DOE/SC-0090

DOE Genomics:GTL Roadmap

Systems Biology for Energy and Environment

August 2005

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

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

Preface

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

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

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

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

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

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

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

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

Executive Summary

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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