are controlled, interact, and function. From a picture of the "software" of the single cell, we can look to the future when researchers will begin building, with as fine a degree of resolution, an integrated view of the universe of cell-cell interactions, differentiation, and development from single cell to organism. The availability of complete sequences of *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana* is already beginning to revolutionize such studies, and this list may soon include significant sequences from other biological models of metazoan development.

THE DEPARTMENT OF ENERGY AS A PIONEER IN GENOMICS RESEARCH

The U.S. federal system of support for science contains no central department or ministry for science. Mission-oriented research and development (R&D)

programs in defense, health, energy, environment, space and aeronautics, oceans and atmosphere, agriculture, transportation, and other fields are, instead, supported by a diverse array of agencies and departments. This pluralistic system of support is regarded as a great strength of the U.S. system and as something to be maintained and safeguarded (NRC, 1995). Under this system, allocation of funding for science is handled mainly by agencies that understand the purpose and content of R&D programs and the value of their results.

DOE is charged with promoting scientific and technological innovation in support of its overarching mission to advance the national, economic, and energy security of the United States (DOE, 2005a). As noted by Martha Krebs, former director of the Office of Energy Research (DOE and NRC, 1998) "DOE is a science agency and . . . our science enables us to meet the energy challenges ahead. All too often, DOE is the forgotten science agency, despite its ranking among the top federal supporters of basic, applied, academic, and overall research."

Many observers (for example, Kenneth I. Shine in DOE and NRC, 1998) have remarked that while the 20th century was the century of physics and astronomy, the 21st century will be the century of biology in all its ramifications. DOE's contributions to the life sciences began with health physics and radiation biology but expanded into many other fields of health and environmental research relevant to its missions. Today, DOE's participation in the pluralistic system of federal research funding means that some non-health-related life-science fields that are unfunded or underfunded by other agencies have become central and essential to DOE's science portfolio, for example, research in many fields of environmental biology, as typified by the Genomics: GTL program.

DOE has played a critical role in the development of genomics research. Under the leadership of Charles DeLisi, it initiated discussion of the Human Genome Project (HGP) in 1986. Scientists at the DOE national laboratories recognized that their long-term studies of radiation-induced mutation could be fully understood only in the context of the genetic variation that existed normally in the world's human populations. Therefore, DOE provided $5.3 million to initiate the HGP at its national laboratories. The National Institutes of Health joined DOE in the HGP in 1988 because it recognized that genomic tools could be important in understanding human genetic disorders. DOE, through efforts at Los Alamos National Laboratory (LANL), had been engaged in early DNA sequence analysis. The Genbank DNA sequence database, now operated by the National Center for Biotechnology Information at the National Library of Medicine, began as a project of Walter Goad at LANL. Many of the important tools for sequence analysis (for example, the Smith-Waterman analysis algorithm) were also developed as projects at LANL. Because of the interdisciplinary culture of the national laboratories, pioneering projects of this type were able to flourish.

Applications of Genomics at the Department of Energy

In addition to the HGP, DOE invested in other programs and facilities for genomics. In 1994, DOE began its microbial genome program. In 1996, it established JGI in Walnut Creek, California, to integrate work based at the three major DOE human genome centers. After completion of the HGP, JGI refocused its mission to align with three of DOE's primary missions: energy production, carbon management, and bioremediation. JGI's massive sequencing capabilities have served the DOE microbial genome program by sequencing the entire genomes of many microorganisms. In addition, JGI began the Community Sequencing Program, which solicits genome sequencing proposals for organisms that are relevant to DOE missions, and other organisms important to other community dynamics. In 2005, 23 projects executed by JGI will have produced complete draft sequences of genomes of diverse organisms, including plants, insects, and fishes. JGI can be characterized as a production facility that serves a broad community of scientists by providing sequence information on diverse organisms, and it has become one of the largest such facilities in the world. Development of new technology is part of the mission of JGI, and it has resulted in remarkable reductions in the time needed to obtain sequence information.

Over 50 years of nuclear-weapons research and production in the United States at DOE sites has resulted in radionuclide, metal, and organic-chemical contamination that is difficult and expensive to remove with physical decontamination methods. Microorganisms offer a biological alternative to cleaning up DOE wastes. DOE's Natural and Accelerated Bioremediation Research (NABIR) program, established in 1995, funds research aimed at providing solutions to bioremediation of contaminants in the subsurface at DOE sites. However, not all NABIR projects depend on genomics; they also involve molecular biology, microbial physiology, geochemistry, microbial ecology, and mathematical modeling. Research supported by other DOE programs on microbial systems has resulted in sequencing of microorganisms that are important in decontamination, such as geobacters, *Shewanella oneidensis*, and *Desulfovibrio vulgaris* (Heidelberg et al., 2002; Methé et al., 2003; Heidelberg et al., 2004). A number of projects use genome-based information on those important microorganisms to elucidate metabolic pathways and their interactions with other members of their ecological community. DOE is also participating in an interagency program in phytoremediation research that supports basic science; much of this work focuses on understanding molecular mechanisms of remediation of metals or organic materials by plants.

Burning fossil fuels has increased the concentration of atmospheric carbon dioxide (CO_2), a heat-trapping greenhouse gas, from the preindustrial 280 ppm to about 375 ppm today (EEA, 2004). Projections are that concentrations will more than double over the next 50 years unless emissions are reduced (IPCC, 2001). Because marine and terrestrial ecosystems play major roles in global carbon

cycling, knowledge of the key feedbacks and sensitivities of those systems are necessary to devise carbon sequestration strategies and alternative response strategies. A current example of DOE carbon-cycle management research is the work of a team of researchers at the Oak Ridge, Pacific Northwest, Argonne, and Sandia National Laboratories, and the University of North Carolina at Chapel Hill. The team is investigating cellular function in *Rhodopseudomonas palustris*, a metabolically versatile bacterium that converts CO_2 into cell material and nitrogen into NH_3, and produce hydrogen. In parallel, a team of researchers at Harvard, the Massachusetts Institute of Technology, Brigham and Women's Hospital (in Boston, Mass.), and Massachusetts General Hospital is studying proteins, protein-protein interactions, and gene regulatory networks of *Prochlorococcus marinus*, a marine cyanobacterium that is important in global photosynthesis. The group is taking a systems approach to understanding the metabolic activity of this microorganism under various environmental conditions.

Charged with securing the nation's energy supply, DOE's Office of Energy Efficiency and Renewable Energy (EERE) has a Biomass Program and a Hydrogen, Fuel Cells, and Infrastructure Technologies Program, both of which substantially involved the National Renewable Energy Laboratory. The Biomass Program aims to develop advanced technologies that transform biomass into biofuels, biopower, and high-value bioproducts (DOE-EERE, 2005a). The hydrogen program supports research on and development of low-cost, highly efficient technologies to produce hydrogen from diverse domestic sources (DOE-EERE, 2005b). Both programs fund research on genomics, but their primary focus is on applied science, so they could benefit from complementary fundamental research aimed at elucidating biological mechanisms.

Current and planned DOE research programs strive to strike a balance between discovery science, exemplified by genomics, and hypothesis-driven science, often identified with single-investigator projects. The benefits of the hybrid approach in subjects related to the DOE mission are apparent in the development of metagenomics. Microbial metagenomics involves the analysis of DNA obtained en masse from environmental samples (Handelsman, 2005a). In a sense, it is "reverse genomics" in that the structure or function of individual genomes or genes is deduced from complex mixtures of microbial consortia rather than with the classical purify-first, characterize-second approach. Metagenomics can be divided into two general categories: (1) shotgun sequencing and assembly of environmental DNA (Tringe and Rubin, 2005), typically resulting in fragmentary genome assemblies of the most abundant organisms, and (2) functional analysis of cloned DNA fragments to determine biochemical properties of interest in heterologous systems (for example, Daniels, 2005). Using metagenomics methods, scientists can study the multitude of species in an environmental system without having to culture the organisms under study. Metagenomics constitutes a huge advance over culture-dependent methods because it allows a glimpse into the nature of organisms that are inaccessible by more traditional methods.

Metagenomic analysis has given new insights for our understanding of genetic diversity in a number of environments, notably the world's oceans, estuaries, and soil communities (Tringe et al., 2005; Venter et al., 2004).

Using Systems Biology to Find Solutions for Carbon Sequestration, Environmental Remediation, and Energy Security

Although scientists often gain insight into microorganisms or microbial processes one at a time, such studies, even when pieced together, do not provide a global picture of how a biological system works. The lack of knowledge of how microbial systems work hinders our ability to harness microbial processes for bioremediation, carbon sequestration, and bioenergy production (Box 1-1). Systems biology has been defined by Ideker et al. (2001) as an approach to studying "biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational

BOX 1-1
Cost and Benefit of Understanding the Systems Biology of an Organism in Bioengineering

Obtaining an understanding of the systems biology of an organism or community of organisms may seem complex, but the cost of ignorance can be enormous. DuPont, in collaboration with Genencor International, recently succeeded in engineering the common bacterium *Escherichia coli* to produce 1,3-propanediol (PDO), a chemical building block for the new fabric Sorona (also called 3GT), which is softer and more stretchable than polyester. Chemical and biological approaches to make PDO were already known when the project began, but they were not well suited for industrial-scale production, because they were energy-intensive and required expensive starting materials. Thus, there was a need to develop a new process that would use one microorganism with the ability to convert an inexpensive basic carbon source into the desired PDO product. Such a microorganism did not exist, so one was created by inserting genes that code for enzymes that catalyze the missing chemical steps into an easily grown bacterium. The metabolic-pathway engineering could have involved, in theory, the insertion of only four foreign genes, from the bacterium *Klebsiella pneumoniae* and the yeast *Saccharomyces cerevisiae*, into *E. coli* to enable it to make PDO from glucose. However, because scientists did not have a systems biology understanding of how *E. coli* would respond to the introduction of the new enzyme activities into its metabolic systems, achieving efficient "green" production of PDO actually required modification of more than 70 different genes. Most of the modified genes were from the host organism and were needed to fine-tune critical pathways, eliminate undesired enzymes, and carefully deregulate ancillary metabolic systems in *E. coli* (Sanford, 2004). The entire process took a team of 40 people more than 7 years.

pathway responses; integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its responses to individual perturbations." Systems biology uses comparative, high-throughput assays, and mathematical or computational models to generate a picture of systemwide activities. That approach can be applied to studying systems at the subcellular level (multiprotein metabolic processes), the cellular level (integration of various functions within a cell), and the community level (interactions within multispecies communities).

Systems biology focuses on the challenge of understanding at high resolution the interlocking metabolic and molecular context for physiological activity and responses to environmental conditions. Systems biology will realize its full potential only when the properties of individual components are tied to variations at the system level. The recent emergence of synthetic biology (see Box 1-2) also provides a new and powerful approach to understanding biological systems. Synthetic biology combines knowledge from various disciplines—including molecular biology, mathematics, engineering, and physics—to develop new cellular components that are based on fundamental design concepts and that will lead to new cellular behaviors. The emerging field of synthetic biology will provide fundamental insights into cellular systems, improve our understanding of natural phenomena, and promote the development of a new engineering discipline focusing on the design and development of complex cell behaviors with predictable and reliable properties.

Using the two complementary approaches to study microorganisms and microbial communities to understand their structure and function, predict their behavior accurately, and manipulate them for desired functions is the key theme of DOE's Genomics: GTL program. The program seeks to combine discovery science with hypothesis-driven research so that an investigator with a well-formulated research question can mobilize the resources of a high-throughput facility to obtain large amounts of data on genes, gene regulation, gene products, and protein-protein interactions.

GENOMICS: GTL PROGRAM

The Genomics: GTL program was conceptualized in 2000 after Martha Krebs, director of DOE's Office of Science (formerly Office of Energy Research), charged DOE's Biological and Environmental Research Advisory Committee (BERAC) to define the agency's potential scientific roles after the HGP was completed. In response to its charge, BERAC prepared the report *Bringing Genomes to Life* (BERAC, 2000), which formed the basis of the first roadmap, "Genomes to Life," prepared by the Human Genome Management Information System at the Oak Ridge National Laboratory (ORNL) in April 2001 (Table 1-1). That first roadmap argued that the availability of genomic sequences of entire organisms would enable us to gain "a new, comprehensive, and profound under-

BOX 1-2
Synthetic Biology

Synthetic biology has been defined by some researchers as "the design and fabrication of biological components and systems that do not already exist in the natural world, and the re-design and fabrication of existing biological systems for useful purposes" (MIT Synthetic Biology Working Group, 2005). Researchers in the synthetic biology community believe that it is time to create a scientific and technical infrastructure that supports the design and synthesis of biological systems and are working to "(a) specify and populate a set of standard biological parts that have well-defined performance characteristics and can be used (and re-used) to build biological systems, (b) develop and incorporate design methods and tools into an integrated engineering environment, (c) reverse engineer and re-design pre-existing biological parts and devices in order to expand the set of functions that we can access and program, and (d) reverse engineer and re-design a 'simple' natural bacterium" (MIT Synthetic Biology Working Group, 2005).

Researchers are exploring a broad range of applications of synthetic biology to manipulate information, fabricate materials, process chemicals, and produce energy, including:

- Inexpensive biosynthesis of artemisinin, the most effective anti malaria drug.
- The design of microorganisms that can efficiently convert sunlight into other forms of energy.
- The engineering of microorganisms that can move toward contaminants and remediate heavy metals, actinides, and nerve agents.
- Embedding of the equivalent of digital circuits in bacteria and programming of communities of bacteria to perform specific tasks, such as sensing and communications.

Synthetic biology is already attracting undergraduate researchers, many of whom have participated in iGEM (intercollegiate Genetically Engineered Machine) competitions, an initiative of the Massachusetts Institute of Technology's iCampus funded by Microsoft Research. Teams of students have developed reusable "parts" for chemical control of bacterial chemotaxis and two-way cell-cell communication using DNA, which could establish the foundation for a bacterial network akin to the Internet.

Although the leading researchers in synthetic biology are in the United States, the European Union has moved aggressively to support synthetic biology as an emerging discipline. Japan is also beginning to fund synthetic biology research. The United States needs a more aggressive strategy for supporting synthetic biology.

TABLE 1-1 Major Events Leading to the Release of the 2005 Genomics: GTL Roadmap

Year	Event
1999	
November 24	Martha Krebs, director of DOE Office of Science, charges BERAC to define the department's potential scientific roles after the HGP is completed
2000	
August	BERAC publishes *Bringing the Genome to Life* in response to Krebs's 1999 charge
October 29-November 1	Genomes to Life roadmapping workshop
2001	
January 25-26	Genomes to Life roadmapping workshop
June 23	Genomes to Life workshop on role of biotechnology in mitigating greenhouse-gas concentrations
August 7-8	First Genomes to Life computational biology workshop
September 6-7	Visions for computational biology and systems biology workshop for Genomes to Life program
December 10-11	Genomes to Life: Technology assessment for mass-spectrometry workshop
2002	
January 22-23	Computing infrastructure and networking workshop for Genomes to Life
March 6-7	Computer science for Genomes to Life workshop
March 18-19	Mathematics for Genomes to Life workshop
April 16-18	Imaging workshop for Genomes to Life program
April 16-19	Computing-strategies workshop
June 19-20	Genomes to Life systems biology facilities planning workshop I
July 23	DOE awards $103 million for post genomics research.
August 16-17	Genomes to Life systems biology facilities planning workshop II
October 14-15	Genomes to Life systems biology facilities planning workshop III
December 3-4	Genomes to Life draft facilities strategy and plan submitted to BERAC by Life Sciences Division of Biological and Environmental Research program
2003	
April 1-2	GTL facility for whole-proteome analysis workshop
April 23	DOE awards $9 million for energy-related genomics research
May 12-14	Bioinformatics in GTL facility for whole-proteome analysis
May 29-30	GTL facility for production and characterization of proteins and molecular tags workshop

continued

TABLE 1-1 Major Events Leading to the Release of the 2005 Genomics: GTL Roadmap

Year	Event
June 2-4	Facility user interactions workshop
June 17-18	Characterization and imaging of molecular machines facilities workshop
July 22-24	Three Genomes to Life workshops: data infrastructure, modeling and simulation, and protein structure and prediction
September 10-11	GTL and beyond: data-standards workshop
2004	
February	Program name changed from Genomes to Life to Genomics: GTL
February 29-March 4	Genomics: GTL contractor-grantee workshop II
March 3-4	Planning study I: Genomics: GTL program science and capability needs for DOE missions
June 14-16	DOE Genomics: GTL roadmap planning phase 2
2005	
February 6-9	Genomics: GTL contractor-grantee workshop III
October 3	Genomics: GTL roadmap released

SOURCE: Adapted from http://doegenomestolife.org/program/timeline.shtml.

standing of complex living systems." High-throughput data and high-performance computing are the two key elements to achieve the goal. Large amounts of data would need to be collected to characterize proteins, molecular machines, gene regulatory networks, and entire microbial communities in natural environments at the molecular level. Computational methods and capabilities would need to be developed to integrate the data and to gain a predictive understanding of these complex biological systems. The 2001 roadmap called for program managers to "meet with stakeholders in a series of workshops, scientific society symposia, and other exchanges on scientific topics to guide program development."

In 2002, DOE put out the first request for proposals (RFP) under the Genomes to Life program, now called Genomics: GTL. The RFP called for applications for "research from large, well integrated, multidisciplinary research teams that support the Genomes to Life research program." The theme of the program was to develop the experimental and computational capabilities necessary to enable a predictive understanding of the behavior of microorganisms and microbial communities of interest to DOE (Box 1-3). Since its launch, the Genomics: GTL program has funded some 75 research projects and subcontracts, including basic research and outreach programs. It has also funded two infrastructure projects at the national laboratories and facilitated 22 workshops on topics ranging from genomics-enabled geomicrobiology to high-performance computing.

BOX 1-3
Selected Highlights of Genomics: GTL Research to Date

Bioenergy Alternatives

During the next 2 decades, U.S. energy demand is expected to outpace substantially the increase in domestic production of fossil fuels. Concurrent with an increased requirement for energy is a need to reduce dependence on foreign sources of oil and thereby increase energy security.

One pillar of DOE's missions is to explore and facilitate development of renewable, environmentally safe, biological sources of energy. Among the topics to be addressed by Genomics: GTL are biological production of liquid (ethanol, methanol, and biodiesel) and gaseous (hydrogen and methane) fuels. One key is an increased understanding of microbial enzyme consortia that participate in degradation of biological polymers, such as lignin and cellulose, which are major chemical components of plant life. JGI has determined the DNA sequence of a fungal species that has an unusual capacity for degradation of cellulose and lignin biomass. In the genome of that fungus, scientists at the DOE national laboratories discovered genes for the enzymes involved in biomass conversion, making the goal of improving enzymes for biomass conversion to ethanol-based fuels more tangible. Other energy-related plans of the Genomics: GTL bioenergy program include efforts to redirect microbial photosynthesis to generate hydrogen fuel in a process that uses energy derived from sunlight to convert water into hydrogen and oxygen and research into the remarkable ability of some soil microorganisms to produce electricity from simple organic compounds.

