Executive Summary

Biology has entered a systems-science era with the goal to establish a predictive understanding of the mechanisms of cellular function and the interactions of biological systems with their environment and with each other. Vast amounts of data on the composition, physiology, and function of complex biological systems and their natural environments are emerging from new analytical technologies. Effectively exploiting these data requires developing a new generation of capabilities for analyzing and managing the information. By revealing the core principles and processes conserved in collective genomes across all biology and by enabling insights into the interplay between an organism's genotype and its environment, systems biology will allow scientific breakthroughs in our ability to project behaviors of natural systems and to manipulate and engineer managed systems. These breakthroughs will benefit Department of Energy (DOE) missions in energy security, climate protection, and environmental remediation.

The Genomics:GTL Systems Biology Knowledgebase Workshop

To promote development of a data and information management system, or *knowledgebase,* DOE's Office of Biological and Environmental Research (OBER) hosted a workshop May 28–30, 2008, in Washington, D.C. Experts from scientific disciplines relevant to DOE missions and from the enabling technologies (e.g., bioinformatics, computer science, database development, and systems architecture) met to determine the opportunities and requirements for developing and managing this knowledgebase for OBER's Genomics:GTL program (GTL).

Workshop participants defined the proposed GTL Knowledgebase, or GKB, as an informatics resource that would focus on DOE science-application areas yet also be widely and easily applicable to all systems biology research. Also discussed were requirements for effective development of data capabilities for systems biology that could be applied specifically to plants and microbes (i.e., bacteria, archaea, fungi, and protists—unicellular eukaryotes such as microalgae) as well as to three areas of science related to DOE missions: (1) researching and developing biofuels, (2) advancing fundamental understanding of the global carbon cycle, and (3) understanding and using biological systems for environmental remediation. Participants were organized into working groups based on four knowledgebase themes: data, metadata, and information; data integration; database architecture and infrastructure; and community and user issues.

Summary Findings

The workshop highlighted DOE's unique and extensive data-management needs as a foundation of mission-inspired systems biology research. These needs require a principal GTL data resource, the GKB, with critical links to complementary systems supported by other agencies and community organizations worldwide. This knowledgebase would facilitate a new level of scientific inquiry by serving as a central component for the integration of modeling, simulation, experimentation, and bioinformatic approaches. The GKB also would be a primary resource for data sharing and information exchange among the GTL community. Furthermore, not only would the GKB allow scientists

