DOE Systems Biology Knowledgebase

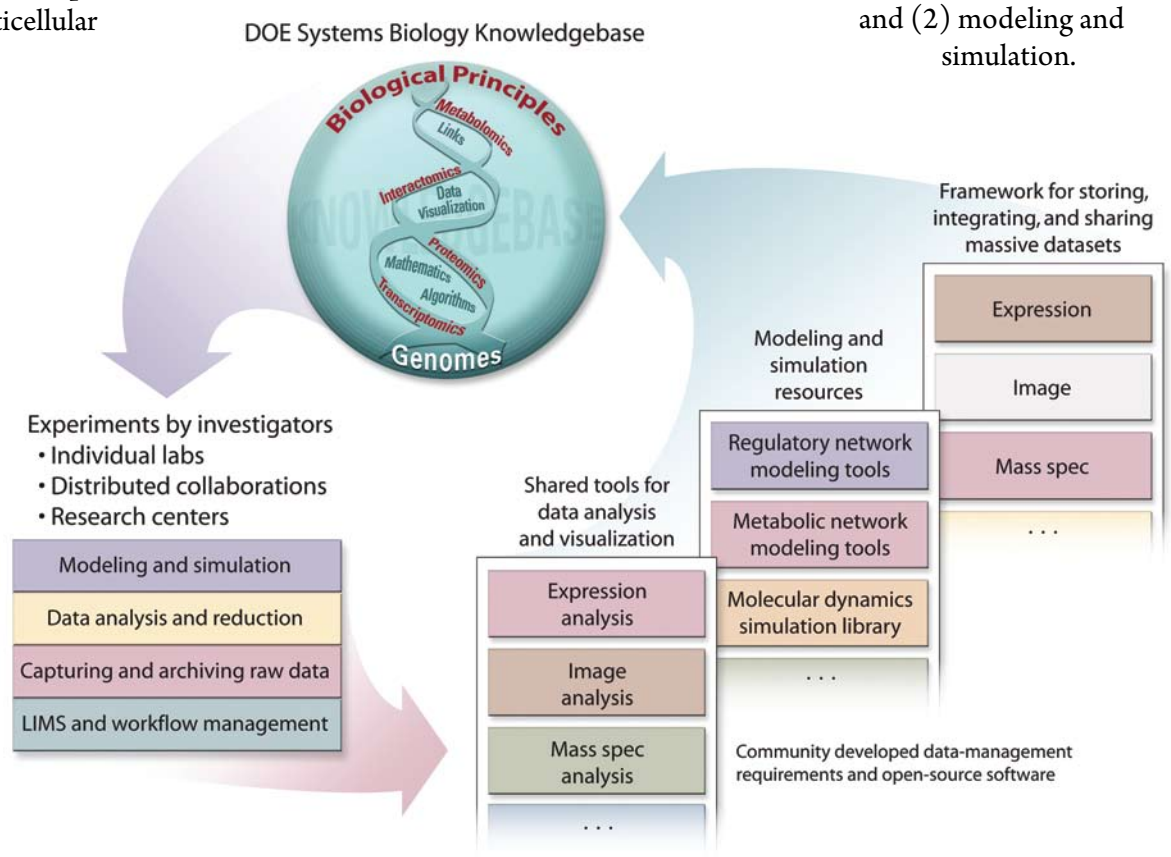
Community-Driven Cyberinfrastructure for Sharing and Integrating Data and Analytical Tools

Systems Biology: Enabling an Integrated Approach to Biological Research

Historically, most biological research has focused on parts of biological systems. The specialized nature of many projects has resulted in data, methods, and scientific advances developed in isolation—separated from other disciplines and often disconnected from other biological research areas. Major advances in experimental and computational capabilities are enabling researchers to merge insights from many different individual efforts into a more complete understanding of how biological components work together as a system. This systems biology approach uses computational science and modeling to collectively investigate and integrate all the dynamic components and interactions underlying the behavior of a living system—whether a single cell, complex community, or multicellular organism.