Bioremediation

DOE is charged with remediating thousands of our nation's most contaminated landscapes, many of which are the legacy of a diverse network of defense facilities. The scale of several of those landscapes, some of which exceed 1 million cubic meters of contaminated earth, will require innovative, biologically based remediation strategies. DOE-funded scientists are working to increase knowledge of microbial systems involved in the remediation of toxic metals and radionuclides. Researchers in a project funded through ORNL are meeting the challenge to understand those complex systems by developing computational models that predict the behavior of key regulatory networks involved in bioremediation. In parallel, DOE has funded research on the genetic potential of a microbial species that has a documented capacity for uranium bioremediation and the ability to produce electric energy from organic matter.

Carbon Cycling and Sequestration

Atmospheric greenhouse-gas concentrations have increased steadily over the last 2 centuries; massive quantities of carbon are released into the atmosphere each year because of human activity. The Intergovernmental Panel on Climate Change predicts a doubling of CO_2 concentrations by the middle of the 21st century with potentially serious consequences for the quality of our environment.

Earth's marine environments and in particular their microbial inhabitants constitute a potential tool to change the balance of the CO_2 equation. A key to realiz-

continued

BOX 1-3 Continued

ing that potential is an increased understanding of the planet's biological carbon cycle, including microbial photosynthesis—a process that uses light energy to convert atmospheric CO_2 into the organic molecules that make up life on Earth. Manipulating photosynthetic systems on a grand scale may offer a means to decrease atmospheric CO_2. To that end, researchers at the Sandia National Laboratory are developing experimental and computational methods to understand the genes and proteins of the photosynthetic marine microorganisms of the genus *Synechococcus*, which play a key role in Earth's carbon cycle, and their colleagues at ORNL and Pacific Northwest National Laboratories are working to characterize the multiprotein machines involved in the microbial carbon cycle.

Research on Enabling Technologies

To achieve the long-term goals of Genomics: GTL, it is essential that technical limitations and knowledge gaps be addressed. Much of the research funded by the program aims to lay the foundation for future study by solving key issues in genome-directed science. For example,

- Several research projects aim to develop computational models to understand complex microbial systems, and other researchers are developing data warehouses and computational tools to organize and relate genomic information for bench scientists.
- Other scientists, distributed among several projects, are working to develop novel methods to image biological systems, including visualization of DNA-protein interactions that regulate an organism's genetic potential and monitoring of life's processes on the microscopic scale of single living cells.
- DOE-funded scientists are devising innovative methods to culture recalcitrant species of microorganisms; such breakthroughs will greatly facilitate the study and manipulation of these species in a laboratory setting.
- Although the term *genomics* typically conjures images of genes and proteins, the ultimate effect of many genes and proteins is to cause changes in the small-molecule complement of a cell, otherwise known as the metabolome. Metabolites can serve a practical role as building blocks of other cellular molecules, or they may have more intriguing roles as signal molecules that orchestrate microbial behavior. In any case, understanding how microbial metabolism influences microbial function is an important goal, and it is the focus of several projects funded by the Genomics: GTL program.

DOE has committed about $240 million from FY 2002 to FY 2006. Of that amount, 60 percent has funded scientists at DOE-operated national laboratories, and 40 percent has funded scientists at academic and private research institutions. The majority of funding awarded to scientists in academic and private laboratories has gone to three institutions (see Appendix C).

Taken together, the funded research projects are addressing some of the most

pressing issues in microbial genomics. Several have direct application to DOE's energy-related mission, and others are developing enabling technologies and datasets that are necessary for the advancement of microbial genomics generally. All funded projects are relevant to energy security, environmental remediation, or carbon cycling and sequestration.

In parallel to the Genomics: GTL program, a series of workshops was held to discuss facility needs. Those workshops led to a working paper presented to BERAC in April 2002 that called for the creation of "unique, high-throughput research facilities to translate the new biology, embodied in the Genomes to Life (GTL) program, into reality for the nation." Those facilities would integrate high-throughput biology and computation and information management and would be resources for the broad scientific community.

Later that year, BERAC provided a draft implementation plan for four user facilities for the Genomics: GTL program. The plan was developed in a series of workshops in 2003. All the workshops ultimately resulted in the outline of the four facilities that are described in the *2005 Roadmap for DOE Genomics: GTL. Systems Biology for Energy and Environment* (DOE, 2005b). The facilities would be constructed sequentially and complement each other.

- *Facility for production and characterization of proteins and molecular tags*. This facility would produce all proteins encoded in any genome on demand, including molecular tags to identify, locate, and manipulate proteins in living cells. The core facility instrumentation will consist of high-throughput technologies for protein-production screening and robotic systems for affinity-reagent production and characterization. Computational capabilities will allow data capture and management, genomic comparative analysis, and control of high-throughput and robotic systems.
- *Facility for characterization and imaging of molecular machines.* This facility would identify and analyze molecular-machine components from microbial cells, including their structure, function, assembly, and disassembly. Facility instrumentation will include mass spectroscopy to characterize molecular machines and imaging capabilities to localize them in cells. Computational capabilities will allow for modeling and simulation of molecular interactions to understand how these complex structures arise.
- *Facility for whole proteome analysis*. This facility will enable the identification of all proteins and other biologically significant molecules (such as lipids, carbohydrates, and enzyme cofactors) that a microbial cell produces under different, but controlled, environmental conditions to identify responses to various environmental influences and to elucidate pathways. The core facility instrumentation would include large numbers of chemostats to grow microbial systems under various environmental conditions and instrumentation to analyze the molecular makeup of microbial cells, such as nuclear magnetic resonance spectrometer and mass spectrometer. Computational capabilities would allow for data

analysis and modeling and simulation of microbial systems to inform experiments and predict their outcomes.

- *Facility for modeling and analysis of cellular systems.* This facility will focus on the study of microbial communities under highly controlled conditions that mimic natural environments. The goal would be to gain an understanding of microbial communities through analysis of functional properties of individual species or multispecies consortia by using imaging techniques that allow nondestructive monitoring of the molecular makeup of cells within the communities. Instrumentation would include cultivation technologies for microbial communities under highly controlled environmental conditions and imaging instrumentation to resolve the molecular makeup of cells spatially and temporally. Computational capabilities would focus on data analysis and modeling, including simulating complex microbial communities.

The committee examined the current Genomics: GTL program and the challenges that it faces in achieving DOE's mission goals. The committee enthusiastically concluded that the case for DOE to play a leading role in systems biology is extremely strong. On the basis of that assessment, the committee considered whether high-throughput capabilities in protein production, proteomics, molecular imaging, and systems biology would facilitate the advancement of Genomics: GTL research in a cost-effective, efficient and scientifically optimal manner. Finally, the committee examined the current plan for the four proposed user facilities, its own proposed alternative plan, and discussed the pros and cons of the two plans.

2

Role of Genomics: GTL in Achieving the Department of Energy's Mission Goals: Promise and Challenges

The Department of Energy (DOE) has the mission of protecting our energy and economic security and our environment by promoting a diverse, reliable, affordable and environmentally sound domestic energy system. In carrying out that mission, DOE has recognized that genomics and systems biology research will enable development of novel strategies to address the agency's three strategic challenges (DOE, 2005b):

- To develop biofuels as a major secure energy source.
- To develop biological solutions for remediation of soil, sediment, and groundwater contaminated with metals, radionuclides, and organic hazardous wastes.
- To understand relationships between climate change and Earth's microbial systems and to generate options for carbon sequestration.

The Genomics: GTL program is expected to provide the scientific underpinning for predicting and manipulating the behavior of complex biological systems, particularly systems that may play a central role in developing biotechnology solutions to fulfill DOE's energy and environmental mandates. The Genomics: GTL program therefore becomes critical for strengthening the nation's scientific leadership in systems biology and supporting an evolving industrial biotechnology sector that is essential for the nation's economic competitiveness in the global economy. The following discussion is offered to illustrate how the science of the Genomics: GTL program can be used to address the three strategic challenges.

BIOENERGY

The Genomics: GTL program is addressing the needs for new sources of energy that could

- Reduce the risk of global climate change by dramatically lowering the emission of greenhouse gases.
- Have a favorable energy balance.
- Have the potential to compete effectively with fossil fuels in the marketplace.
- Reduce the adverse environmental effects of today's pattern of energy production and consumption.
- Meet a substantial fraction of U.S. (and global) energy demand.

One source of energy that could eventually meet those criteria is bioenergy produced by a variety of plants and microorganisms. The Genomics: GTL program could play a key role in realizing the potential of bioenergy by generating the fundamental knowledge that would make it technologically and economically feasible. Although it is premature to pick a "winner," the research community has identified a number of promising directions, including

- Genetic modification of crops to increase yields of usable energy per unit of cultivated land by a factor of 3-5 while maintaining nutrient and water requirements.
- Conversion of cellulosic biomass to fuels by depolymerizing cellulose and hemicellulose into their component sugars and then converting the sugars to fuel.
- Design of algae or bacteria that cost-effectively produce hydrogen or hydrocarbons.

Energy from biomass is the largest source of renewable energy in this country; it has surpassed hydropower and makes up 3 percent of the total energy consumed in the United States (Perlack et al., 2005). A recent study conducted by the Natural Resources Defense Council (NRDC, 2004) concluded that scientific and technological advances and sound public policies could rapidly expand the use of plants and plant-derived materials for energy. By 2050, biofuels could displace more than 7 million barrels of oil per day, the equivalent of nearly half of the oil that the United States use in the transportation sector. In that scenario, the United States would be able to reduce emissions of greenhouse gases by nearly 1.7 billion tons per year (as measured in tons of carbon dioxide [CO_2] equivalents)—more than 22 percent of U.S. greenhouse-gas emissions in 2002. A transition to biofuels could also lead to improvements in air quality in that biofuels have almost no sulfur and produce fewer particles and toxic air pollutants (NRDC, 2004).

Development of advanced biological conversion processes (enzymatic, microbial, and plant processes) is central to the DOE biomass program and to the expanding industrial biotechnology sector. Biological processes are the preferred path because they tend to have higher reaction specificity, require milder reaction conditions, and produce fewer toxic byproducts. Those characteristics are consistent with the goal of developing industrial processes and systems that are environmentally friendly. However, the challenges are to increase rates and extents of conversion in an array of microbial and biochemical processes, to accelerate commercial development of biofuels, and to expand the portfolio of industrial enzymes, microorganisms, and plants for an expanding bioeconomy.

The Genomics: GTL program can promote the development of more-effective bioconversion processes and plant-based feedstocks by enhancing our understanding of biological conversion processes from a systems perspective. Understanding of systems biology will lead to better methods and tools for manipulating and controlling metabolic pathways that are important for bioenergy and industrial chemical production, for prospecting for novel industrial enzymes and microorganisms, and for bioengineering to enhance plants' usability as feedstocks for energy and industrial chemicals. For example, one goal of the Genomics: GTL program is to discover functions of genes that could contribute to cheaper biofuels. The development of more-efficient and cost-effective enzymes is a critical step in making the abundant and diverse array of plant-derived polysaccharides available for the production of energy and industrial chemicals. A study commissioned by the National Research Council (NRC, 2000) and the roadmap from the Biomass Technical Advisory Committee (2002) established by the Biomass R & D Act of 2003 have each identified enzyme engineering as having one of the top three priorities in biological research to support the development of microbial or plant-based biofuels and industrial chemicals. The high priority is based on the recognition that enzymatic conversion of biomass is the preferred path for processing microbial or plant-based resources into industrial products because enzymes exhibit specific catalytic activities and enzymatic processes are environmentally more benign. It is also recognized that the industrial development of enzymes itself occupies an important industrial biotechnology sector that holds the promise of expanded economic growth. The enzyme market was estimated to be $2 billion per year in 2004 and to have an annual growth rate of 4-5 percent (Business Communication Company, 2004).

Plant cell walls comprise a highly complex matrix of polysaccharides, including cellulose, lignins, pectins, and diverse hemicelluloses. Microorganisms found in soils, compost piles, and other environments have been shown to produce enzymes effective for degrading each one of those polysaccharides into fermentable sugars. The Genomics: GTL program can play an important role in increasing understanding of the structure and functions of genes associated with degradation of polysaccharides by those microorganisms. It can also contribute to the discovery of new polysaccharide-degrading enzymes by prospecting for novel

microorganisms in exotic environments. For example, termites degrade polysaccharides. Bacteria that live in a termite's hindgut break down plant matter and release hydrogen as a byproduct. The mechanism of hydrogen production in the termite hindgut is not yet known. The DOE Joint Genome Institute (JGI) is scheduled to sequence the community of microorganisms in the termite hindgut by 2006. That would enable the Genomics: GTL program to identify and characterize the enzymes associated with hydrogen production in the termite hindgut.

Another example involves the production of microbial polysaccharide-degrading enzymes by plants (Nuutila et al., 1999; Dai et al., 2000; Ziegler et al., 2000; Ziegelhoffer et al., 2001). Today, the National Renewable Energy Laboratory estimates that enzyme production costs account for $0.10 of the price of a gallon of ethanol. The challenge is to reduce that by half by increasing the activities of enzymes or reducing the production cost. One avenue toward that goal is the use of plants as biomolecular farms for the production of the enzymes. The concept has been demonstrated in a study in which a corn plant was used to produce an *Acidothermus cellulolyticus* endogluconase. In light of methods of bioconfinement of recombinant crops in the field (NRC, 2004), that technology could well become America's standard technique for production of cellulases and other polysaccharide-degrading enzymes in biomass crops that are converted into fermentable sugars that can be fermented into ethanol biofuel. It is important to remember that the knowledge developed by producing polysaccharide-degrading enzymes in plant biomass should also be largely applicable to other proteins, including those which produce such valuable industrial products as 1,3-propanediol, a precursor of many important industrial polymers.

Cost-effective and efficient microbial conversion processes are necessary to convert low-cost sugars derived from plant-based resources to ethanol and other industrial chemicals. Although there are several key technological differences in how ethanol is produced from corn or cellulosic feedstock, both paths to ethanol production require a fermentation step that involves the conversion of glucose and other sugars to ethanol. Currently, baker's yeast, *Saccharomyces cerevisiae,* provides the primary microbiological system used by the corn-based ethanol industry. As we seek to increase the amount of ethanol produced from biomass, we will have to increase our knowledge of the metabolism of important microbiological systems, particular those with the potential to enhance the production of useful biobased products. That need was clearly articulated in the National Research Council report on biobased industries (NRC, 2000) and in DOE's Genomics: GTL roadmap (DOE, 2005b). Two of the key research activities identified by the Research Council are relevant to the Genomics: GTL program:

- "Analysis of biochemical pathways that integrate basic intracellular measurements. Such analysis will provide fundamental understanding of the microbial metabolism and physiology necessary to focus metabolic engineering manipulations on enhancing organisms' overall productivity."

- "Basic research on principles of intermediate microbial metabolism to gain a better understanding of how concentrations of substrate or product can inhibit rates of product formation. Such understanding will aid in engineering bioreactors control to enhance the rate and conversion of raw materials into useful products."

Rapid advances in genomics have facilitated the manipulation of metabolic pathways to engineer organisms that can efficiently produce a desired metabolic product or reduce unwanted byproducts. Metabolic engineering allows a more directed and rational use of classical genetic or molecular biology tools to optimize the production of metabolites and proteins of interest. For example, the complete genome sequence of *S. cerevisiae* was published in 1996, and a complete collection of deletion mutants of yeast is commercially available (Goffeau et al., 1996). They create many opportunities to customize systems biology research and develop metabolic engineering tools to characterize the metabolic networks of wild-type *S. cerevisiae* and newly constructed mutant strains. For example, the Genomics: GTL program could provide new insight into the role of cellular myo-inositol in the physiological and metabolic behavior of *S. cerevisiae* that might reveal a clear link between high phosphatidylinositol concentration and ethanol tolerance.

Another important subject would be metabolic engineering of bacteria—such as *Thermotoga neapolitana*, *Enterobacter aerogenes*, and *Clostridium butyricum*—that are hydrogen producers that use fermentative pathways. Such microorganisms are known to ferment sugars to hydrogen at a relatively high rate by glycolytic breakdown of sugars through the anaerobic metabolism of pyruvate (Hallenbeck, 2005). The generation of hydrogen by fermentative bacteria is accompanied by the formation of organic acids as metabolic products that are not used by the microorganisms (Nath and Das, 2004). Thus, altering the metabolic pathway to shift more of the pyruvate to hydrogen is an important step for improving fermentative hydrogen production. The science, methods, and tools of the Genomics: GTL program would strengthen our understanding of the regulatory and metabolic pathways that influence hydrogen production and create opportunities for more-informed engineering of those pathways and others.

Although the focus of Genomics: GTL bioenergy research is on microbial processes, it should be clear from the preceding paragraphs that biomass for bioenergy is derived from plants. Better understanding of the mechanisms and regulation of polysaccharide and cell wall synthesis in plants is critical to meeting the goals of the nation's bioenergy research agenda. For example, it may be possible to engineer plants for novel cell wall structures that enhance the efficiency of biomass conversion. The committee believes that bioenergy research through Genomics: GTL should include a parallel focus on polysaccharide and cell wall synthesis in plants. To that end, *Arabidopsis thaliana* provides an outstanding experimental platform for developing a systems-level analysis of plant

function. The resulting knowledge could greatly enable bioengineering applications involving biomass species, such as corn and poplar. The committee's suggestion is to focus a portion of the Genomics: GTL program on specific aspects of plant biology (in this case, aspects relevant to biomass conversion), and not to develop a broad-based effort in plant biology.

CARBON SEQUESTRATION

About 6 billion tons of CO_2, a greenhouse gas, is released into the atmosphere by anthropogenic activities each year. Atmospheric CO_2 concentrations have increased because of human use of fossil fuels and changes in land use, such as deforestation. It is also known that microorganisms can be used to mitigate global change due to human activities, such as agriculture, mining, and waste treatment (ASM, 2004). It is estimated that atmospheric CO_2 and methane concentrations are now increasing at about 0.4 percent and 1 percent each year, respectively. The growing scientific evidence that CO_2 and other greenhouse gases are altering our climate has stimulated interest in CO_2 sequestration as a means to counteract global climate change. The DOE mission with respect to carbon cycling and sequestration is to "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." Photosynthetic terrestrial and aquatic organisms naturally perform biosequestration, and understanding how this is achieved at the whole-organism and microbial-community levels is one of the important roles of the Genomics: GTL program.

Microorganisms have a much greater role in mediating biogeochemical activities than previously thought, given that they outnumber all other forms of life on land and in rivers, lakes, and oceans. Therefore, it is of great importance to understand the genetic regulation behind these biogeochemical activities and the role of microorganisms in carbon sequestration. Many of the critical questions surrounding the role of microorganisms in biosequestration were addressed by the American Academy of Microbiology (ASM, 2001):

- Which microorganisms are responsible for producing and consuming specific environmentally important compounds, and how does the diversity of microorganisms affect soil, water, and atmospheric concentrations of various chemicals?
- How and to what extent do microorganisms and their recycling processes respond to climate change and other disturbances?
- How can information about activities occurring on the scale of microorganisms (micrometers to millimeters) be integrated across scales of communities, landscapes, and ecosystems to help to explain phenomena observed on a global scale?