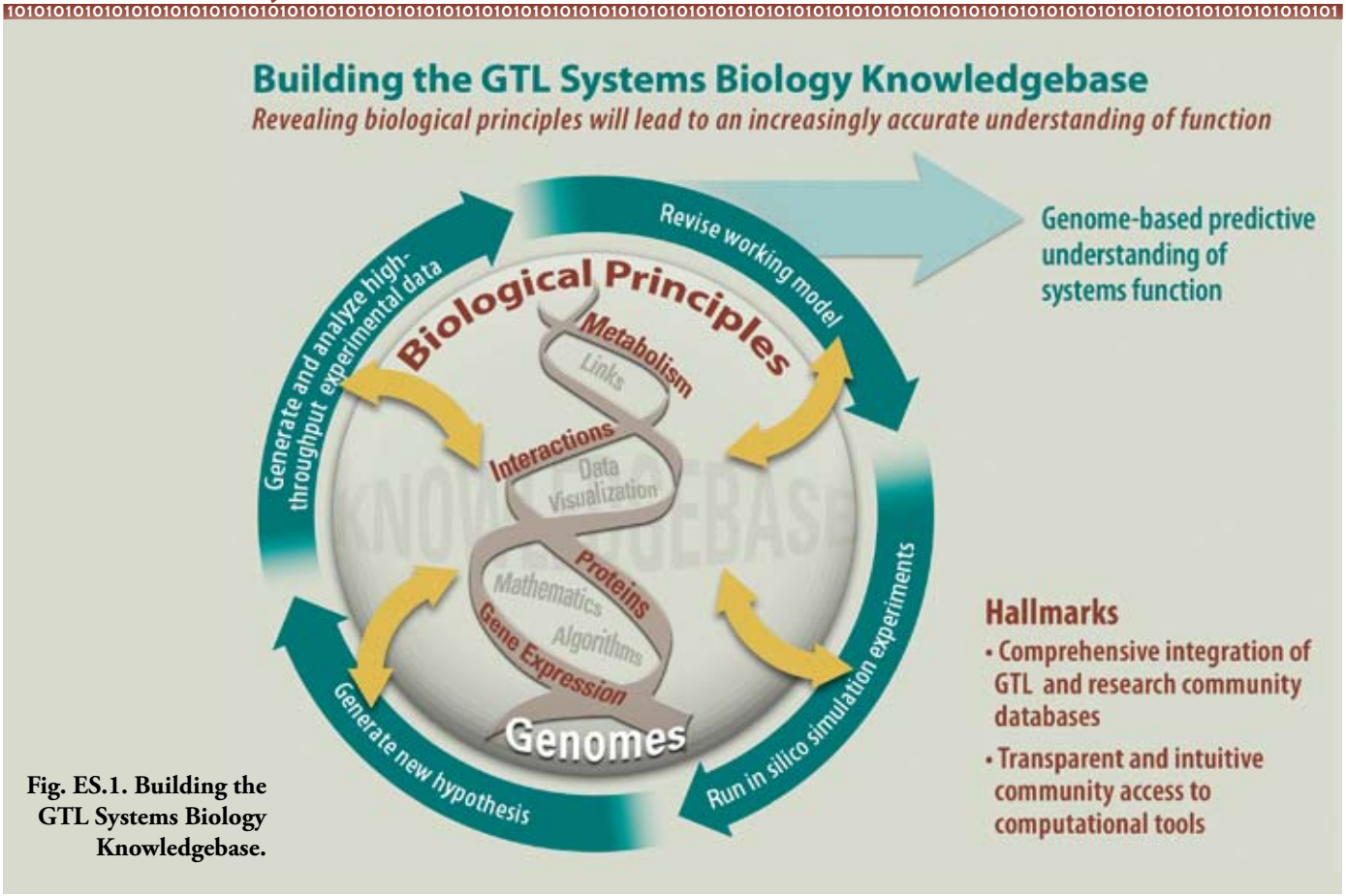

Fig. ES.1. Building the GTL Systems Biology Knowledgebase.

to expand, compute, and integrate data and information program wide, it also would drive two classes of work: experimental design and modeling and simulation. Integrating data derived from computational predictions and modeling, as envisioned in the knowledgebase project, would increase data completeness, fidelity, and accuracy. These advancements in turn would greatly improve modeling and simulation, leading to new experimentation, analyses, and mechanistic insight. Scientists' ever-increasing exploitation of the dynamic linkages among data integration, experimentation, and modeling and simulation—aided by the GKB—will advance efforts to achieve a predictive understanding of the functions of biological systems. The knowledgebase, therefore, must serve multiple roles, including (1) a repository of data and results from high-throughput experiments; (2) a collection of tools to derive new insights through data synthesis, analysis, and comparison; (3) a framework to test scientific understanding; (4) a heuristic capability to improve the value and sophistication of further inquiry; and (5) a foundation for prediction, design, manipulation, and, ultimately, engineering of biological systems to meet national needs in bioenergy, environmental remediation, and carbon cycling (see Fig. ES.1. Building the GTL Systems Biology Knowledgebase, above).

Summary Recommendations

The Department of Energy should establish the GTL Knowledgebase as a growing and extensible system of open and integrated biological, ecological, and environmental databases uniquely suited to DOE missions and distinct from, but linked to, other biological databases. To guide users and developers, the knowledgebase must be framed by a governance model with a set of user and programmatic interfaces;