DOE's Vision for a Systems Biology Knowledgebase

The Department of Energy (DOE) Genomic Science program supports systems biology research to ultimately achieve a predictive understanding of microbial and plant systems for advancing DOE missions such as sustainably producing biofuels, investigating biological controls on carbon cycling, and cleaning up contaminated environments. To manage and effectively use the exponentially increasing volume and diversity of data resulting from its projects, the Genomic Science program is developing the DOE Systems Biology Knowledgebase (see figure below). Envisioned as an open cyberinfrastructure to integrate systems biology data, analytical software, and computational modeling tools that will be freely available to the scientific community, the Knowledgebase will drive two classes of work: (1) experimental design and (2) modeling and simulation.



DOE's Vision for a Systems Biology Knowledgebase. The systems biology research community requires the integration of a wide range of high-volume data and an open computational environment designed to support modeling, derivation of predictions, and exchange of data and analytical software.

Community-Driven Design of the DOE Systems Biology Knowledgebase

The success of the Knowledgebase will rely largely on its ability to meet the dynamic information needs of different user communities and the willingness of these communities to support open sharing of data, science, and software (see figure, The Community-Driven DOE Systems Biology Knowledgebase). When research data and information are not publicly available to the scientific community, a corresponding price is paid in missed opportunities, barriers to innovation and collaboration, and lost productivity resulting from inadvertent repetition of similar work.

Guiding Principles for the DOE Systems Biology Knowledgebase

- **Open access:** Data and methods are available for anyone to use
- **Open source:** Source code is freely available to access, modify, and redistribute
- **Open development:** Anyone can contribute to the development of Knowledgebase resources by following guidelines defined by the community

Genomic science and systems biology research represent a cultural change from isolated individual investigations to open community science. The central importance of embracing open biological science was presented in the 2009 National Research Council report, *A New Biology for the 21st Century*:

> "Traditional dissemination of results through publications in journals can convey only a fraction of the information that is generated in most experiments. To capture the full benefit of funded scientific work, one must maximize the ability to share that information. Information about research results that is not made accessible is lost to the rest of the research community…"

Input and feedback from the scientific community are needed to ensure that as the Knowledgebase develops, this infrastructure provides tools and services that are valuable and useful to researchers. To specify requirements for the Knowledgebase, the Genomic Science program is sponsoring a series of community-building workshops to engage experts from microbial genomics, plant genomics, supercomputing, and other disciplines. Output from these workshops and opportunities to contribute will be available from the DOE Systems Biology Knowledgebase Wiki site (see information box, p. 4).