• What new technologies and computational systems are needed to facilitate integration and understanding across scales?

Those are complex questions and can be addressed with a systems biology approach for generating realistic strategies for biosequestration. We provide some examples on how each of the above questions can be addressed by the Genomics: GTL program. First, it is known that phytoplankton photosynthesis in the oceans is an important subsystem in the recycling of CO_2 in the biosphere. Variation in the species composition or population sizes of the ocean's phytoplankton could theoretically have a great effect on the oceans' ability to take up atmospheric carbon. The focus of the Genomics: GTL program is to understand how those microorganisms affect ocean ecosystems by cycling carbon and other important elements, such as nitrogen. For example, the program is supporting studies on and the sequencing of *Emiliania huxleyi.* DOE has also sequenced several species of ocean carbon-sequestering phytoplankton, such as the diatom *Thalassiosira pseudonana* and the cyanobacteria *Prochlorococcus* and *Synechococcus*. Diatoms account for about 20 percent of global carbon fixation, and the genome sequence of *T. pseudonana* sheds light on the silicic acid metabolism and energy storage and use strategies that allow diatoms to flourish in marine systems (Armbrust et al., 2004; DeLong and Karl, 2005). Many scientists believe that *Prochlorococcus* constitutes the most abundant photosynthetic organisms on Earth. Understanding those organisms' roles in global carbon cycles is central to the issue of carbon sequestration.

As to the second question, this is being addressed by DOE studies on soil microorganisms. Soil respiration accounts for 75 percent of the carbon in the terrestrial ecosystem, and according to Rosenberg, Metting, and Izaurralde (2004), it returns nearly 10 times as much CO_2 to the atmosphere as emissions from fossil-fuel combustion. Agriculture and fire also contribute substantially to the carbon being released to the atmosphere from soils. Metagenomic studies of soil samples indicate that microbial communities have a wide range of mechanisms and biochemical pathways for carbon metabolism, some of which may emerge as targets for the application of carbon management strategies. However, the mechanisms by which microbial populations adjust to climate change in an ecosystem are not well understood. Genomics, proteomics, and metabolomics could elucidate such mechanisms and thereby increase our understanding of the important role of microorganisms in carbon cycling. Moreover, the resulting knowledge could enable predictive models of system function that might presage changes in the global carbon economy.

Although currently missing from the Genomics: GTL research plan, plants contribute substantially to nutrient cycles in the soil through both photosynthesis and nitrogen fixation. As a consequence of these autotrophic processes, the soil zone around plant roots (the "rhizosphere") is among the most important and diverse, yet least understood, of ecosystems. To fully address soil microbiology

and its relevance to carbon sequestration, nitrogen cycling, and response to climate change, it will be essential to include substantive efforts on rhizosphere biology in Genomics: GTL. By definition, such efforts will need to include studies of both plant and microbial systems, extending the systems biology analogy to multiorganism and cross-kingdom interactions.

Working across scales from whole organisms to the biosphere is daunting. The Biological and Environmental Research Division in DOE's Office of Science established CSiTE (Carbon Sequestration in Terrestrial Ecosystems)—a research consortium—to perform fundamental research that will lead to acceptable methods of enhancing carbon sequestration in terrestrial ecosystems as one component of a carbon-management strategy. Three national laboratories are members of CSiTE—the Argonne, Pacific Northwest, and Oak Ridge National Laboratories. The goal of CSiTE (DOE-ORNL, 2002) is "to discover and characterize links between critical pathways and mechanisms for creating larger, longer-lasting carbon pools in terrestrial ecosystems. Research is designed to establish the scientific basis for enhancing carbon capture and long-term sequestration in terrestrial ecosystems by developing:

- "Scientific understanding of carbon capture and sequestration mechanisms in terrestrial ecosystems across multiple scales from the molecular to the landscape,
- "Conceptual and simulation models for extrapolation of process understanding across spatial and temporal scales,
- "Estimates of carbon sequestration potential,
- "Assessments of environmental impacts and economic implications of carbon sequestration."

The fourth question, on needed new technologies and computational systems, highlights the need to develop infrastructure for advancing the Genomics: GTL program. For example, metagenomic methods can document the makeup and activity of ocean communities involved in CO_2 recycling. The sequencing of microbial communities in the Sargasso Sea appears to have revealed 1.2 million previously unknown genes, including almost 800 genes coding for rhodopsins that are presumed to be involved in phototrophy (Venter et al., 2004). Those data serve as a starting point for further Genomics: GTL studies on the mechanism of this potential energy-yielding process. It is believed that the dynamics of carbon-assimilation and anabolic pathways that sequester carbon or return it to the atmosphere, respectively, will be elucidated and that biological models of carbon-sequestration activity can be developed to assess the effects of carbon-cycle perturbations on climate change. DOE can collaborate with and build on the joint National Science Foundation and U.S. Department of Agriculture program on microbial-genome sequencing that supports research on the diversity of microorganisms and their roles in complex ecosystems and in global geochemical cycles

(NSF, 2005). DOE is also exploring the use of biosensors based on genomic information that would detect changes in the levels of DNA, RNA, proteins, and metabolites in response to stress or population shifts.

Obviously, answers to those questions require a broad interdisciplinary approach and will benefit from genomic studies to identify key genes and pathways. Understanding the complexity of ecosystems in which many functions are being carried out simultaneously by millions of microorganisms of diverse species is no small task; it will take years of study, including the development of novel experimental and computational approaches. Effective implementation of the Genomics: GTL program would facilitate the broad interdisciplinary and multidisciplinary research collaborations needed to address the questions on a variety of biological and physical scales.

BIOREMEDIATION

During the 50 years of the nuclear era, the United States invested in facilities to do research on, develop, manufacture, and test nuclear weapons and materials. The environmental-remediation legacy left from those manufacturing and testing activities is staggering: DOE has the responsibility for monitoring and cleaning up more than 7,000 sites at 100 facilities. The groundwater and soil at those sites are contaminated with radionuclides, which are often mixed with other wastes, such as chlorinated hydrocarbons. Collectively, 2 trillion gallons of contaminated groundwater and 75 million cubic meters of soil and subsurface sediment must be remediated at these sites. For comparison, the groundwater volume equals 4 times the U.S. daily water consumption, and the sediment volumes would fill 17 professional sports stadiums.

With current technology, cleanup costs would run to $300 billion over a 70-year period. Hence, alternative strategies, methods, and technologies are being investigated throughout the many DOE remediation programs. The one with the most promise is bioremediation. Bioremediation is the use of microorganisms to contain or eliminate hazardous and radioactive wastes or decrease them to environmentally safe levels. Enhancing it with the modern tools of biotechnology could save 30-50 percent of the cost. The committee notes that, although they are not covered in the Genomics: GTL plan, plants have also been shown to have utility in bioremediation and thus should be considered among the targets for Genomics: GTL research. Enhancing the capacity and quality of bioremediation by means of the modern tools of biotechnology could lead to savings in the range of 30-50 percent. For example, at DOE's Savannah River site in Aiken, South Carolina, bioremediation of subsurface solvent contamination cost two-thirds as much as a pump-and-treat method and was 40 percent more efficient. An Environmental Protection Agency study of 150 sites that use bioventing (a form of in situ bioremediation) showed cost savings of 50-90 percent.

Bioremediation of organic contaminants involves transforming them to be-

nign products, such as CO_2. Metals and radionuclides must be immobilized to prevent subsurface travel to rivers or groundwater. All bioremediation systems have a common goal of stimulating and maintaining microbial metabolism (Hughes et al., 2002). Stimulation might involve optimizing the metabolic pathways of whole organisms or of a community of microorganisms to achieve the desired transformation. The optimization of metabolic activities of whole organisms and microbial communities is the key to converting hazardous materials to nonhazardous materials or nonbioavailable forms and is consistent with the research and development activities of the Genomics: GTL program.

The Genomics: GTL roadmap listed several research needs:

- Assessment of benefits and effects.
- Establishing links between biology and geochemistry.
- Using genome sequences as a launching point for understanding communities.
- Modeling microbial metabolic activities.
- Merging metabolic and field-scale models.

Assessment of benefits and effects is going on through four DOE programs managed by the Environmental Remediation Sciences Division: the Natural and Accelerated Bioremediation Research (NABIR) program, the Environmental Management Science Program, the Environmental Molecular Sciences Laboratory, and the Savannah River Ecology Laboratory. Of all the programs reviewed, NABIR has had the most promising results for field applications and NABIR programs are linked to the Genomics: GTL program now (COV, 2004). The goal of the NABIR program is "to provide the fundamental science to serve as the basis for the development of cost-effective bioremediation of radionuclides and metals in the subsurface at DOE sites." The program focuses on intrinsic bioremediation and accelerated bioremediation through the use of biostimulation (the addition of inorganic or organic nutrients).

Links between biology and geochemistry are being established through research focused on the survival of environmental microorganisms under stressful conditions, such as those at bioremediation sites. Researchers are integrating fields of biology—for example, genomics, ecology, molecular biology, proteomics, bioinformatics, and metagenomics. By understanding processes that allow specific bacteria to exploit different environments—such as water, air, soil, and the subsurface—scientists are identifying critical mechanisms for survival. For example, Gary Andersen and his group focus on understanding mechanisms of bacterial diversity by using 16S rRNA gene sequences to measure the relative abundance of individual members of microbial communities. In partnership with the DOE-JGI in Walnut Creek, California, they have developed novel microarray systems to measure dynamic changes and rapid systems for classifying the thousands of individual sequences from clone libraries that are being constructed.

Another example of establishing links between biology and geochemistry is the research under way on the ubiquitous aquatic bacterium *Caulobacter crescentus* (Box 2-1). That organism was selected for extensive study by DOE because of its ability to survive in low nutrient habitats where contamination may be present. The completed genome sequence of strain CB15 has provided information needed to study genomewide response to heavy-metal stress. A customized 500,000-probe Affymetrix array was designed by Harley McAdams's group at Stanford University to measure transcription levels of all 3,763 putative open reading frames (DOE, 2005c), both strands of genes for hypothetical proteins, and the intervening intergenic regions. The microarray was used to study transcriptional response to heavy-metal stress.

The work of Derek Lovley, of the University of Massachusetts, Amherst, on optimizing in situ bioremediation of uranium and harvesting electrical energy from waste organic matter by *Geobacter* species is an example of how genome sequences can be used as a launching point of understanding. His project addresses not only the identification and validation of the microbial community involved in the bioremediation of uranium in contaminated subsurface environments but also the use of this microbial community to harvest electricity from waste organic matter and renewable biomass. He is engaged in subsurface environmental studies in Colorado at the Old Rifle Uranium Mill Tailings Remedial Action site. His studies are supported primarily by the Genomics: GTL program, which focuses on detailed geochemical and microbiological characterization of the site.

Lovley's research and the research of countless other environmental researchers across the country would benefit from a new generation of molecular ecology tools that might evolve from the Genomics: GTL program. For example, current methods for profiling 16S rDNA can provide information about the structure of a microbial community, but they tend to sample the most abundant species. Similarly, transcriptional profiles may provide information about community composition and their corresponding transcription and metabolic activities, but the methods are limited by the difficulty in reliably sampling environmental RNA. Moreover, RNA from only the most abundant organisms can be sampled, and potentially important details (such as spatial information) are lost. Thus, many of the current molecular ecology methods have reached their experimental limits. The Genomics: GTL program provides opportunities to address those and other technological limitations.

Modeling microbial metabolic activities is an important and challenging goal of the Genomics: GTL program. As articulated in the Genomics: GTL roadmap, the Genomics: GTL program enables three key modeling scenarios: microorganism-mineral interactions and resulting molecular structure and charge transfer, microbial-community responses (for example signaling, motility, and biofilm formation), and ensuing community functionality. There is a growing awareness in the systems biology research community that mathematical modeling is an essential tool for exploring those elements because it provides a framework for

BOX 2-1
Radiation-Resistant *Deinococcus*

The ionizing radiation-resistant microorganism *Deinococcus radiodurans*, a strain of which was isolated from a radioactive sediment beneath a leaking waste tank on DOE's Hanford site in Washington state, has remarkable capacity for resisting environmental stresses, such as radiation and desiccation, and can transform contaminants, such as chromate, to less mobile and hazardous forms. Recent research has also revealed that *Deinococcus* accumulates high concentrations of intracellular manganese while limiting its intake of iron—factors hypothesized to be largely responsible for its ability to resist ionizing radiation. The underlying biological mechanisms responsible for those traits, however, remain unknown. Genomics, the study of the genes and their associated functions, has tremendous potential for revealing the underlying mechanisms and the environmental factors that control their expression (DOE, 2002).

FIGURE 2-1 A new species of radiation-resistant *Deinococcus* isolated from radioactive sediment beneath a leaking Hanford waste tank (DOE, 2002).

Metal Resistance in *Caulobacter crescentus*

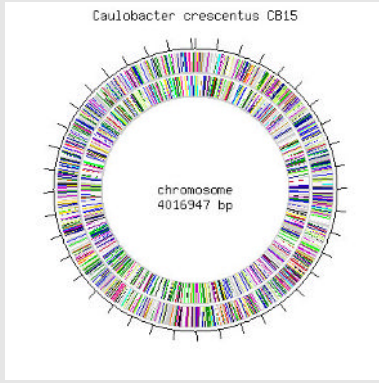

FIGURE 2-2 *Caulobacter crescentus* (left) and diagram of its genome. (Diagram of *C. crescentus* genome provided by Gary Andersen. Photo of *C. Crescentus* provided by Yves Brun.)

The bacterium *Caulobacter crescentus* is known for its broad habitat, its distinctive ability to live in low-nutrient environments, and its being a model organism for studying cell-cycle regulation. Gary Andersen's group at the Lawrence Berkeley National Laboratory, in collaboration with Harley McAdams's group at Stanford University, identified the pathways responding to heavy-metal toxicity in *C. crescentus* to provide insights into the possible application of *Caulobacter* to environmental restoration. They exposed *C. crescentus* to five heavy metals (chromium, cadmium, lead, selenium, and uranium-238) and analyzed genomewide transcriptional activities after exposure with a high-density GeneChip® array. In addition to the unexpected finding that *C. crescentus* CB15N is tolerant to uranium, the combination of biochemical, transcriptional, and imaging analysis, showed that *Caulobacter* formed a calcium-uranium-phosphate precipitate extracellularly in contrast with the intracellular sequestration mechanism of other resistant bacteria, such as *Arthrobacter* spp. That was consistent with the limited response to oxidative stress seen with other metals. Combining the results of transmission electron microscopy, the team speculated that a possible mechanism for this precipitation is a biotic formation of extracellular calcium-uranium-phosphate precipitates. A key aspect of this research has been the interrogation of the whole genome on both strands. That has allowed the identification of antisense transcripts that are differentially regulated for each metal and that play an important part in the response model (G. Andersen, Lawrence Berkeley National Laboratory, personal communication on December 14, 2005).

structuring the understanding of complex biological systems; it can be used to extract insights and mechanisms from a rich set of empirical studies that have been sponsored by DOE, and it provides mathematical and computational tools that can be used in the design, analysis, and optimization of bioremediation strategies.

Thus far, the mathematical modeling and computer simulation tools being used to study bioremediation systems lack substantial biological detail. Bioremediation is an inherently spatial problem that operates under nonequilibrium and highly nonlinear conditions. It involves species that are potentially diverse genetically, and this results in complex and multi-dimensional models. Strong selection is likely to act on populations that are far from equilibrium and can result in considerable changes in the genetic composition of the population and thus potentially unpredictable changes in responses. Including evolutionary factors into standard ecological population models can yield behavior that differs from that of models in which evolutionary factors are absent (Neuhauser et al., 2003), but considerable experimental effort would be needed to develop such models and make them accurate and predictive. Beyond simulations, there is little available mathematical theory that could be applied to the transient behavior of such systems. Thus, there is a tremendous opportunity for the Genomics: GTL program to lead the effort to bring systems mathematics to the challenges of bioremediation.

Merging metabolic and field-scale models is a daunting task that requires a strong multidisciplinary approach. The Genomics: GTL program can address several essential elements, including the following problems:

- Identify and characterize the multiprotein complexes—"protein machines"—that perform most cell functions in microorganisms.
- Determine how the operations of the machines are orchestrated to allow organisms to thrive in diverse environments.
- Describe the metabolic capabilities of complex microbial communities in their natural environments.
- Develop new computational methods and tools to increase the understanding of complex biological systems and predict their behavior.

The Genomics: GTL program will also provide valuable data for improving field site treatments, for example, by developing computer models that would test and elucidate the activities and interactions occurring in microbial communities. Observations made with respect to syntrophic relationships, anaerobic degradation consortia, and shifts in the dominant terminal electron-accepting process observed in sediments could be integrated into models that describe the dynamic flow and transport regimes found at most DOE contaminated sites.

As stated earlier, bioremediation is considered to be the least expensive and

most versatile means of dealing with soil and groundwater contamination. Genomic studies being carried out by the Genomics: GTL program on microorganisms that have remediation potential will enable an assessment of the capability of individual species or strains and inform scientists and engineers how the bioremediation process might be better managed or improved. In addition, Genomics: GTL provides synergy with the NABIR program goals and supports research into the capabilities of microbial communities to promote metal and radionuclide precipitation. Finally, Genomics: GTL data may provide insight to guide the development of biosensors to monitor bioremediation over large areas and long durations and thus to help to sustain bioremediation activities in the field.

CHALLENGES TO THE ACHIEVEMENT OF THE DEPARTMENT OF ENERGY'S MISSION GOALS THROUGH SYSTEMS BIOLOGY

The primary cross-cutting theme of Genomics: GTL is systems biology—with the goal of developing predictive models of system function. The committee strongly endorses the notion that being able to predict the properties of DOE's target systems would revolutionize energy-related and environment-related biotechnology. However the challenges to the achievement of the Genomics: GTL mission are immense; and, although they are likely to be solvable, the precise route to be taken is uncertain and subject to debate.

Where to begin depends largely on how one defines and sets priorities among the specific factors that limit progress. In its current form, Genomics: GTL has two parallel tracks: use of a traditional research funding process to identify and fund relatively large-scale, often multi-investigator projects focused on specific biological problems and a multidecade plan to construct and operate facilities that target high-throughput production and analysis of proteins, protein complexes, and microbial systems within which the proteins express their potential. The current facilities model assumes that progress in microbial systems biology is limited by lack of knowledge about proteins and their derived attributes within biological systems of interest to DOE, that acquiring such knowledge will require high-throughput facilities that can solve the problem by applying appropriate technology, and that knowledge of and access to all proteins in a range of target systems will revolutionize microbial systems biology in a way analogous to how genome sequencing has transformed biology in general.