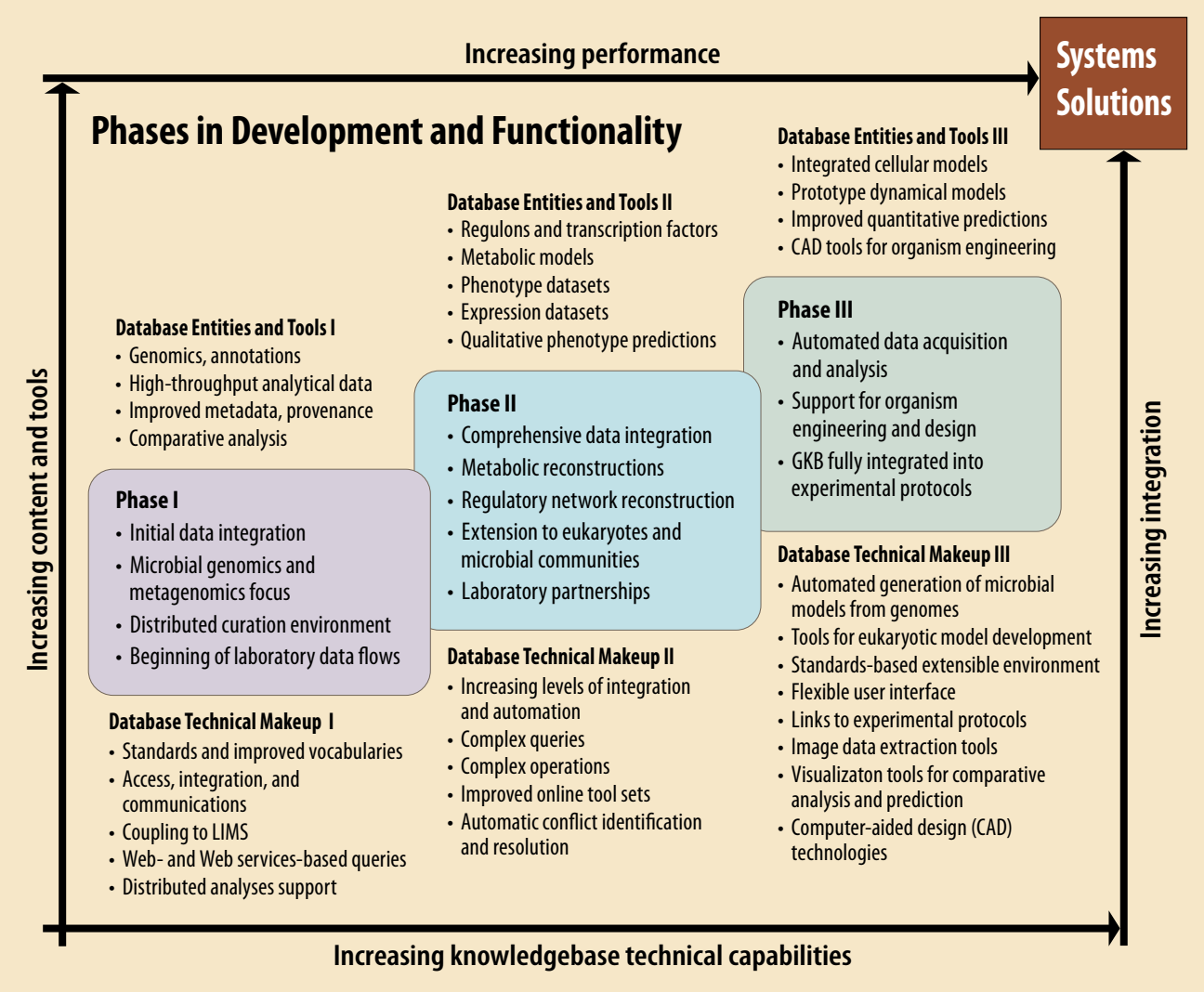

Fig. ES.2. Phases in DOE GTL Knowledgebase Development and Functionality. Phase I is centered around pulling together the components and developing functional elements. In Phase II, the components become more integrated, coupled, and automatic. In the final, mature phase, the knowledgebase becomes fully integrated, automatic, and transparent.