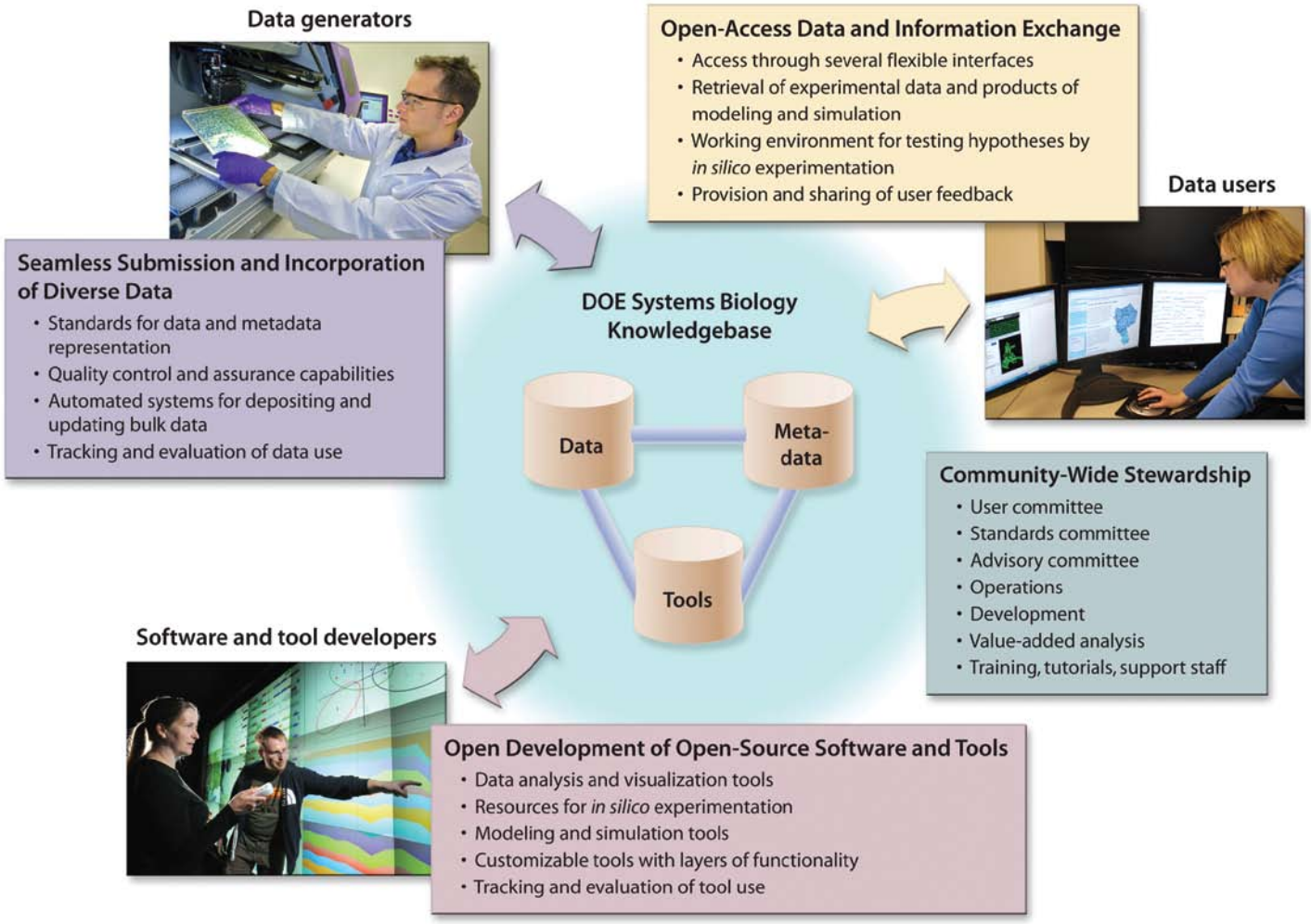
Facets of DOE Systems Biology Requiring a Knowledgebase

The emergence of systems biology as a research paradigm and approach for DOE missions has resulted in dramatic increases in data flow from a new generation of genomics-based technologies. In addition to accessing and managing this torrent of data, other aspects of DOE systems biology also compel the development of a Knowledgebase to facilitate the open sharing of this program's data, information, and analytical software. These aspects are described in the following paragraphs.

Integrating Data from Diverse Biological Systems Relevant to DOE Missions. Living systems possessing capabilities important to addressing the Department's missions in energy production, carbon cycling, and environmental remediation represent a vast range of biological, environmental, and biochemical diversity not observed in more traditional research targeting model organisms. Biological systems for DOE missions include microbial communities from the deep subsurface, thermal springs, cow rumen, guts of wood-eating insects, and other environments; root fungi and bacteria that influence carbon accumulation in plants; and trees and grasses for bioenergy feedstocks. The heterogeneous mix of data emanating from these investigations spans diverse environmental conditions and wide-ranging scales of time (nanoseconds to decades) and space (nanometers to kilometers).

Capturing the Environmental Context of DOE Biological Research. Not only are the biological systems diverse, but the environments with which these organisms intimately interact are even more varied. Understanding the environmental conditions that influence an organism's biological function requires knowing and describing the specific microenvironment immediately surrounding that organism—average conditions over larger scales of space and time are not sufficient. To enable comparisons among data and experimental results, each dataset from an environmental sample must be accompanied by metadata that provide contextual information. Having a common resource such as the Knowledgebase for collectively gathering data and experimental results will stimulate a community-wide effort for establishing guidelines and standards needed to adequately capture environmental metadata.

Accessing the Torrent of Data from High-Throughput Analyses. The large-scale genomic methods and high-throughput instrumentation used to study microbial and plant systems are generating enormous amounts of data



The Community-Driven DOE Systems Biology Knowledgebase. Using new open-source infrastructures and community-defined contribution standards, the Knowledgebase will facilitate the open development and sharing of data, metadata, and analysis tools to advance systems biology. *[Photo credits: Top, Lawrence Berkeley National Laboratory; bottom, Argonne National Laboratory; right, Oak Ridge National Laboratory]*

and information, much of which is archived in individual laboratories that often are inaccessible to the larger research community or impossible to search collectively. The rate of data production is rapidly outpacing analysis, and much of the information already generated could be more fully utilized to maximize biological discovery and reveal higher-level insights and trends occurring across research results from multiple labs.