Whether one agrees with that model depends largely on how one defines the primary barriers to progress. This section provides an overview of important gaps in knowledge and technology that must be filled to facilitate and expedite the achievement of the long-term goals of Genomics: GTL.

The Challenges

Multiple Scales of Systems Biology

In the context of DOE's bioenergy mission, the relevant properties of a given system or molecular machine may span essentially all scales of biological organization, from simple binary interactions that may function at the subcellular level as regulatory switches to superdimensional interactions in microbial consortia, where the emergence of a particular property may depend on organism-organism or organism-environment interactions.

Diversity of Microorganisms

Many of the applications of bioremediation, carbon sequestration, and biofuel production will occur in situ, using endemic taxa. Biodiversity in terrestrial and marine environments is poorly understood. At the taxonomic level, recent estimates suggest that 1g of pristine soil may harbor as many as 10^6 distinct prokaryotic taxa and that most of the taxa are rare (Gans et al., 2005). Current sampling strategies, such as those based on metagenomic phenotype analyses (Williamson et al., 2005) or sequence analyses (Venter et al., 2004; Tringe et al., 2005), are adequate for sampling only the most abundant of these organisms with any certainty. Rare but stable components of ecosystems may contribute important properties to system function, but we lack routine methods for characterization of ultra-rare genomes, let alone for understanding the majority of such species.

Diversity within species is also relevant to DOE's bioenergy mission. In recent years, high-quality assemblies of many microbial genomes have revolutionized microbiology. However, the sequences themselves are blueprints of reference genotypes, and in most cases the fraction of natural diversity that such reference genomes encompass is uncertain. Metagenomic sequencing is a means of surveying DNA from complex consortia. The results of several metagenomic projects highlight the wealth of information likely to accrue from ecosystem-level genome sequencing, both within and among species (Venter et al., 2004; Tringe et al., 2005). Nevertheless, current strategies for shotgun sequencing limit analysis to the most abundant genomes and typically yield only fragmentary assemblies. Recognition of that has sparked renewed interest in developing methods for enriching and culturing recalcitrant and rare species. Similar gains are likely to be realized by implementing nucleic acid normalization methods, such as Cot enrichment or suppressive-subtractive hybridization, which are well established in other genome investigations (for example, Yuan et al., 2003; Galbraith et al., 2004). The combination of new culture-independent technologies—such as single-cell sequencing, for example, the work of DOE-funded investigator George Church (DOE, 2005c)—with efforts to enrich and set priorities among specific genomes for analysis may have potential to expand our view of ecosystem biocomplexity.

Inferring Function from Sequence and Structure

Implicit in many of the arguments about the usefulness of data-gathering exercises, such as genome-sequencing projects and structural genomics initiatives, is the assumption that such data will illuminate the functions of many gene products. Several problems lead us to question the wisdom of accepting that assumption uncritically. First, the term function is imprecise (Ning et al., 2003; Fraser and Marcotte, 2004). Depending on who is the consumer of the information, function may refer to the biochemical activity of the isolated gene product, its role in metabolic or signal-transduction pathways in the cell, the phenotype of its knockout in a cell or model organism, or any combination of these. Even the seemingly simplest to ascertain, biochemical function, turns out to be loosely coupled to simple "determinants," such as sequence and structure (see below). We are sympathetic with the need to obtain such information, given that 30-50 percent of the genes in most newly sequenced genomes have no established biochemical or cellular functions, but the goals of the Genomics: GTL program seem too sophisticated to adopt a single, restrictive view of function as a guiding principle.

The second problem is that much functional annotation depends on relating the sequence of a gene product to other sequences of known biochemical function, and the database of annotated functions is simply not as reliable as it needs to be (Gerlt and Babbitt, 2000). It is estimated that as many as 50 percent of the functional annotations based on sequence comparison may be wrong, at least in part; as more sequences are determined and annotations increase, the problem is likely to be compounded and should be rectified as soon as possible.

The third problem is that sequence and structure information rarely, if ever, increases understanding of whether a gene product has important interactions with others in the cell and, if it does, how those interactions affect its biochemical and cellular roles. Relying too heavily on such data-gathering for functional annotation risks taking a step backward, away from the more complex pictures demanded by systems biology.

Fourth is the increasing recognition that many or most of the gene products in higher organisms—and many in bacteria—have more than one function, no matter how one defines the term. Sequence and structural analyses hardly ever provide information on more than one prospective function and are usually silent about the conditions under which that function is biologically relevant. Focusing on a single function misses the point that systems biology is meant to address.

The most serious problem in using sequence and structure information to deduce function is that function changes much more rapidly than the other properties. In many instances in the database, two gene products that have more than 80 percent sequence identity—and correspondingly high structural identity—have biochemical functions that are completely different because of one or two changes in critical amino acid residues. Even when both sequence and tertiary structure are very similar, biochemical function may change. The ThiJ/DJ-1 su-

perfamily of proteins—which has members that are bacteria, archaeans, and eukaryotes—is a striking example. Members of that superfamily can share 40 percent sequence identity and monomer protein folds that superimpose to less than 1 angstrom of root-mean-square deviation and yet they have completely different functions as a result of different oligomerization states that bring completely different residues from one subunit into contact with those from the other (Wilson et al., 2005). Natural selection, which guides genome evolution, guarantees that examples like that will be common. The sequence of a protein can change easily. Overall structure is more robust but can still be affected considerably by a few mutational events. And neither is under the control of natural selection, which acts only on function.

The attempt to infer function from sequence and structure alone is an exercise in futility, particularly where complex phenomena are concerned. However, the committee does not wish to leave the impression that such data are of no value. On the contrary, they are a valuable part of the panoply of information that must be obtained to understand function. But they are data at the lowest level of complexity, involve the most routine and readily available technologies, and should not be a cornerstone of a program designed to advance the cutting-edge field of systems biology.

Computational Challenges

The Genomics: GTL roadmap states that "the goal is to create increasingly accurate mathematical models of life processes that enable predictions of cell and community behavior and create new and modified systems tailored for mission applications." The systems biology approach of Genomics: GTL will integrate experiments, data acquisition and processing, modeling, and simulations in an iterative process in which model predictions inform experiments and experiments inform model development. The development and analysis of increasingly accurate models at all levels of biological organization pose mathematical and computational challenges. Vastly different time scales of the different biological processes can pose additional numerical challenges in simulation. The appropriate level of model complexity needs to be found because of tradeoffs between level of detail and computational complexity. The more detailed a model, the larger the number of variables and parameters, which further increases the difficulty of model validation and inference. Stochastic noise inherent in many of the processes makes accurate parameter estimation difficult. Many of the processes depend heavily on environmental conditions, so experiments with a wide array of environmental conditions will need to be integrated. We outline some of the modeling and simulation challenges on the different scales (see also *Mathematics and 21st Century Biology*, NRC, 2005).

At the cell level, the ultimate goal is to predict the cell phenotype from its

genotype and environmental conditions. A diverse set of mathematical and statistical approaches have been developed to unravel the topologies of networks (for example, "wiring diagrams") that link cellular components, including gene networks, regulatory networks, and metabolic networks. Those networks provide a static view of cellular interactions. To model these accurately, spatiotemporally resolved data on numerous cellular processes must be integrated into dynamic models. That involves a large number of diverse components that range from rare to abundant. Many of the mathematical models that describe cellular processes are based on systems of ordinary differential equations. It is a mathematical framework that assumes that components are spatially homogeneous and can be approximated by continuously varying densities. That works well as long as the components under consideration are abundant in the entire cell with little spatial variation. Many components, however, are produced only in small quantities that are spatially localized within the cell and that exhibit considerable stochastic variability. Thus, accurate models of cellular processes will be mixtures of deterministic and stochastic models of discrete and continuous variables that vary both spatially and temporally. Little mathematical theory is available to deal with such models.

Microbial populations consist of genetically diverse cells, and individual cell responses even to the same environmental stimulus may vary. Ecological models of populations consider average responses and do not take genetic variation into account. Adding genetic heterogeneity to mathematical population models can considerably increase the complexity and dimensionality of the models. A considerable body of work deals with quantitative characters (see, for instance, Turelli and Barton, 1990 or Nuismer and Kirkpatrick, 2003). Including sequence variation in ecological models, however, would require a model framework that has not been established.

Modeling of ecological communities has a long history in mathematical ecology. Classical models use the simple framework of ordinary differential equations and consider few interacting species that are often highly unstable, exhibiting oscillations and chaotic behavior. Simulations of systems of larger numbers of interacting species have shown that they exhibit rich behavior but can be stabilized through interactions (see, for instance, Williams and Martinez, 2004). Attempts to gain a better understanding of the behavior of communities that consist of many interacting species have begun only recently, and there is no general theory that would allow prediction of the behavior of communities of thousands of interacting species. In addition, spatial heterogeneities and stochastic effects are rarely taken into account. Both are probably important, especially in soil microbial communities. There has been no attempt to integrate across all scales from molecules to ecosystems. Such multiscale models would span spatial and temporal scales of many orders of magnitude and would need to incorporate the genetic variation that is present in a community of interacting species, which

would require considerable computing power. Given the importance of this type of modeling activity to DOE's missions, it is essential that the Genomics: GTL program seek to rectify the situation.

In addition to modeling and simulation challenges, some of which have been described above, there will be a need to develop bioinformatics tools further. The Genomics: GTL program will produce data at unprecedented rates and diversity, and they will need to be captured, archived, and annotated, preferably in an automated way. The types of data will go far beyond simple sequence data. They will include structural data from which three-dimensional models of molecules will be built, imaging data that track individual molecules in cells, and data that will track physiological responses of communities that consist potentially of thousands of different microorganisms in a wide array of environmental conditions. New methods in mathematics and computational biology to analyze such complex data will need to be developed, as will software and hardware to allow researchers to use these diverse datasets.

Issues to Be Addressed by Genomics: GTL Program and Facilities

A variety of issues will need to be addressed in the course of achieving the long-term goals of Genomics: GTL. Among these is the need to improve and implement genomics-enabled, high-throughput studies of genetic diversity in Genomics: GTL environments. The resulting information would contribute greatly to understanding aspects of ecosystem-level population biology, evolution, and function that are currently lacking and are critical to the mission of the Genomics: GTL program. The following are insights:

- Description and then development of predictive models for how complex microbial consortia respond to natural and imposed selection.
- Identification of the genomic diversity best suited to manipulation of Genomics: GTL target processes, for example, remediation of specific contaminants in unique environments.
- Characterization of genotypes and the genes and proteins that most strongly influence system function.
- Understanding how human intervention may alter community structure and function, and identifying and quantifying related risk factors, if any. In particular, if genetically modified organisms are to be released into open field settings for bioremediation, DOE should make strong efforts to gain public acceptance for such release.

Central to the Genomics: GTL mission is the need to identify the molecular machines that underlie target processes. The challenge is not simple, in that what we conceive of as distinct molecular entities may exist on any of a number of scales, from coherent protein complexes, to physically unrelated complexes in a

single cell, to proteins present in unrelated taxa but where complementary activities yield a desired outcome. Major challenges include the following:

- Identification and functional characterization of the proteins and the complexes that underlie Genomics: GTL target processes.
- Formulation of models that predict the function of these key cellular or organismal components in situ.
- Development of strategies to improve the efficiency of these "molecular machines."
- Improving methods for analysis and interpretation of gene and protein function in heterologous systems, including both computational and experimental approaches.

Much of the progress envisioned under Genomics: GTL will require derivation and application of novel technologies, principles, and computational approaches that permit biologists and engineers to understand and manipulate the Genomics: GTL ecosystems. Key milestones toward this broad goal are

- Improved technologies for surveying taxonomic and genetic diversity in target environments, including the development of tools for both culture-dependent and culture-independent methods and strategies to deal with ultrarare genomes.
- Development of experimental tools, concepts, and mathematical methods that can model transient and stable states and identify the control points for particular system parameters.
- Establishment of predictive models of microbial behavior during discrete phases of development and in response to external biotic and abiotic stimuli.
- Development of new methods and instrumentation to measure key biological parameters that may be relevant to system function, including metabolite flow and protein function in vivo and in situ.
- Establishment of methods to reproduce native ecologies in the laboratory or to analyze them in situ.
- Understanding of the consequences and frequency of events that may alter population function, such as horizontal gene transfer, alterations of physical-chemical environments, and introduction of nonnative species.

Broadly stated, the goal of systems biology is to uncover properties of organisms and communities that would not be made apparent by analysis of their components in isolation. Few would argue that our current understanding and methods are adequate to develop a quantitative model of even one bacterium, much less a collection of genotypes in a single species, and even less an entire ecosystem. Systems biology suffers from a dearth of general principles that can guide further study. Nevertheless, although the challenges are substantial, the

situation is not impossible, and the general goal of understanding system function is worthy. Evolution solved the problem by making systems work. A question that is relevant to the proposed Genomics: GTL facilities is, What is the best route to an understanding of these systems?

Despite the temptation to draw analogies, biological systems are not similar to electrical circuits. Electrical circuits are composed of a rather small diversity of entities, whereas biological systems are composed of a multitude of dissimilar parts, even to the point of adaptive variation in apparently common components. Given the scale and complexity of the challenge, it is not obvious that a complete catalog and partial analysis of all proteins in a few target genomes would be a major advance toward understanding and predicting the function of complex microbial systems.

In going forward, some questions need to be considered:

- Do we need a complete catalog of all parts, or do we want only to describe some parts?
- If we want to focus only on some parts, what are they, and how do we identify them?
- Given a complete catalog of parts, do we care equally about all interactions between parts?
- Do we care equally about where every part resides, in every genotype, and under every condition, or do we care more about some parts under specific conditions?

It stands to reason that for any given situation only a small portion of the "parts" need to be understood in great detail to model system function. Such *control points* might be genes, alleles, proteins, metabolites, genotypes, or even taxa and their relative spatial distribution. Moreover, the specific control points may vary between different systems and situations. Developing methods of identifying such control points would be a major step toward predicting system function. Moreover, it would allow research to focus on relevant aspects of systems rather than all aspects irrespective of their relative importance.

It is not clear that describing the protein components of individual cells or multiorganism consortia is a necessary first step toward systems biology. Forefront science requires taking a step beyond that, into a detailed characterization of target systems by highly interdisciplinary teams of scientists. In the course of such an endeavor, enumeration of individual components and their interrelationships emerges naturally because it is driven by the complexity of the specific systems under study, as revealed by an integrated approach to their analysis. The technologies (both new and existing) are more effectively limited to those which are appropriate, and this is both time-efficient and cost-effective.

3

Implementation of the Genomics: GTL Program Plans

As noted in Chapter 1, the Department of Energy (DOE) is charged with promoting scientific and technological innovation in support of its overarching mission to advance the national, economic, and energy security of the United States (DOE, 2005a). DOE is a key supporter of non-health-related biological research, and DOE's scientific strategic goal emphasizes both the production of new knowledge through science and the creation of new research capabilities.

It is the assessment of the committee that the goals of DOE's Genomics: GTL program are consistent with both elements of DOE's scientific strategic goal. But as currently envisioned, Genomics: GTL is focused almost exclusively on microorganisms. Although work on microbial systems is well justified, plants also represent a major pathway to the production of bioenergy, they play an important role in carbon sequestration and global nutrient cycles, and they are potentially useful for bioremediation. Thus, the absence of targeted research within Genomics: GTL on relevant aspects of plant biology is a serious omission. Consistently with that view, the Energy Basic and Applied Sciences Act of 2005 calls for an emphasis on both plants and microorganisms. DOE already has a modest investment in energy-related aspects of plant biology in its Energy Biosciences Program. The committee suggests the inclusion of plant biology research in the Genomics: GTL program where appropriate.

In Genomics: GTL, DOE proposes to use systems and synthetic biology approaches to achieve a predictive understanding of microorganisms and to mine these untapped resources. The approaches are well matched to DOE's history as the founder of the Human Genome Project (Roberts, 2001) and represent a logical extension of DOE's scientific vision and capabilities. No other federal agency

is playing a lead stewardship role with respect to microbial systems and synthetic biology. In addition, Genomics: GTL's somewhat unique practice of funding large teams with large grants is essential, and the committee believes that it should be maintained. The Genomics: GTL program has produced valuable scientific results (DOE, 2003, 2004, 2005c), and its planned research promises to generate additional important and useful results. The committee enthusiastically endorses DOE's plan to enlarge the program to $200 million per year for basic research and further endorses the focus of this research on long-term goals for energy production, environmental remediation, and the mitigation of global climate change. In particular, the committee notes that the ability to produce cost-competitive ethanol from cellulose and hydrogen biophotolytically from water or fermentatively from other carbon substrates, the development of biological solutions to the many recalcitrant problems of legacy wastes, and the attainment of an increased understanding of the role of microbial communities in global carbon cycling to enable the development of carbon-sequestration techniques for addressing climate change are all worthy goals that are highly suitable to DOE's missions. The committee wholeheartedly supports those goals and the conclusion that the best way to achieve them is through a systems biology approach. We further endorse, with enthusiasm, the ambition of the Genomics: GTL initiative to place DOE at the forefront of systems biology research, as it has been in genomics.

Recommendation 1: The committee recommends that DOE and the nation give high priority to genomics research aimed at achieving DOE's mission goals.

However, the committee disagrees with the plan to create four facility types—for protein production and characterization, characterizing and imaging biomolecular machines, proteomic analysis of microorganisms, and modeling of microbial community cellular systems—sequentially. In place of that plan, we propose a set of integrated, problem-oriented genomics-enabled facilities that will focus on pioneering technologies rather than duplicating existing technologies.

THE PROPOSED GENOMICS: GTL USER FACILITIES

The committee agrees that the technologies associated with each of the four proposed facilities are critical to providing fundamental and quantitative understanding of biological processes relevant to the long-term missions of DOE and the sought-after predictive capability for biological behaviors at all levels from molecules to whole microorganisms. However, we believe that to advance DOE's missions and to strengthen the nation's systems biology research capacity, a

parallel and integrated approach to creating facilities should be taken instead of a sequential approach.

First, DOE's argument for the sequential creation of the four new facilities is flawed. It rests on an assertion that the facilities will greatly speed the achievement of the program's long-term goals for research to advance energy production, environmental remediation, and carbon sequestration. In their absence, DOE asserts that the achievement of the goals could take as long as a century because the scientific advances needed to reach productive applications of new technologies to the problems will be severely retarded by the lack of appropriate tools. Department officials have expressed to the committee the view that the creation of the facilities could direct and speed the research efforts to achieve the desired goals in far less time—for example, in 30 years. But in the current plan, the time for each facility to be designed, constructed, and come on line for operation is estimated to be about 6 years because DOE assumes that budgets for the facilities will be appropriated one at a time. That suggests that the complete program capability cannot be achieved until the completion of the fourth and final facility roughly 24 years from now. If for any reason there is a delay in the appropriation of funds for the creation of any of the four facilities, even 24 years may not be enough time to reach full program capability. The committee therefore finds DOE's estimation of the time for the program to come to full capability and produce the desired result highly unrealistic under the current plan. The reductionist approach, moving from simple to increasingly complex systems in a period of 24 or 25 years, is not an efficient way to achieve the program's goals. Building four individual facilities in sequence is fraught with too many uncertainties and requires much too long a time. Even if we assume that the conglomerate of different facilities will, in the end, be able to come together to achieve useful progress toward solving the cutting-edge problems being addressed by the Genomics: GTL program, the time to completion is so long that there is considerable risk that the systems biology train not only will have left the station but will be at some other station when everything is finally on line. In this fast-changing era of genome-enabled science, it is not sufficient to require that many years be spent in pulling together the tools needed to make progress. That approach is unlikely to place DOE in a leadership position, either intellectually or in terms of research results.