data-quality practices and standards; policies for data submission and data access; and a supporting communications, computing, and informatics infrastructure. Robust knowledgebase use among members of the scientific community would require a consonant suite of algorithms, tools, and services for data analysis, visualization, annotation, curation, extraction, and mining of datasets. Providing these resources would involve capturing a rapidly growing flow of data, correcting errors, and enlisting the expertise of researchers skilled in data integration, analysis, and extraction. Moreover, to support the ultimate goals of systems biology and DOE missions, the GTL Knowledgebase should be the focal point for a set of capabilities to reconstruct, model, and simulate biological and ecological systems. Workshop participants prioritized development of these integrated capabilities and outlined a strategy to implement each in phases to span a 5-year period (see Fig. ES.2. Phases in DOE GTL Knowledgebase Development and Functionality, above).

Implementation: Levels of Capabilities, Content, and Functionalities

The first and arguably most straightforward phase to implement involves the establishment of capabilities to gather new data for high-quality genomic annotation. This important process identifies and assigns biologically meaningful descriptions to DNA sequences, for example, by identifying genes, developing metabolic reconstructions, and creating estimates of regulons and regulatory circuitry. Additional challenges are associated with annotating the genome sequences in eukaryotes (e.g., protists, plants, and fungi) and metagenomic samples relevant to DOE missions. To meet these annotation challenges, the GTL Knowledgebase would require substantial automation and greater data and information depth (e.g., increased accuracy, consistency, and coverage) and breadth (e.g., expansion from hundreds to thousands of genomes). In this paradigm, the concept of annotation must be extended to encompass the growing suite of datasets and objectives of high-throughput experimental systems biology.

Building on the initial data integration and annotation of Phase I, the next developmental phase involves coupling knowledgebase components to bridge the spatial and process scales in an organism. This second phase includes draft reconstructions to progressively encompass greater complexity that would characterize individual proteins; their interactions to form molecular machines; and, ultimately, metabolic and regulatory pathways and networks in functioning cells. A synergistic, two-way, and iterative workflow is envisioned for this phase, in which annotations provide the foundation for reconstruction, and reconstruction imposes consistency on annotations. The knowledgebase must comprehensively integrate all relevant data and information for Phase II to be viable.

These integrated genome-scale (whole system) reconstructions will describe progressively more complex cellular networks, contributing to predictive modeling of physiological properties, behavior, and responses at the organismal level. The two mission-relevant and readily tractable layers of reconstruction to be developed at this stage would be metabolic and transcriptional regulatory networks in bacteria, archaea, and unicellular eukaryotes. Such reconstructions would lay the foundation for applying similar techniques to more complex systems—including plants and microbial communities—and have the potential to capture temporal and spatial aspects of systems behavior. They also would provide a natural framework for integration of various types of genomic and postgenomic data (e.g., proteomic).

The third phase involves predicting and manipulating the functions of biological systems. Accomplishing these objectives requires integrating different layers of reconstruction (e.g., metabolic and transcriptional regulatory networks) to generate more realistic, predictive models of the "stable states" of organisms. Enhanced modeling of these states would allow prediction of organism behavior in response to environment, support a new generation of hypotheses, and reveal novel insights for systems design and engineering. Furthermore, the dynamic modeling of transitions between stable states—resolved for space and time—would contribute to multiscale exploration and prediction of the behavior of systems. Critical to such research is the modeling of microbial communities (i.e., prokaryotic or eukaryotic organisms such as protozoans, bacteria, archaea, algae, and fungi) and ecosystems, which includes representing associations among biota, such as plants and microbes, and their interactions with the environment. Together, the modeling of stable states, communities, and ecosystems

will enable system investigation spanning all scales—from molecular to global. To achieve advanced modeling and predictive capabilities, Phase III of the knowledgebase must include acquisition of the experimental data needed to validate physiological and functional predictions.

In summary, the long-range goals of the GTL Knowledgebase are twofold: (1) enabling and providing support for progressively more inclusive, predictive modeling of various cellular processes, organisms, and communities and (2) facilitating the use of these capabilities to inform ecosystem-level models and engineering applications. Attaining these goals would require a knowledgebase framework that precisely and comprehensively integrates data and information critical to DOE missions.