Sharing Data and Information Across Large, Distributed Research Collaborations. The biological challenges addressed by DOE are complex and often require large multidisciplinary teams of researchers that approach similar problems from different directions to accelerate scientific progress. Several large research collaborations supported by the DOE Genomic Science program are examples of this team approach to biology that requires the well-coordinated sharing and use of large datasets among scientists in diverse locations.

Benefits of the DOE Systems Biology Knowledgebase

When fully deployed, the Knowledgebase will assume a new role for biological data management systems—from one traditionally perceived as bioinformatics support of mainstream experimental research to one in which computational analysis, modeling, and simulation capabilities drive a new era of *in silico* experimentation and hypothesis testing. As a unified framework linking otherwise disparate systems, the Knowledgebase will be an important tool to accelerate biological discovery for DOE missions and provide insights and benefits that can ultimately serve numerous application areas.

Democratizing Access to Experimental Data and Computational Capabilities. Biological research efforts (large and small) would gain access to dramatically more data and robust analytical and modeling tools that may not be available to smaller, individual projects. Scientists could integrate knowledge from their own research and also draw upon data generated from the entire research community.

Leveraging New Biological Insights to Advance Multiple Applications. The power of the systems approach to biology is rooted in the fact that—at the molecular level—all life is based on similar sets of fundamental processes and principles. Knowledge gained about one biological system, therefore, can advance the understanding of other systems when information is readily available in an integrated and transparent format. For example, the discovery of new regulatory pathways that influence plant biomass accumulation in bioenergy crops could also shed light on how these pathways affect carbon cycling in terrestrial vegetation or impact the productivity of agricultural crops.

Establishing the Foundation for Predictive Modeling of Biological Systems. For the first time, genomic sequence will be directly linked to the many downstream, multimodal analytical measurements of biochemical, cellular, and organismal activities. Only by developing an open infrastructure for mining, comparing, and interconnecting large biological and environmental datasets will we begin to build the comprehensive understanding needed to predict how the complex interplay between genomes and environments controls the behavior of biological systems.

Initial DOE Knowledgebase Development Efforts

Several efforts are under way to establish "proof of concept" for the Knowledgebase. DOE will gather community input and build support for the Knowledgebase by sponsoring workshops and an open Wiki to share ideas about requirements for this cyberinfrastructure. In addition, several small projects will explore a range of computational challenges facing DOE systems biology research. These projects will initiate strong collaborations among experimental researchers, bioinformatics specialists, computational biologists, and computer scientists to develop computational methods that could contribute to Knowledgebase development. Another effort is focusing on data management and information sharing issues underlying biological challenges common to all three DOE Bioenergy Research Centers. Each center is a multi-institution partnership working on high-risk, high-return breakthroughs in cellulosic biofuel production. By experimenting with prototype designs and documenting failures and successes, collectively, these activities will help lay the founding principles and design requirements for the Knowledgebase.

Contacts and Websites for More Information

Susan Gregurick
susan.gregurick@science.doe.gov
Office of Biological and Environmental Research
U.S. Department of Energy Office of Science

Wiki for Drafting DOE Systems Biology Knowledgebase Requirements
- sites.google.com/site/doekbase/

DOE Systems Biology Knowledgebase May 2008 Workshop Report
- genomicscience.energy.gov/compbio/

DOE Genomic Science Program
- science.doe.gov/ober/BSSD/genomicsgtl.html
- genomicscience.energy.gov

Genomic Science Program Information and Data Sharing Policy
- genomicscience.energy.gov/datasharing/

DOE Bioenergy Research Centers
- genomicscience.energy.gov/centers/

DOE Office of Biological and Environmental Research
- science.doe.gov/ober/

DOE Office of Science
- science.doe.gov

U.S. Department of Energy
- energy.gov

Printed with soy ink on recycled paper