Second, the proposed configuration for the Genomics: GTL facilities is based on an underlying assumption that large-scale generation of reagents and data is the best way to apply a systems biology approach to the grand challenges of the Genomics: GTL program—bioremediation, bioenergy, and carbon sequestration. That approach first parses the organisms that might be subjects for study into their components, assembles a parts list, inventories the parts, and then reassembles them into interacting complexes, networks, and pathways. Underlying the plan is the belief that data and reagents abstracted from the individual organ-

isms will enable DOE's missions. The implied assumption for the current facility model is that the best way to solve the problems in systems biology is to operate several separate facilities that concentrate on specific aspects of the question. The committee finds that assumption questionable. Systems biology is a cutting-edge field because the properties of organisms and groups of organisms cannot be predicted from simple considerations of the behavior of their parts. Therefore, obtaining much more information about the parts is not the best way to reach the goal of systems biology. A more consistent view would be that the properties of complex systems can best be understood by studying complex systems. Such studies need to be conducted by interdisciplinary teams of experts that include genomicists, geneticists, physiologists, biochemists, biophysicists, computer scientists, engineers, and mathematicians.

The rationale for the current plan appears to be a direct analogy with genomics, especially large-scale sequencing, as practiced during the last decade. The committee finds, however, that the analogy is weak at best. The economies of scale that made large genome projects successful have not yet been obtained in work with proteins or in most aspects of systems and synthetic biology. It is not clear when technological advancements that deliver such economies of scale will become available for protein analyses and systems and synthetic biology. Hence, DOE should be cautious about embarking on construction of large-scale infrastructure that assumes that such economies are readily available.

We also find a lack of equivalence, for example, between the high-throughput production and characterization of individual proteins as a means to advance systems biology and the effect of DNA sequencing on characterizing the genome. Understanding what the various parts of an organism do is one aspect of understanding the organism as a whole; and knowledge of component functions is useful, and perhaps essential, for modeling the behavior of the organism, but it is insufficient for predicting how various component functions work and interact together. Hence, if the short-term goal of the Genomics: GTL program is to be able to predict the behavior of systems so as eventually to engineer microorganisms to serve the nation's energy-related needs, this model does not address the most important needs of such a program, nor does it push the research frontier forward. The goals of the protein-production facility are modest compared with its size and budget and can easily be met much less expensively either by several components in different types of facilities or by outsourcing to academic and commercial facilities (Box 3-1). The proposed protein production facility would be doing today's research, not tomorrow's.

It is true that the Human Genome Project was both important and timely because it provided reams of data, but that was before we had sequence information. Now that we have it, additional data-gathering and reagent production, although necessary for the field as a whole, will not propel the field forward in the same way that genome sequences did, because the cutting-edge questions are very different. Furthermore, sequence information can lead to immediate leaps in

BOX 3-1
Results of a Survey of Groups Engaged in or Contemplating Large-Scale Protein Production

The committee contacted personnel at different entities and asked each of them to estimate the cost of setting up a large-scale protein-production facility in an existing building that would begin operation 5 years from now. The committee also asked for estimates of the annual operating costs to produce 10,000 proteins per year in quantities of 3-10 mg with purity suitable for characterization or crystallization. No cost estimate was requested for producing specific reagents for the proteins, because the costs of production depend heavily on the nature of the reagents, and the Genomics: GTL roadmap (DOE, 2005b) did not provide sufficient specifications for such estimates. Responses were received from five entities representing small and large companies and academic research centers (Center for Advanced Biotechnology and Medicine, New Jersey; CODA Genomics, Inc., California; Invitrogen, Inc., California; Modular Genetics, Inc., Massachusetts; and The Scripps Research Institute, California). There was remarkable unanimity in the responses. All agreed that the task was approachable with contemporary technology. All agreed that the infrastructure investment would be far less than 1 year's running costs. The range of running costs quoted was $3-$8 million per year. Furthermore, two of the entities seemed on the verge of scaling up their current efforts to near or at the ranges specified by tapping their existing funding sources. Most responders felt that the process was scalable across a wide range and that economies of doing this in a single location, as opposed to a group of locations in parallel, were modest at best.

Although the committee acknowledges the need for infrastructure for protein production, these findings suggest that a nine-figure construction project for a protein-preparation facility that would not come on line for 6 years may not represent the best strategy for meeting the infrastructure need. Given the magnitude of current Genomics: GTL activities, it appears that program needs could be satisfied in a cost-effective way by using available technology and getting open bids from currently active players. Thus, if funds could be made available at the magnitude originally specified in the Genomics: GTL program, there ought to be sufficient resources to fund a number of smaller vertically integrated efforts that could not only produce the proteins but also amply apply them toward the research goals of the program. The best ways to produce the necessary specific reagents require further study, but it is worth noting that some approaches, such as antibody display libraries, potentially offer enormous economies of scale.

understanding through analysis of individual genomes but especially through comparison of genomes. Such understanding does not come from a shopping cart full of proteins, however useful they may be to bench scientists. The payoff from such resources is a long-term one, and this greatly lessens the intellectual impact of DOE's contribution. Simple service functions provide useful tools but will not themselves advance understanding in a major way. They reduce DOE to the role of enabler rather than leader. As stated above, the knowledge gaps in systems biology go beyond an understanding of the biochemical role of each protein in an

organism and beyond an understanding of how the proteins associate with one another and how the associations modulate individual function. The greatest progress in systems biology can be achieved only by solving the problems of integration of function at the level of pathways, organisms, and colonies and communities of disparate organisms. The committee concurs that for the Genomics: GTL program to achieve its mission goals, the facilities model must directly aim to address those needs.

AN ALTERNATIVE MODEL FOR IMPLEMENTATION

The sequential-facilities model appears to be based on a belief that the barriers to achieving the short-term and long-term goals of the Genomics: GTL program are related primarily to the ability to obtain material (such as proteins) for study. As a consequence, DOE's proposed plan is technique-driven. It is the view of the committee, however, that the real barriers are related to our ability to study and understand complexity and that a more problem-oriented approach is needed. Our list of key barriers includes barriers to measuring metabolite flow and other biological characters in vivo; to an understanding of interactions in communities of microorganisms that would enable prediction of the effects of introducing new metabolites or new or engineered organisms into the community; to a predictive understanding of how organisms respond to the introduction of new enzymes and pathways; to our ability to model microbial behavior in many developmental states, including stationary phase and sporulation, and in proliferation; to an understanding of transient states; to the acquisition of tools, including mathematical models and concepts, to enable prediction; to the use of general principles; and to adequate characterization of microbial diversity in the target ecosystems. The timely execution of the Genomics: GTL program and the achievement of DOE's mission goals are much better served by investing in comprehensive research programs that drive technology development than by investing primarily in infrastructure.

Because the currently proposed model is technique-driven, the facilities proposed for it are fairly specialized, single-purpose facilities. We believe that it is incorrect to assume that an interdisciplinary problem can be addressed best with a set of cooperating, but independent, specialized facilities, especially if they are created sequentially. We believe that interdisciplinary problems require interdisciplinary approaches from the outset.

To understand what an optimal model might look like, we consider the facilities needed to meet the short-term Genomics: GTL goal of predicting the properties of microorganisms. Examining several aspects of alternative models in the light of that mission provides a framework for a different model.

- *Vertically integrated facilities vs. reductive specialization.* It seems inherently contradictory to assume that a coherent approach to modeling the properties

of microorganisms will "emerge" of its own accord from the activities of a set of autonomous facilities. If we assume that such an interdisciplinary problem requires, at a minimum, good communication among researchers from many backgrounds, that alone would favor a vertically integrated structure in which such scientists all work on a common set of problems at the same facility. Another advantage of integrated facilities is that basic scientists would be working with engineers, so discoveries would be translated into applications efficiently.

- *Institute vs. service model.* As anyone who has watched the transition of a startup biotechnology company from a research-oriented to a customer-driven business can tell, a service facility is concerned primarily with solving day-to-day technical problems and meeting the demands of those it serves. If its technology becomes outdated or less important, the service facility is marginalized in its field. An institute, in contrast, is concerned with the big picture, can contain people of many backgrounds, is not wedded to a particular technology, and can adjust its strategies more easily as its field progresses. It seems logical that more-talented people will want to work in the institute model and that an agency that operates the institute will be more prestigious and have more influence than one that runs a service center.

- *Partnerships and integration with a local academic or research community vs. an autonomous DOE-only model.* One of the biggest obstacles to progress in any organization is the "not invented here" effect. The tendency of many institutions to look inward restricts progress and retards the integration of new technologies. Thus, a facilities model in which entities outside DOE participate is likely to be more open-minded, flexible, and innovative than one that draws its personnel, tools, and infrastructure exclusively from one source. In addition, a DOE initiative that is isolated from the broader biological community may not be taking full advantage of outside expertise or of opportunities to communicate its own expertise to others. The committee suggests that the optimal placement for Genomics: GTL facilities would be close to strong academic or industrial research establishments. That should be a key criterion in selecting the location of one or more of the proposed Genomics: GTL facilities; proximity to a community of cutting-edge biological research is essential to the long-term success of the Genomics: GTL program. The exchange of ideas and information among academe, industry, federal agencies, and other research entities will leverage human, scientific and financial resources. Although it might be more feasible for DOE to locate facilities near a national laboratory, the committee notes that not all national laboratories are close to strong academic or industrial research establishments.

- *Parallel vs. sequential.* The advantages of a sequential model are that one can learn from mistakes and build on success. But we conclude that the disadvantages of the sequential model outweigh those potential advantages. One disadvantage is the impossibility of correct timing. Built into a sequential model are assumptions about the pace of advances and the changing needs of a field over

time. No one can predict either with any confidence. Thus, there is a great risk that by the time the first facility finally comes on line, it will be obsolete technically or irrelevant to the needs of the later facilities, which will have changed because of outside developments over which the program has neither influence nor control. Another disadvantage is that only the last facility in such a model interacts with all the others from its inception; the growth of many facilities in isolation from others breeds a culture not of cooperation but of independence. "Parallel" facilities would avoid those pitfalls.

What alternative model follows from those considerations, and why? The committee proposes a model that has the following features:

- It consists of institute-like facilities, each focused on a particular problem or theme in systems biology that will advance both the short- and long-term goals of the Genomics: GTL program. Some possible research areas include remediation of plumes of toxic waste, understanding and predicting the consequences of adaptive evolution (including that originating from horizontal gene transfer), modeling microbial communities, and developing a systems-level understanding of microbial stationary phases, sporulation, and other non-proliferative developmental states.
- Each facility develops technology to support its research. Technology development is driven by research needs.
- The model is vertically integrated to address a problem in concert with many techniques, some of which are developed there. Some work will be collaborative, and some will be contracted. Outsourcing to acquire materials and technology could be a cost-efficient strategy that allows more rapid acquisition of critical capabilities. It is important that a vertical focus directs priority-setting toward specific ends in a manner that is likely to be more or less peculiar to individual challenges.
- Each facility is near major academic, private, or federal research centers. Being close to institutions that have excellent biology, mathematics, and computer science programs is important because access to researchers with related interests allows the facility to draw more completely on and leverage the expertise and interests of surrounding universities, institutes and industries. It involves public-private partnerships with both academe and industry. And it integrates the Genomics: GTL program into the larger biological research community.

In facing the reality of rapid scientific progress and the timeliness of the proposed approaches, the committee strongly encourages DOE to rethink its user facility construction plans and create institute-like facilities each of which combines the capabilities of the original four planned types in a vertically integrated manner. That would allow each facility to tackle all aspects of a problem or small set of problems in parallel and potentially to achieve goals more quickly. Table 1

in Section 5 of the Genomics: GTL roadmap document provides three "conceptual science roadmaps" for microbial energy and environmental processes, and Table 2 in Section 5 provides three more for natural systems (DOE, 2005b). The first vertically integrated facility could reduce its focus from all six conceptual science roadmaps and concentrate on only one or two. For example, the first facility could concentrate on fuels by working on the systems biology of cellulose conversion and the production of hydrogen and high-hydrogen fuels by using sunlight. That focus would be consistent with the committee's view that investment in bioenergy is inadequate despite the urgent need to find alternatives to petroleum-based fuels.

A concentration on the science underpinning fossil-fuel replacements would have two advantages. First, technologies would be developed with a greater emphasis on the needs of alternative-fuel research than on serving the scientific community as a whole. The synergy between researchers and technology developers should also speed technology outcomes and keep them at the cutting edge. Selection of organisms to be studied should be based on their application to bioenergy production, irrespective of whether they are common or rare. Therefore, DOE could take a lead role in identifying and developing key organisms into model systems for systems biology research relevant to bioenergy production.

Second, experience gained from building and running the first facility could be used to refine approaches and improve planning and execution of the second facility. The second facility would focus on one or two additional aspects of the items listed in the tables in Section 5 of the roadmap document. A third and perhaps fourth facility, if deemed necessary for continued progress, could be planned and constructed by using further design and execution refinements in the longer term. In that way, work on some of the initial problems could be reaching long-term goals long before the estimated 24 years proposed for full design and construction of consecutive facilities had elapsed.

It would be necessary for DOE to select problems for each facility in the chain to focus on as it comes on line. The committee suggests that there are compelling reasons for DOE to give top priority to the creation of a "bioenergy institute" that also focuses on carbon cycling in the context of bioenergy. First, the U.S. (and global) economy is increasingly vulnerable to oil shocks caused by political unrest, terrorism, and natural disasters. A recent analysis conducted by Securing America's Future Energy and the National Commission on Energy Policy concluded that even "small incidents" that reduced global oil supply by 4 percent would cause oil prices to increase dramatically to more than $161 per barrel (SAFE and NCEP, 2005). Second, as several National Research Council reports have concluded, we need to act now if we are to have any chance of stabilizing greenhouse-gas emissions (NRC, 1992, 2003a), given the 100-year residence time of carbon dioxide (CO_2) in the atmosphere and the 30- to 50-year lifetime of capital stock in the energy industry. Because reduction of CO_2 in the

atmosphere and carbon cycling are related to the use of bioenergy, the bionergy institute should have a secondary focus on carbon sequestration. Third, recent advances in biology—such as rapid sequencing, directed evolution, and whole-genome synthesis—may enable us to design biological systems that can generate affordable, carbon-free energy (see Chapter 2), and thereby to reduce the costs of mitigating CO_2 emissions (IPCC, 1997). Later facilities could have carbon sequestration and bioremediation as their primary foci.

The revised model would greatly improve the cost effectiveness and efficiency of DOE's investment and optimize the achievement of useful scientific results. The committee agrees that there is a sound scientific case to be made for systems biology at DOE, but the approach to the order, scope, and scale of the facilities needs rethinking. The committee's alternative model could deliver some of the hoped-for scientific output more quickly and more efficiently.

Recommendation 2: DOE should revise its plans for creating four single-purpose technology-driven facilities in sequence. Instead, DOE should create up to four institute-like facilities that each contain all the capabilities of the original planned facility types—protein production, molecular imaging, whole-proteome analysis, and systems biology—in a vertically integrated manner. Each facility should focus on one or two of the DOE mission objectives and develop short-term, medium-term and long-term goals to chart a course for the program. Short-term milestones should be used as a metric for independent evaluation.

The committee concurs that having a single physical space for the first integrated facility is important, although some of the investigators should be encouraged to participate from remote sites. Proximity will enable a large-scale research program with efficient, coordinated, and complementary activities among a team of investigators who have diverse expertise and no physical barriers. Day-to-day interactions among the assembled scientists in various disciplines (biology, chemistry, physics, nanoscience, computing, and engineering) are essential to overcome cross-disciplinary barriers and to generate novel approaches to and productive outcomes for the intended goals. The Institute for Systems Biology and the Department of Systems Biology of the Harvard Medical School are two examples of integrated facilities that facilitated and advanced biomedical sciences.

Later phases of facility implementation will be based on the experience gained in the first facility and on progress in the proposed research at the time of funding. It is conceivable that there will be more than four facilities at the end, with different emphases and strengths and consistent with DOE's goals. Some of the future facilities should be flexible enough to include more than one site per

facility so as to leverage additional resources and partnerships that would broaden the portfolio of the Genomics: GTL program.

Our proposed alternative plan of implementation is based on the following merits.

- Vertically integrated facilities would establish the Genomics: GTL program in a leadership position to launch a world-class, comprehensive, integrated research and training program in systems and synthetic biology. They would create a paradigm shift in biological research that will integrate data from a broad spectrum of spatial and temporal scales to advance understanding of biological phenomena to be able to predict or alter capabilities for optimal performance under field conditions. The facility would provide an intellectual and physical environment for both multidisciplinary teams and individual-based research.
- The research programs of the facilities would be built on overarching biological themes relevant to the DOE missions in energy production, environmental remediation, and carbon sequestration. They would involve diverse disciplines, including genomics, genetics, physiology, biochemistry, structural and computational biology, nanoscience, and engineering. The facilities would provide an intellectual and physical environment for both multidisciplinary teams and individuals pursuing research in relevant missions.
- New technologies would be developed in the facilities on the basis of well-justified scientific problems. The technologies will be aimed at particular ends rather than being ends themselves.
- The successful development of the integrated facilities would attract investigators around the country to use them. The resulting scientific discoveries and technology development can be expected to benefit not only a subset of biologists but a broad spectrum of scientists and engineers in different disciplines. Because of the diverse disciplines of the investigators, the integrated facilities are likely to have complex organization charts. That will leverage the experience of DOE to administer this new research enterprise.
- Modeling plays a central role in studying and understanding complexity. New computational approaches and tools would be developed in the facility to promote synergy between modeling and experimentation at both bench and field level.
- The first facility constitutes a pilot to validate the hybrid systems approach and to identify roadblocks to be addressed in the later facilities. All facilities can be designed so that they will not be outdated by the rapid pace of scientific discovery and technological development.
- The vertically integrated facilities lend themselves to a staged investment with expandable bases, flexibility to shift directions without losing prior investments, leverage, and open-source positioning.

Criteria for Selecting Contractors and Locations for Genomics: GTL Facility Awards

The committee has concerns about other aspects of DOE's plan. The locations for the user facilities should be selected in an open and all-inclusive competitive process that provides adequate opportunities for universities and industry to partner with DOE and its national laboratories. The committee strongly believes that such an open and inclusive competition will provide DOE with a robust and enabling facility that could strongly support systems biology research. The criteria for the selection of contractors should include innovations of the project plan relevant to DOE missions, investigator qualifications, management organization, educational outreach, technology dissemination, strategy for maintaining the interface between Genomics: GTL research and industry and other entities that conduct translational research, intellectual-property management plan, proximity to a concentration of high-caliber participating scientists, and possibly provision of matching funds by the applicant institutions.

The committee encourages DOE to consider cost sharing by applicant institutions for design and construction because timely establishment of the proposed facilities is crucial in the fast-moving field of systems biology. For example, allowing the successful applicant to fund new construction or renovation of an existing facility upfront could greatly speed up the process. As federal funds become available through the appropriations process, DOE could then "lease-purchase" the facility to eventually acquire it from the private sector partner. If DOE lacks the authority to support that kind of transaction, the committee recommends that it request a waiver from current policy. That could be consistent with Congress's recent conference-report language directing DOE to accelerate the deployment of all four Genomics: GTL facilities.

The committee notes that there is much vacant space in buildings in localities that were once targeted for major development by the biotechnology industry. Reuse of existing space, rather than new construction, would have the advantage of speeding the establishment of new facilities. It could also help to improve access to the facilities for academic and industry scientists by locating the facilities on private land off DOE reservations. The Joint Genome Institute, which is on non-DOE land in Walnut Creek, California, is an excellent model for a facility with an appropriate degree of openness to encourage the scientific community to regard it as a user facility. The committee feels that the greatest mistake would be to create new user facilities behind the fences of some of the more remote existing DOE laboratories (for example, Hanford and Los Alamos) that lack proximity to major centers of biotechnology research. The Genomics: GTL program will not achieve DOE's mission goals unless it is embedded in a culture of strong basic biology and innovative biotechnology. To ensure the program's success, DOE should consider locating the facilities close to universities or federal or

private research institutions that have established centers of excellence in biosciences and biotechnology.

Recommendation 3: DOE should consider locating user facilities on private land off DOE reservations to allow an open-access policy and close to research institutions that have established programs or centers of excellence in biosciences and biotechnology. The locations for the user facilities should be selected in an open and all-inclusive competitive process that provides adequate opportunities for universities and industry to partner with DOE and its national laboratories.

The committee recognizes that the proposed integrated Genomics: GTL facilities would have a logical linkage to various large-scale enterprises, such as the nanoscience and high-performance computing programs supported by DOE, NSF, the National Institutes of Health (NIH), and the National Institute of Standards and Technology. Those programs are developing novel technologies that can be of immediate value to the proposed integrated Genomics: GTL facility. For example, breakthroughs in nanofabrication techniques for engineering single-molecule confinement devices coupled with optical interrogation systems are making it possible to explore life processes on a new scale of physical and biological reality (Box 3-2).

The committee recommends that the major effort of an integrated Genomics: GTL facility should be to pioneer new technologies. For example, purifying a soluble protein from a bacterium is now relatively straightforward. Therefore, a facility should focus on developing experimental protocols for understanding and manipulating proteins that are resistant to purification. Membrane proteins will be an important target of investigation because over 30 percent of the human genome is thought to code for them, but membrane proteins have not been well characterized; they make up only about 0.3 percent of the solved high-resolution structures. Such proteins play critical roles in various cellular processes, including signal transduction, ion and metabolite transport, and maintenance of chemical and electrical balance inside the cell. Membrane proteins have been of immense interest to biotechnology companies because of their potential use as drug targets. Their biophysical characterization is urgently needed.

A facility will enable scientists and engineers to address important and intricate metabolic subsystems that are used in such activities as converting sunlight into cellular energy. Manipulation of such an intricate metabolic subsystem inevitably involves understanding of a subset of protein components and coupled reactions that define a major subsystem. The identification and purification of those complexes, which may be transient and labile, would require a major research effort to extract them as functional units for various biophysical-biochemical characterizations. A full understanding of a complex will allow scien-

BOX 3-2
Nanobiotechnology

Nanobiotechnology is an exciting, challenging, and rapidly evolving field of scientific and technological exploration. Its evolution is being catalyzed by breakthroughs in nanofabrication techniques, material science, and molecular biology and genetics and by the development of advanced optical systems. In addition, synergistic interactions between the various physical and biological sciences are yielding engineered scientific devices and systems with well-defined structure and function at the nanometer level. Such synergy is opening up intriguing opportunities to explore and manipulate biological systems on the community, cellular, subcellular, and molecular scales. Participants in the National Nanotechnology Initiative Workshop held in Washington, DC, in 2003 identified four general categories of research and development opportunities that can catalyze new discovery at the interface of physical and biological systems (Vogel and Baird, 2005):

- Advanced imaging technologies.
- In vivo analysis of cellular processes.
- Understanding how cells work through bottom-up assembly of biological nanosystems ex vivo.
- Nanotechnology and human health.

Already-advanced imaging technologies are being coupled with single-molecule confinement devices for optical interrogation enzyme systems. Such experimental systems are being used to accelerate the development of a new generation of molecular ecology tools for DNA fingerprinting and quantitative hybridization probing (Ugaz et al., 2003). The tools are essential for high-throughput analysis of complex microbial communities and for prospecting for novel industrial microorganisms and enzymes. Bottom-up assembly using self-organizing molecules into higher-order assemblies is being explored by the National Renewable Energy Laboratory, which is using cellulosomal proteins as a scaffolding material and quantum dots to observe the assembly of the proteins on cellulose (Din et al., 2005). In vivo analysis with nanoprobes and optical detection is being used to study the state and activities of proteins and other biomolecules in whole cells (Zang et al., 2003).

tists to implement a rational engineering design to make the subsystem more predictable, controllable, and efficient. To introduce the engineered subsystem with a higher efficiency into a cell system, it is equally important to understand how a particular subsystem interacts with other subsystems inside the cell. A whole-cell approach will be necessary to understand the networking among different subsystems in a living cell. That argues strongly for the importance of simultaneously studying molecular components in the cellular context.

All the projects in the proposed integrated facilities would involve multiple investigators, so a user-friendly data-management system across domains of in-

vestigation will be essential to keep track of all the new information and will not only guide next experiments but allow the tracking of progress. It is a challenging project in itself to generate an electronic notebook and database for the heterogeneous data and to cope with the volumes of data that will be generated. We encourage computational engineers and investigating scientists to codevelop the laboratory data-management system to meet specifications of the research. Constant and iterative discussions and software testing between the computational engineers and the scientists will be necessary to produce a practical product.

Overlap with Other Federal Agencies

Because there is no such vertically integrated facility dedicated to studying biofuel, bioremediation, and carbon management, there is no direct competition with other agencies or anyone else. On the contrary, it is conceivable that a successful applicant will be able to partner with other investigators to leverage some of the existing but smaller-scale operations of various modules of technologies in protein and machine purification, mass spectroscopy, crystallography, electron cryomicroscopy, light microscopy, laboratory information management, and computer simulation and modeling.

The recent NIH roadmap initiatives for medical research have a number of programs—such as structural biology, computational biology, bioinformatics, molecular imaging, nanomedicine, building blocks, and biological pathways and networks—that would be complementary to the Genomics: GTL Program. Although NIH's missions are peculiar to human health, many of the methods developed with NIH support are generic and adaptable for studying other organisms of interest. NSF has also been supporting nanoscience and technology research centers and programs, which have technology-development components relevant to the Genomics: GTL program. Neither NIH nor NSF is supporting a large-scale and integrative approach as Genomics: GTL is planning to do. A synergistic coordination among those funding agencies would have the potential to push the technologies more efficiently in advancing our fundamental understanding of life and in improving our quality of life and environment. DOE has had collaborative programs with other agencies in the past (for example, the Human Genome Program with NIH, and the maize-sequencing program with NSF and USDA). For Genomics: GTL, DOE should be strongly encouraged to coordinate with and leverage the programs of the other federal agencies with common interests in microbial biology (NSF), bioremediation (Environmental Protection Agency), biofuels (U.S. Department of Agriculture), and genomics (NIH) (DOE, 2005d). The complex and often labile nature of the biological systems to be studied will present new challenges for any of the technologies and will justify investment in further development and refinement.

At the same time, it will be important for Genomics: GTL to coordinate effectively with related programs in DOE. For example, as Genomics: GTL

acquires additional emphasis on relevant aspects of plant biology, recommended by this committee throughout this report, it will be critical to achieve close coordination with plant-related research funded through DOE's Energy Biosciences Program. The DOE Energy Biosciences Program has a long history of supporting basic research in plant and microbial biology. Although the program is small, the quality of the research programs supported by it has been very high. Because both the Energy Biosciences and the Genomics: GTL programs are administered by the Office of Science, it should be easy for these two programs to cooperate. In fact, as the Genomics: GTL program begins to add plant biology to its research portfolio, it makes sense to use the close connections with academic plant scientists already established by the Energy Biosciences Program, as opposed to reinventing such expertise in Genomics: GTL.

In addition to collaborations within the United States, DOE should also consider international collaborations with other countries that have similar genomics programs—for example, Genome Canada and the Netherlands Genomics Initiative. Moreover, bioenergy and carbon management are subjects of global interest, so international collaborations would reduce duplication of effort and leverage each country's resources and expertise.

Technology Dissemination and Educational Outreach

Because of the rapid development of various technologies in the proposed Genomics: GTL facilities, part of the role of the Genomics: GTL program is to sponsor regular workshops and symposia to disseminate the new experimental and computational methods to the broad community. Such activities have been well conducted in the DOE laboratories in various disciplines—for example, contractor-grantee workshops of the Human Genome Project and Genomics: GTL program and Sandia National Laboratory's Workshop on Computational Molecular Biology.

The proposed Genomics: GTL facilities could provide unique and unusually rich technology environments for junior scientists to be introduced into research, to get excited about research related to our environment and energy production, to learn specialized techniques, and to develop research careers in microbial biology. Therefore, training components that reach out to high-school students, undergraduates, graduate students, and postdoctoral fellows should be included in the Genomics: GTL program. The training program will focus on recruiting the most talented young people from various educational and ethnic backgrounds.

Such training programs would benefit not only the trainees but also the research staffs of the Genomics: GTL facilities. Because the proposed facilities are not didactic teaching institutions, the presence of trainees will provide an intellectual environment in which senior researchers will be confronted by students who question and challenge the basic premises of a scientific approach. Often research ideas can be generated from simple questions born of curiosity.

Such training components will add important intellectual dimensions to the proposed facilities.

DOE is already sending brochures about its programs to high schools (for example, *Your World: Biotechnology and You*), but a selective and well-constructed summer training program for high-school students and teachers will not only reach out to the students but also enhance the quality of science education. Immersion, in which teachers join scientists to conduct experiments, has been identified as a key strategy in the professional learning of teachers (Loucks-Horsley et al., 1998). DOE already has the Laboratory Science Teacher Professional Development program (DOE, 2006), which provides an immersion experience for highschool teachers. Genomics: GTL facilities could be included in the settings for the program. Training for students and teachers would allow the Genomics: GTL program to identify talented young people and ensure the nurturing of their talent. Student participants could become leading scientists, and their training would help to ensure the quality and leadership of bioresearch in DOE and other bioscience enterprises in both academe and industry.

The undergraduate research experience is valuable for students who are exploring career options. Many graduate students who enroll in highly competitive graduate programs in the United States have already had research experience in their undergraduate years through specialized summer research programs or through faculty-supervised research in universities. Both DOE and NSF have sponsored competitive and successful summer research programs—for example, Science Undergraduate Laboratory Internships at DOE and Research Experiences for Undergraduates at NSF. The Genomics: GTL program should create a prestigious undergraduate summer research program that would introduce students to the importance and excitement of microbial systems biology as applied to energy or environmental problems. Such experience may attract this pool of talented undergraduates to select a research career path consistent with the vision of the Genomics: GTL program.

DOE's national laboratories have a long tradition of providing laboratory space and research resources for graduate students to conduct their PhD thesis research in physics and chemistry aimed at degrees from academic institutions hundreds or even thousands of miles away. Similar arrangements could be adopted in the Genomics: GTL program. The senior staff of the Genomics: GTL facilities would play the role of comentors of the graduate students enrolled in participating universities. NIH has recently implemented the Graduate Partnerships Program (GPP), which links NIH with graduate programs of various universities around the country (NIH, 2005); students can work in one of the intramural laboratories at NIH to conduct part of their thesis work for a period of weeks to months. NSF's Integrative Graduate Education and Research Traineeship (IGERT) program also seeks to train PhD candidates to become scientists and engineers who have the interdisciplinary background and the technical, professional, and personal skills needed to address the global questions of

the future. GPP and IGERT are models that DOE could consider adapting to its own purposes. Programs similar to those will allow the Genomics: GTL program researchers to gain access to a broad pool of talented graduate students. A successful graduate training program is an effective way of disseminating the Genomics: GTL resources and technologies to the broader biological community.

A postdoctoral fellow is a scientist at the stage of developing his or her own independent research. Postdoctoral work takes place during a critical period of typically 3-5 years during which one must make the transition from graduate student to independent investigator. The productivity of postdoctoral fellows is generally high. Because systems biology is a developing field, many universities and industries will probably be seeking qualified scientists with relevant expertise to fill their new faculty and staff positions. The Genomics: GTL program will have the opportunity to become a primary source of such people. Supplying talented scientists to the marketplace of ideas is one of the best ways to establish the credibility and reputation of the Genomics: GTL program.

The committee suggests that DOE consider innovative arrangements to encourage the use of the Genomics: GTL program and its facilities as training grounds for the next generation of scientists. This is an especially important aspect of a program that promises to deliver a high degree of interdisciplinary cooperation.

In addition to training students and postdoctoral follows, the Genomics: GTL facilities should provide long-term and short-term sabbatical fellowships for faculty from universities and industries. The constant influx of visiting scientists will provide expertise and perspectives complementary to those of the Genomics: GTL staff. The intellectual contribution of well-established investigators not only will benefit the science done today but also will generate the new approaches and ideas of tomorrow in the Genomics: GTL program.

Recommendation 4: DOE should consider partnering with universities and other federal agencies to develop programs that use Genomics: GTL institute-like facilities as training grounds for the next generation of scientists.

SUMMARY

The grand challenge articulated in the Genomics: GTL program mission is to understand the "molecular machines of life" that underlie key processes in *bioremediation*, *carbon sequestration*, and *biofuel production*. The more immediate goal of the program is to develop predictive models of system function. If robust, such models would enable efficient re-engineering or optimizing of molecular machines to solve the nation's energy and environmental challenges. The committee wholeheartedly supports the goal of the program and concludes that the best way to achieve them is through a systems biology approach. We further

endorse, with great enthusiasm, the ambition of the Genomics: GTL program to place DOE at the forefront of systems biology research as it was in genomics.

The committee finds, however, that the proposed configuration of the user facilities is not optimal for achieving the goals. Worse, it risks leaving DOE as a follower, rather than a leader, in systems biology. We offer, as a constructive suggestion, an alternative model that we feel would better serve the Genomics: GTL mission. The alternative model would consist of several problem-oriented facilities that are vertically integrated and institute-like (Box 3-3). Each facility would work on research problems or themes chosen to propel the field of systems biology as a whole, and its applications to the grand challenge in particular, rapidly forward. To the greatest extent possible, the facilities would be built in parallel and brought on line rapidly. Although technology development would be an important part of all the facilities, it should be driven by the scientific questions being addressed, not by the need to produce reagents in bulk to serve a wider community. Our model places the facilities close to and in collaboration

BOX 3-3
Vision for an Institute-like Genomics: GTL Facility for Bioenergy

An institute-like Genomics: GTL facility would

- Facilitate use-inspired fundamental research, motivated by a "grand challenge," such as the replacement of oil with affordable, carbon-neutral biofuels.
- Have a portfolio of large-scale interdisciplinary science projects and some smaller efforts.
- Support simultaneous projects that explore competing approaches to the goal.
- Develop technology driven by biological problems.
- Serve as a "summer institute" for graduate students and postdoctoral fellows from around the country to expose them to the scientific, technological, and societal implications of systems and synthetic biology.
- Have a list of potential projects suitable for high-school and undergraduate research projects.
- Serve as a shared facility available to broader communities.
- Have programs for visiting scholars and industrial fellows modeled after the ones at the Institute for Theoretical Physics.
- Host community building activities—for example, annotation jamborees.
- Include an "e-science" or "cyberinfrastructure" component.
- Award an annual prize or host an annual competition to encourage innovative ideas—for example, a competition in predicting organism behavior from existing systems biology data.
- Have Ethical, Legal, and Social Implications Research Program activities as an integrated component of the research.

with other centers of front-line biological research so that they can draw on and exchange expertise with colleagues in the wider community.

DOE has done an exceptional job in leading the country in many fields of research that were at the cutting edge, in particular the Human Genome Project. DOE now has an opportunity to become a world leader in systems biology through the Genomics: GTL program and by integrating or connecting fundamental research data to other programs in DOE and other national and international agency programs. The committee commends DOE for its development of the Genomics: GTL program and encourages the DOE administration to consider the committee's proposed alternative plan for the Genomics: GTL facilities. The recommendations in this report, if implemented, will enhance DOE's potential for success in its three critical mission areas: bioenergy, bioremediation, and carbon sequestration.

Glossary

Annotation[1]	Identification of the locations and coding regions of genes in a genome and determination of what they do.
Bioinformatics[1]	The study of genetic and other biological information using computer, mathematical, and statistical techniques.
Codon	A set of three adjoined nucleotides (triplet) that codes for an amino acid or a termination signal.
Functional genomics[1]	The study of genomes to determine the biological function of all the genes and their products.
Gene expression[1]	Conversion of the information encoded in a gene first to messenger RNA and then to a protein.
Genome[1]	The entire chromosomal genetic material of an organism.

Definitions taken from [1]Science Vol. 291; [2]NRC, 2003b; [3]Handelsman, 2005b; [4]NML, 2005; [5]http://www.biochem.northwestern.edu/holmgren/Glossary/; [6]http://www.medicine.net.com.

Genomics[1]	The comprehensive study of whole sets of genes and their interactions rather than single genes or proteins.
High throughput[2]	Rapid (and simultaneous) processing of large sample sets.
Metabolomics[2]	The global analysis of metabolites and metabolic networks in cells, tissues, and organ systems.
Metagenomics[3]	The genomic analyses of assemblages of uncultured microorganisms.
Microarray[2]	A microscope slide or other solid support on which many distinct cDNAs or DNA oligonucleotides are patterned at high density in an addressable array. Microarrays are interrogated by hybridization to fluorescently labeled cDNAs or RNAs to detect the genes that are actively transcribed.
Open reading frame[4]	A reading frame in a sequence of nucleotides in DNA that contains no termination codons and so can potentially translate as a polypeptide chain.
PCR[1]	Polymerase chain reaction: a technique for amplifying a piece of DNA quickly and cheaply.
Proteomics[1]	The study of full set of proteins encoded by a genome.
Sequencing[5]	Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein.
Shotgun assembly[1]	Breaking DNA into many small pieces, sequencing the pieces, and assembling the fragments.
Structural genomics[1]	The effort to determine the 3D structures of large numbers of proteins using both experimental techniques and computer simulation.
Upregulation[6]	An increase in the number of receptors on the surface of target cells, making the cells more sensitive to a hormone or another agent.

References

Armbrust, E.V., J.A. Berges, C. Bowler, B.R. Green, D. Martinez, N.H. Putnam, S. Zhou, A.E. Allen, K.E. Apt, M. Bechner, M.A. Brzezinski, B.K. Chaal, A. Chiovitti, A.K. Davis, M.S. Demarest, J.C. Detter, T. Glavina, D. Goodstein, M.Z. Hadi, U. Hellsten, M. Hildebrand, B.D. Jenkins, J. Jurka, V.V. Kapitonov, N. Kröger, W.W.Y. Lau, T.W. Lane, F.W. Larimer, J.C. Lippmeier, S. Lucas, M. Medina, A. Montsant, M. Obornik, M.S. Parker, B. Palenik, G.J. Pazour, P.M. Richardson, T.A. Rynearson, M.A. Saito, D.C. Schwartz, K. Thamatrakoln, K. Valentin, A. Vardi, F.P. Wilkerson, and D.S. Rokhsar. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306:79-86.

ASM (American Society for Microbiology). 2001. Global Environmental Change. Microbial Contributions. Microbal Solutions. American Academy of Microbiology, Washington, DC.

ASM. 2004. The Global Genome Question: Microbes as the Key to Understanding Evolution and Ecology. American Academy of Microbiology, Washington, DC.

BERAC (Biological and Environmental Research Advisory Committee). 2000. Bringing the genome to life: Energy related biology in the new genomic world. Accessed online 11/21/05 at http://www.sc.doe.gov/ober/berac/genome-to-life-rpt.html.

Biomass R&D Technical Advisory Committee. 2002. Roadmap for biomass technologies in the United States. Accessed online 11/21/05 at http://72.14.203.104/search?q=cache:0GE_a5SC4gYJ:www.bioproducts-bioenergy.gov/pdfs/FinalBiomassRoadmap.pdf+roadmap+for+biomass+technologies&hl=en&gl=us&ct=clnk&cd=1.

Bonifati, V., P. Rizzu, M.J. van Baren, O. Schaap, G.J. Breedveld, E. Krieger, M.C. Dekker, F. Squitieri, P. Ibanez, M. Joosse, J.W. van Dongen, N. Vanacore, J.C. van Swieten, A. Brice, G. Meco, C.M. van Duijn, B.A. Oostra, and P. Heutink. 2003. Mutations in the DJ-1 gene associated with autosomal recessive early-onset Parkinsonism. *Science* 299(5604):256-259.

Business Communication Company, Inc. 2004. Global market for industrial enzymes to reach $2.4 million by 2009. Accessed online 11/17/05 at http://www.bccresearch.com/editors/RC-147U.html.

COV (Committee of Visitors for the Review of the Environmental Remediation Sciences Division). 2004. Report to the Biological and Environmental Advisory Committee (BERAC). Accessed online 12/13/05 at http://www.sc.doe.gov/ober/berac/ERSDCOV.pdf.

Dai Z., B.S. Hooker, D.B. Anderson, and S.R. Thomas. 2000. Improved plant-based production of E1 endoglucanase using potato: expression optimization and tissue targeting. *Molecular breeding* 6:277-285.

Daniels, R. 2005. The metagenomics of soil. *Nature Reviews Microbiology* 3(6):470-478.

DeLong, E.F. and D.M. Karl. 2005. Genomic perspectives in microbial oceanography. *Nature* 437:336-342.

Din, S.-Y., S. Smith, Q. Xu, J. Sugiyama, M. Jones, G. Rumbles, E.A. Bayer, and M.E. Himmel. 2005. Ordered arrays of quantum dots using cellulosomal proteins. *Industrial Biotechnology* 1:198-206.

DOE (U.S. Department of Energy). 2000. On the Shoulders of Giants: Private Sector Leverages HGP Successes. *Human Genome News* Vol. 11, No. 1-2. The U.S. Department of Energy, Germantown.

DOE. 2002. Innovative approaches for cleaning and treating hazardous wastes at DOE sites. Accessed online 12/13/05 at http://doegenomestolife.org/cleanup.pdf.

DOE. 2003. DOE: Genomics: GTL Contractor-grantee workshop I. The U.S. Department of Energy, Germantown.

DOE. 2004. DOE: Genomics: GTL Contractor-grantee workshop II. The U.S. Department of Energy, Germantown.

DOE. 2005a. Mission. Accessed online 10/10/05 at http://www.doe.gov/engine/content.do?BT_CODE=ABOUTDOE.

DOE. 2005b. DOE: Genomics: GTL roadmap. The U.S. Department of Energy, Germantown.

DOE. 2005c. DOE: Genomics: GTL Contractor-grantee workshop III. The U.S. Department of Energy, Germantown.

DOE. 2005d. Programs with microbial research component. Accessed online 9/12/05 at http://www.microbialgenome.org/research.shtml.

DOE. 2006. About laboratory science teacher professional development. Accessed online 1/18/06 at http://www.scied.science.doe.gov/scied/LSTPD/about.htm.

DOE and NRC. 1998. Serving Science and Society in the New Millennium. National Academies Press, Washington, DC.

DOE-EERE (U.S. Department of Energy-Energy Efficiency and Renewable Energy). 2005a. Biomass program. Accessed online 11/23/05 at www.eere.energy.gov/biomass.

DOE-EERE. 2005b. Hydrogen fuel cells, and infrastructure technologies program. Accessed online 11/23/05 at www.eere.energy.gov/hydrogenandfuelcells/.

DOE-ORNL (Oak Ridge National Laboratory). 2002. The DOE consortium on enhancing carbon sequestration in terrestrial ecosystems. Accessed online 12/1/05 at http://csite.esd.ornl.gov/.

EEA (European Environment Agency). 2004. Impacts of Europe's changing climate. European Environment Agency, Copenhagen.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.-F. Tomb, B.A. Dougherty, J.M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J.D. Gocyne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J.M. Kelley, J.F. Weidman, C.A. Phillips, T. Spriggs, E. Hedblom, M.D. Cotton, T.R. Utterback, M.C. Hanna, D.T. Nguyen, D.M. Saudek, R.C. Brandon, L.D. Fine, J.L. Fritchman, J.L. Fuhrmann, N.S.M. Geoghagen, C.L. Gnehm, L.A. McDonald, K.V. Small, C.M. Fraser, H.O. Smith, and J.C. Venter. 1995. Whole-genome random sequencing and assembly of *Haemophilus Influenzae* Rd. *Science* 269(5223):496-498.

Fraser, A.G. and E.M. Marcotte. 2004. A probabilistic view of gene function. *Nature Genetics* 36(6):559-564.

Galas, D. 2001. Sequence interpretation: making sense of the sequence. *Science* 291(5507):1257-1260.

Galbraith, E.A., D.A. Antonopoulos, and B.A. White. 2004. Suppressive subtractive hybridization as a tool for identifying genetic diversity in an environmental metagenome: the rumen as a model. *Environmental Microbiology* 6(9):928-937.

Gans, J., M. Wolinsky, and J. Dunbar. 2005. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* 309:1387-1390.

Gates Foundation. 2005. New prospects in the fight against malaria. Accessed online 11/16/05 at http://www.gatesfoundation.org/GlobalHealth/Pri_Diseases/Malaria/Related+Info/MalariaControl.htm.

Gerlt, J.A. and P.C. Babbitt. 2000. Can sequence determine function? *Genome Biology* 1(5):reviews0005.1-0005.10.

Gibson, G. and S.V. Muse. 2002. A Primer of Genome Science. Sinauer Associates, Sunderland, Massachusetts.

Goffeau, A., B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S.G. Oliver. 1996. Life with 6000 genes. *Science* 274(5287):546-567.

Hallenbbeck, P.C. 2005. Fundamentals of the fermentative production of hydrogen. *Water Science & Technology* 52:21-29.

Handelsman, J. 2005a. Metagenomics or megagenomics. *Nature Reviews Microbiology* 3(6): 457-458.

Handelsman, J. 2005b. How to find new antibiotics. *The Scientist* 19:20.

Heidelberg, J.F., I.T. Paulsen, K.E. Nelson, E.J. Gaidos, W.C. Nelson, T.D. Read, J.A. Eisen, R. Seshadri, N. Ward, B. Methe, R.A. Clayton, T. Meyer, A. Tsapin, J. Scott, M. Beanan, L. Brinkac, S. Daugherty, R.T. DeBoy, R.J. Dodson, A.S. Durkin, D.H. Haft, J.F. Kolonay, R. Madupu, J.D. Peterson, L.A. Umayam, O. White, A.M. Wolf, J. Vamathevan, J. Weidman, M. Impraim, K. Lee, K. Berry, C. Lee, J. Mueller, H. Khouri, J. Gill, T.R. Utterback, L.A. McDonald, T.V. Feldblyum, H.O. Smith, J.C. Venter, K.H. Nealson, and C.M. Fraser. 2002. Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nature Biotechnology* 20:1118-1123.

Heidelberg, J.F., R. Seshadri, S.A. Haveman, C.L. Hemme, I.T. Paulsen, J.F. Kolonay, J.A. Eisen, N. Ward, B. Methé, L.M. Brinkac, S.C. Daugherty, R.T. Deboy, R.J. Dodson, A.S. Durkin, R. Madupu, W.C. Nelson, S.A. Sullivan, D. Fouts, D.H. Haft, J. Selengut, J.D. Peterson, T.M. Davidsen, N. Zafar, L. Zhou, D. Radune, G. Dimitrov, M. Hance, K. Tran, H. Khouri, J. Gill, T.R. Utterback, T.V. Feldblyum, J.D. Wall, G. Voordouw, and C.M. Fraser. 2004. The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris Hildenborough*. *Nature Biotechnology* 22(5):554-559.

Hughes, J.B., K. Dustin, and C.H. Ward. 2002. Bioremediation. Published electronically by the Ground-Water Remediation Analysis Center, September, 2002.

Ideker, T., T. Galistski, and L. Hood. 2001. A new approach to decoding life: Systems biology. *Annual Review in Genomics and Human Genetics* 2:343-372.

IPCC (Intergovernmental Panel on Climate Change). 1997. Stabilization of Atmospheric Greenhouse Gases: Physical, Biological and Socioeconomic Implications. Cambridge University Press, Cambridge, UK.

IPCC. 2001. Climate change 2001: The Scientific Basis. Cambridge University Press, Cambridge, UK.

Lan, N., G.T. Montelione, and M. Gerstein. 2003. Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. *Current Opinion in Chemical Biology* 7(1):44-54.

Loucks-Horsley, S., N. Love, K.E. Stiles, S. Mundy, and P.W. Hewson. 1998. Designing Professional Development for Teachers of Science and Mathematics. Corwin Press, Inc., Thousand Oaks, California.

Methé, B.A., K.E. Nelson, J.A. Eisen, I.T. Paulsen, W. Nelson, J.F. Heidelberg, D. Wu, M. Wu, N. Ward, M.J. Beanan, R.J. Dodson, R. Madupu, L.M. Brinkac, S.C. Daugherty, R.T. DeBoy, A.S. Durkin, M. Gwinn, J.F. Kolonay, S.A. Sullivan, D.H. Haft, J. Selengut, T.M. Davidsen, N. Zafar, O. White, B. Tran, C. Romero, H.A Forberger, J. Weidman, H. Khouri, T.V. Feldblyum, T.R. Utterback, S.E. Van Aken, D.R. Lovley, and C.M. Fraser. 2003. Genome of *Geobacter sulfurreducens*: Metal reduction in subsurface environments. *Science* 302:1967-1969.

MIT Synthetic Biology Working Group. 2005. Synthetic biology frequently asked questions. Accessed online 11/24/05 at http://openwetware.org/wiki/Synthetic_Biology:FAQ.

Mount, D.W. and R. Pandey. 2005. Using bioinformatics and genome analysis for new therapeutic interventions. *Molecular Cancer Therapeutics* 4:1636-1643.

Nath, K. and D. Das. 2004. Improvement of fermentative hydrogen production: various approaches. *Applied Microbiology and Biotechnology* 65:520-529.

Neuhauser, C., D.A. Andow, G. Heimpel, G. May, R. Shaw, and S. Wagenius. 2003. Community genetics—a synthesis of community ecology and population genetics. *Ecology* 84:545-558.

NIH (National Institutes of Health). 2005. Graduate Partnerships Program. Linking NIH Laboratories to Universities in the Training of Graduate Students. Accessed online 12/7/05 at http://gpp.nih.gov.

Ning, L., G.T. Montelione, and M. Gerstein. 2003. Ontologies for proteomics Towards a systematic definition of structure and function that scales to the genome level. *Current Opinion in Chemical Biology* 7:44-54.

NML (U.S. National Medical Library). 2006. Genetics Home Reference: Your Guide to Understanding Genetic Conditions. Accessed online 2/1/06 at http://ghr.nlm.nih.gov/ghr/glossary/openreadingframe.

NRC (National Research Council). 1992. Policy Implications of Greenhouse Warming: Mitigation, Adaptation, and the Science Base. National Academy Press, Washington, DC.

NRC. 1995. Allocating Federal Funds for Science and Technology. The National Academies Press, Washington, DC.

NRC. 2000. Biobased Industrial Products: Priorities for Research and Commercialization. National Academy Press, Washington, DC.

NRC. 2003a. The Carbon Dioxide Dilemma: Promising Technologies and Policies. The National Academies Press, Washington, DC.

NRC. 2003b. Frontiers in Polar Biology in the Genomic Era. The National Academies Press, Washington, DC.

NRC. 2004. Biological Confinement of Genetically Engineered Organisms. The National Academies Press, Washington, DC.

NRC. 2005. Mathematics and 21st Century Biology. The National Academies Press, Washington, DC.

NRDC (Natural Resources Defense Council). 2004. Growing Energy. How Biofuels Could Help End America's Dependence on Oil. The Natural Resources Defense Council, Washington, DC.

NSF (National Science Foundation). 2005. Microbial Genome Sequencing Program FY 2006. Program Solicitation. Accessed online 12/12/05 at http://www.nsf.gov/pubs/2006/nsf06513/nsf06513.htm.

Nuismer, S.L. and M. Kirpatrick. 2003. Gene flow and the coevolution of parasite range. *Evolution* 57:746-754.

Nuutila M.A., A. Ritala, R.W. Skadsen, L. Mannonen, and V. Kauppinen. 1999. Expression of fungal thermotolerant endo-1, 4-β-glucanase in transgenic barley seeds during germination. *Plant Molecular Biology* 41:777-783.

Perlack, R.D., L. Wright, A.F. Turhollow, R.L. Graham, B.J. Stokes, and D.C. Erbac. 2005. Biomass as feedstock for a bioenergy and bioproducts industry: the technical feasibility of a billion-ton annual supply. DOE/GO-102005-2135.

Pizza, M., V. Scarlato, V. Masignani, M.M. Giuliani, B. Arico, M. Comanducci, G.T. Jennings, L. Baldi, E. Bartolini, B. Capecchi, C.L. Galeotti, E. Luzzi, R. Manetti, E. Marchetti, M. Mora, S. Nuti, G. Ratti, L. Santini, S. Savino, M. Scarselli, E. Storni, P. Zuo, M. Broeker, E. Hundt, B. Knapp, E. Blair, T. Mason, H. Tettelin, D.W. Hood, A.C. Jeffries, N.J. Saunders, D.M. Granoff, J.C. Venter, E.R. Moxon, G. Grandi, and R. Rappuoli. 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287:1767-1768.

Roberts, L. 2001. The human genome. Controversial from the start. *Science* 291:1182-1188.

Rosenberg, N.J., F.B. Metting, and R.C. Izaurralde (Editors). 2004. Applications of Biotechnology to Mitigation of Greenhouse Warming. Proceedings of the St. Michaels II Workshop, April 2003. Battelle Press, Columbus, Ohio.

Sanford, K. 2004. Bioprocessing: Pathway engineering through rational design. Daily Biotech Updates. Volume 24, January 15, 2004. Accessed online 11/17/05 at http://www.genencor.com/pdf/pathway_011504.pdf.

Securing America's Future Energy (SAFE) and the National Commission on Energy Policy (NCEP). 2005. Oil shockwave: Oil crisis executive simulation. Accessed on December 13, 2005 at http://www.secureenergy.org/reports/oil_shock_report_master.pdf.

Tringe, S.G. and E.M. Rubin. 2005. Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics* 6(11):805-814.

Tringe, S.G., C. von Mering, A. Kobayashi, A.A. Salamov, K. Chen, H.W. Chang, M. Podar, J.M. Short, E.J. Mathur, J.C. Detter, P. Bork, P. Hugenholtz, and E.M. Rubin. 2005. Comparative metagenomics of microbial communities. *Science* 308(5721):554-557.

Trotter, E.W., C.M. Kao, L. Berenfeld, D. Botstein, G.A. Petsko, and J.V. Gray. 2002. Misfolded proteins are competent to mediate a subset of the responses to heat shock in *Saccharomyces cerevisiae*. *The Journal of Biological Chemistry* 277(47):44817-44825.

Turelli, M. and N.H. Barton. 1990. Dynamics of polygenic characters under selection. *Theoretical Population Biology* 38:1-57.

Ugaz, V.M., R. Lin, N. Srivastava, D.T. Burke, and N.A. Burns. 2003. A versatile microfabricated platform for electrophoresis of double- and single-stranded DNA. *Electrophoresis* 24:151-157.

Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, D.E. Fouts, S. Levy, A.H. Knap, M.W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.H. Rogers, and H.O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667):66-74.

Vogel, V. and B. Baird. 2005. Nanobiotechnology: Report of the National Nanotechnology Initiative Workshop, October 9-11, 2003, Arlington, VA. National Science and Technology Council.

Williams, R.J. and N.D. Martinez. 2004. Stabilization of chaotic and non-permanent food-web dynamics. *The European Physical Journal B* 38:297-303.

Williamson, L.L., B.R. Borlee, P.D. Schloss, C. Guan, H.K. Allen, and J. Handelsman. 2005. Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. *Applied and Environmental Microbiology* 71:6335-6344.

Wilson, M.A., D. Ringe, and G.A. Petsko. 2005. The atomic resolution crystal structure of the YajL (ThiJ) protein from *Escherichia coli*: a close prokaryotic homologue of the Parkinsonism-associated protein DJ-1. *Journal of Molecular Biology* 353:678-691.

Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2003. High-Cot sequence analysis of the maize genome. *The Plant Journal: for cell and molecular biology* 34(2):249-255.

Zang, J., R.E. Campbell, A.Y. Ting, and R.Y. Tsien. 2003. Creating new fluorescent probes for cell biology. *Nature Reviews Molecular Cell Biology* 3:906-918.

Ziegelhoffer T., J.A. Raasch, and S. Austin-Phillips. 2001. Dramatic effects of truncation and subcellular targeting on the accumulation of recombinant microbial cellulase in tobacco. *Molecular Breeding* 8:147-158.

Ziegler, M., S. Thomas, and K. Danna. 2000. Accumulation of a thermostable endo-1,4 β-D-glucanase in the apoplast of *Arabidopsis thaliana* leaves. *Molecular Breeding* 6:37-46.

Appendixes

Appendix A

Committee Biographies

Jennie Hunter-Cevera (Chair), is the president of the University of Maryland Biotechnology Institute and has served in that position since October 1999. Before then, she was the director of the Center for Environmental Biotechnology at the Lawrence Berkeley National Laboratory (LBL). She was cofounder of two small companies (the Biotic Network and Blue Sky Research) that did contract work for large pharmaceutical and biotechnology companies and consulted for 5 years in a variety of biotechnology fields before working at LBL. She was employed at Cetus Corporation for 10 years, where she served as the director of fermentation, research and development, and before that, at E.R. Squibb and Sons as a research scientist. Dr. Hunter-Cevera holds a PhD in microbiology from Rutgers University and an MS in microbiology, and a BA in biology from West Virginia University. She has given more than 50 invited lectures and seven keynote lectures and is the author of several papers, chapters, and books, as well as the holder of two patents and one pending patent. She was elected to the American Academy of Microbiology in 1995, received the 1996 SIM Charles Porter Award, was elected as a SIM Fellow in 1997, and was the West Virginia University Nath Lecturer in 1999. She was honored by West Virginia University as the 2003 recipient of its Distinguished Alumni Award. She has been honored as one of Maryland's Top 100 Women for 2003. Dr. Hunter-Cevera serves on Governor Ehrlich's Technology Commission for the State of Maryland and was one of six members of the Governor's Executive Council for Transition. In May 2004, she was the recipient of the USFCC/J. Roger Porter Award. Supported by the U.S. Federation for Culture Collections and the American Society for Microbiology, the award recognized Dr. Hunter-Cevera's expertise in collecting, main-

taining, and preserving microbial cultures. Dr. Hunter-Cevera serves as president of the International Marine Biotechnology Association.

Charles R. Cantor is a founder, chief scientific officer, and member of the Board of Directors at Sequenom, Inc. He is also founder of SelectX Pharmaceuticals, a drug-discovery company based in the Boston area. He is codirector of the Center for Advanced Biotechnology at Boston University, where he is a professor of biomedical engineering. Dr. Cantor has held positions at Columbia University and the University of California, Berkeley, and he was also director of the Human Genome Center of the Department of Energy at Lawrence Berkeley Laboratory. He has published more than 400 peer-reviewed articles and is coauthor of a three-volume textbook on biophysical chemistry and the first textbook on genomics, *The Science and Technology of the Human Genome Project*. He has been granted more than 60 patents. He sits on the advisory boards of more than 20 national and international organizations and is a member of the National Academy of Sciences.

Wah Chiu is the Alvin Romansky Professor of Biochemistry at Baylor College of Medicine. He is a leading investigator in the structural determination of macromolecular assemblies using electron cryomicroscopy. He directs the National Center for Macromolecular Imaging (http://ncmi.bcm.tmc.edu/), which is moving into new territories with computational and visualization techniques and is supported by the National Center for Research Resources. His group has deployed this imaging technology to study three-dimensional structures of viruses, chaperonins, ion channels, and actin bundles at unprecedented resolutions. Dr. Chiu is the founding director of a cross-disciplinary and cross-institutional PhD program in structural and computational biology and molecular biophysics (http://www.bcm.tmc.edu/scbmb/). He also served as the first chair of the Gulf Coast Consortia (http://www.gulfcoastconsortia.org), which promotes the collaboration of scientists at the interface between biomedicine and physical, chemical, mathematical, computational, and engineering sciences. Both those research and training organizations have faculty drawn from six academic institutions, including Baylor College of Medicine, Rice University, University of Houston, University of Texas Health Science Center in Houston, MD Anderson Cancer Center, and University of Texas Galveston Medical Branch.

Douglas R. Cook is a professor in the department of plant pathology and the director of the CA&ES Genomics Facility at the University of California, Davis. He is also a professor in the graduate school in bioinformatics and genome research, Universität Bielefeld, Germany. Dr. Cook's research involves the application of genomics, genetics, and molecular biology to study basic and applied issues in legumes and grapes. His specific research interests include regulatory mechanisms of the rhizobium-legume symbiosis, characterizing the determinants

of disease resistance and susceptibility in grapes, genomic and molecular characterization of quality attributes in wine grapes, and translational genomics from model systems to crops in the legume family.

Eric W. Kaler is Elizabeth Inez Kelley Professor in the Department of Chemical Engineering and dean of the College of Engineering at the University of Delaware. He is known for his distinguished study and applications of complex fluids, including advances in the understanding of surfactant mixtures in synthesis of new materials. Dr. Kaler has served on several National Research Council panels, including the subpanel for the National Institute of Standards and Technology Center for Neutron Research, which he chaired, and the panel for materials science and engineering. He was named a fellow of the American Association for the Advancement of Science in 2001. He was one of the first to receive a Presidential Young Investigator Award from the National Science Foundation, in 1984. He also received the Curtis W. McGraw Research Award from the American Society of Engineering Education in 1995 and the 1998 American Chemical Society Award in Colloid or Surface Chemistry. Dr. Kaler earned his PhD from the University of Minnesota.

Thomas Kalil received a BA in political science and international economics from the University of Wisconsin, Madison, and completed graduate work at the Fletcher School of Law and Diplomacy. He is the special assistant to the chancellor for science and technology at the University of California, Berkeley. He has been charged with developing major new multidisciplinary research and education initiatives at the intersection of information technology, nanotechnology, microsystems, and biology. He will also help to develop a broad array of partnerships between two of the California Institutes of Science and Innovation (Center for Information Technology Research in the Interest of Society and California Institute for Bioengineering, Biotechnology, and Quantitative Biomedical Research) and potential stakeholders in industry, government, foundations, and nonprofits. Previously, he served as the deputy assistant to President Clinton for technology and economic policy and deputy director of the White House National Economic Council (NEC). He was the NEC's "point person" on varied technology and telecommunication issues, such as the liberalization of Cold War export controls, the allocation of spectrum for new wireless services, and investments in upgrading America's high-technology workforce. He led a number of White House technology initiatives, such as those on nanotechnology, the next-generation Internet, bridging the digital divide, e-learning, increasing funding for long-term information technology research, making information technology more accessible to people with disabilities, and the growing imbalance between support for biomedical research and support for the physical sciences and engineering. He was also appointed by President Clinton to serve on the G-8 digital opportunity task force (dot force). He is the author of articles and op-ed pieces on

science and technology policy, nanotechnology, nuclear strategy, U.S.-Japan trade negotiations, U.S.-Japan cooperation in science and technology, the National Information Infrastructure, distributed learning, and electronic commerce.

David Kingsbury is the deputy chief program officer at the Gordon and Betty Moore Foundation. He received his PhD from the University of California, San Diego, and his BS and MS degrees from the University of Washington. From 1997 to 2003, he worked in the biotechnology industry at Chiron Corporation and later at Valigen in Paris, France. From 1992 to 1997, he was on the faculty at the Johns Hopkins University School of Medicine, where he was an associate dean of the School of Medicine and the director of the Division of Biomedical Information Sciences, the Genome Data Base, and the Welch Biomedical Library. He also was the chief information officer for Johns Hopkins University. Before joining the Johns Hopkins University faculty, he was professor of microbiology at George Washington University. From 1984 to 1988, he served as the assistant director for Biological, Behavioral, and Social Sciences at the National Science Foundation (NSF), where he was acting director for several months in 1984. At the time of his appointment to NSF, he was a professor of virology at the University of California, Berkeley. While at NSF, he served as the chair of two White House committees on biotechnology policy and regulation. His research has focused on the genetics and biochemistry of viruses and more recently computational biology. He was the founding editor-in-chief of the *Journal of Computational Biology* and is the author of a review text in medical microbiology and nearly 100 research papers and reviews.

Claudia Neuhauser is a professor and head of the Department of Ecology, Evolution and Behavior at the University of Minnesota. She received her PhD in mathematics from Cornell University after completing her undergraduate work in Heidelberg, Germany. Before joining the faculty of the University of Minnesota, Dr. Neuhauser served as a professor at the University of Southern California, the University of Wisconsin, Madison, the University of Minnesota, Minneapolis, and the University of California, Davis. Dr. Neuhauser's work on spatial stochastic processes addresses questions in population genetics, ecology, and evolution. Studies of genealogies under selection are focused on the development of methods for statistical tests for selection in spatially structured and unstructured populations. Work on spatial stochastic processes centers on mechanisms of coexistence in food webs, including disease dynamics. Mathematical models are used to investigate nonequilibrium dynamics after large-scale perturbations in natural and managed habitats with the goal of understanding their evolutionary and ecological consequences. Physiological models of phytoplankton, zooplankton, and bacteria are used to investigate ecosystem consequences of physiological tradeoffs.

Gregory A. Petsko is the Gyula and Katica Tauber Professor of Biochemistry and Molecular Pharmacodynamics and the director of the Rosenstiel Basic Medical Sciences Research Center at Brandeis University. He was elected to membership in the National Academy of Sciences in 1995 and to the Institute of Medicine in 2001. He has developed low-temperature methods in protein crystallography and their use to study enzymatic mechanisms and has pioneered the study of protein dynamics in enzymatic reactions. For over 25 years, he has worked to understand how enzymes achieve their extraordinary catalytic power, developing crystallographic methods for direct observation of productive enzyme-substrate and enzyme-intermediate complexes that led to techniques for studying protein crystal structures at very low temperatures. Recently, he took a sabbatical in Ira Herskowitz's laboratory at the University of California, San Francisco, learning yeast genetics to be able to combine the reductionist approach of biological chemistry with the whole-organism approach of genetics. He is a founding scientist of the combinatorial-chemistry company ArQule, Inc. He hopes to use genetic, biochemical, and biophysical tools to study structure-function relationships as they apply to in vivo and in vitro function.

Mariam Sticklen is a professor in the Department of Crop and Soil Sciences at Michigan State University. Since 1987, she has supervised and advised over 160 scientists (BS, MS, PhD, and sabbatical professors). She has played advisory roles at the international board of trustees level and at the U.S. National Academies level. She is developing systems to eliminate or minimize human and environmental risks posed by transgenic crops. Her expertise includes production of biofuels-related industrial enzymes, polymers, and pharmaceuticals in transgenic plants. Her research activities also center on improvement of agricultural crops of developing countries (Africa, India, Indonesia, Pakistan, Iran, and Turkey), development of crops that are tolerant to extreme abiotic factors (such as drought, high salinity, and low temperature), and reductions of pests and applications of hazardous pesticides through gene discovery, cloning, and genetic engineering.

Larry P. Walker is a professor of biological and environmental engineering at Cornell University. Dr. Walker did his undergraduate and graduate work at Michigan State University, receiving a BS in physics and an MS and a PhD in agricultural engineering. He joined the Cornell faculty in 1979. Dr. Walker's research interests can be grouped into four categories: single-molecule detection and analysis to elucidate hydrolytic and synergistic mechanisms of cellulases acting on microcrystalline cellulose; modeling, analysis, and optimization of solid-state fermentation processes for the production of natural products; modeling and optimization of submerged fermentation processes for the production of enzymes and other products; and single-molecule detection and nanofabricated devices for use in molecular ecology studies of microbial communities in high-solids degra-

dation processes. Dr. Walker's research activities are inherently multidisciplinary, involving researchers from applied and engineering physics, civil and environmental engineering, molecular biology and genetics, microbiology, plant pathology and plant biology. In addition, Dr. Walker is coordinator of the Biomolecular Devices and Analysis Program, the director of the Northeast Sun Grant Initiative, a former member of the National Biomass Research and Development Technical Advisory Committee, and coeditor of *Industrial Biotechnology*.

Janet Westpheling is a professor in the Department of Genetics at the University of Georgia. She earned her PhD in 1980 from the John Innes Institute in Norwich, England. Her primary research involves the control of gene expression in *Streptomyces* with emphasis on the study of carbon use and primary metabolism, and the strategies used by bacteria to regulate genes involved in morphogenesis and antibiotic production. *Streptomyces* is of particular interest because it produces most of the natural-product antibiotics used in human and animal health care. Dr. Westpheling serves on the *Journal of Bacteriology* Editorial Board and was chair of the Gordon Research Conference on Microbial Stress Response in 1996. She serves as a member of the Scientific Advisory Boards of several biotechnology companies interested in natural-product drug discovery and is a consultant to pharmaceutical and biotechnology companies. Dr. Westpheling participates annually in a course on fermentation technology offered by the Chemical Engineering Department at the Massachusetts Institute of Technology.

Appendix B

Presentations to the Committee

Keck Center of the National Academies
September 25, 2005

Ari Patrinos, U.S. Department of Energy, Office of Biological and Environmental Research; Michael Strayer, U.S. Department of Energy, Office of Advanced Scientific Computing Research
Overview of Genomics: GTL program

James Fredrickson, Pacific Northwest National Laboratory
Progress and potential of Genomics: GTL program and challenges it faces

George Church, Harvard Medical School
Role of Genomics: GTL in genome engineering and synthetic biology

Derek Lovley, University of Massachusetts, Amherst
Role of Genomics: GTL in bioremediation and biofuel research

George Pierce, Georgia State University
Role of Genomics: GTL in bioremediation research

Craig Venter, J. Craig Venter Institute
Role of Genomics: GTL in carbon-sequestration research

Elbert Branscomb, Lawrence Livermore National Laboratory
Role of Genomics: GTL in achieving DOE's mission—national-laboratory perspective

Michael Himmel, National Renewable Energy Laboratory (NREL)
Relation of Genomics: GTL to NREL programs

Eric Eisenstadt, The Institute for Genomic Research
Role of Genomics: GTL in advancing genomic research—nonprofit institution's perspective

Joel Parriott, Executive Office of the President, Office of Management and Budget (OMB); Michael Salamon, Executive Office of the President, Office of Science and Technology Policy (OSTP)
Genomics: GTL—OMB and OSTP's perspective

September 26, 2005

James Tiedje, Michigan State University
Role of Genomics: GTL in environmental research

Mark Guyer, National Institutes of Health, National Human Genome Research Institute (NHGRI)
Relation of Genomics: GTL to NHGRI's programs

Joanne Tornow, National Science Foundation (NSF), Directorate for Biological Sciences
Relation of Genomics: GTL to NSF's genomics programs

Ann Lichens-Park, U.S. Department of Agriculture (USDA), Cooperative State Research, Education, and Extension Service
Relation of Genomics: GTL to USDA genomics programs

Martin Godbout, Genome Canada
Genome Canada—successes, lessons learned, and relation to Genomics: GTL

Dawn Field, Oxford Center for Ecology and Hydrology
Environmental Genomics Thematic Programme Data Centre—successes, lessons learned, and relation to Genomics: GTL

Appendix C

Research Institutions That Have Received Funds through Contracts and Subcontracts from the Genomics: GTL Program

Institution	Funding (thousands of dollars)
National Laboratories	
Argonne National Laboratory	3,814
Brookhaven National Laboratory	4,288
Lawrence Berkeley National Laboratory	60,635
Lawrence Livermore National Laboratory	240
Los Alamos National Laboratory	4,068
Oak Ridge Institute for Science and Education	2,686
Oak Ridge National Laboratory	24,248
Pacific Northwest National Laboratory	45,830
Sandia National Laboratory	7,550
Other Institutions	
Biatech	2,370
Biotechnology Industrial Organization	68
Boston University	1,648
Brigham and Women's Hospital	1,288
Brown University	1,304
California Institute of Technology	4,175
Defense Advanced Research Projects Agency	1,300
Diversa Corporation	1,500
Gene Network Sciences	2,519

Institution	Funding (thousands of dollars)
Georgia Institute of Technology	2,203
Harvard University	11,861
Institute for Systems Biology	1,370
J. Craig Venter Institute	20,000
Johns Hopkins University	2,652
Keck Graduate Institute	460
Massachusetts General Hospital	632
Massachusetts Institute of Technology	2,272
Michigan State University	986
Northeastern University	766
Oregon State University	1,037
Scripps Research Institute	288
SoundVision Productions	992
Stanford University	4,300
University of California, Berkeley	1,277
University of California, Los Angeles	2,332
University of California, Merced	2,831
University of California, San Francisco	982
University of Chicago	1,644
University of Georgia	380
University of Illinois	400
University of Maryland Biotechnology Institute	1,510
University of Massachusetts	18,063
University of Missouri	1,863
University of Southern California	120
University of Washington	2,199
University of Wisconsin	2,695
Washington University	395

Appendix D

Examples of Research in Microbial Genomics Supported by Federal Agencies Other Than the Department of Energy

Agencies	**Microbial Genomics Program**
Department of Defense (DOD), jointly with National Institutes of Health (NIH) and Wellcome Trust and Burroughs Wellcome Fund	Sequencing of entire genome of malaria parasite *Plasmodium falciparum*
DOD in collaboration with Department of Energy (DOE) and NIH	Sequencing of entire genome of *Bacillus anthracis*
Department of the Interior (DOI)	Microbial research in natural resources, for instance, to identify microorganisms to control invasive species or to treat municipal sludge

Agencies	Microbial Genomics Program
Environmental Protection Agency (EPA)	Research on harmful and beneficial microorganisms found in environment, including microorganisms that are harmful to humans (mold and biofilms in water systems) or are involved in cleaning up environment (toxins removed from wetlands, microorganisms that can remove chlorinated solvents and other organics, or microorganisms that can sequester heavy metals, such as lead)
Food and Drug Administration (FDA)	Research and surveillance on microbial pathogens, research to develop rapid assessment technologies for pathogens and treatment strategies, and work on antibiotic resistance
National Aeronoutics and Space Administration (NASA)	Research on functional genomics of organisms in extreme environments to understand nature of life and to ensure human health during space travel
National Institute of Standards and Technology (NIST)	Genomics research (bioinformatics, managing data banks, and structural genomics), research in such applied fields as biological production of chemical products that are more economical and environmentally friendly, and protein engineering
National Institutes of Health (NIH)	Microbial research that addresses human health issues—disease-causing microorganisms, Human Genome Project, yeast research, and computational biology to support genome-enabled science

Agencies	Microbial Genomics Program
National Oceanic and Atmospheric Administration (NOAA)	Microbial research in support of sustaining healthy coasts and sustainable fisheries and work on microbial pathogens (algal blooms, parasites, and bacteria)
National Science Foundation (NSF)	Microbial research through a number of foundationwide initiatives and through unsolicited proposals and research on computational biology and bioinformatics
U.S. Department of Agriculture (USDA)	Microbial genomics that is relevant to mission of addressing regional and national problems and opportunities relevant to agriculture, food, forestry, and environment; and efforts to sequence animal and plant pathogens