DOE Systems Biology Knowledgebase Implementation Plan

Microbes

Plants

Metacommunities

Biological Principles

Metabolism

Integration

Interactions

Data Visualization

Proteins

Mathematics

Algorithms

Gene Expression

Computing

Predictive Understanding

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

DOE Systems Biology Knowledgebase Implementation Plan

As part of the U.S. Department of Energy's (DOE) Office of Science, the Office of Biological and Environmental Research (BER) supports fundamental research and technology development aimed at achieving predictive, systems-level understanding of complex biological and environmental systems to advance DOE missions in energy, climate, and environment.

DOE Contact

Susan Gregurick
301.903.7672, susan.gregurick@science.doe.gov

Office of Biological and Environmental Research
U.S. Department of Energy Office of Science

---

www.science.doe.gov/Program_Offices/BER.htm

Acknowledgements

The DOE Office of Biological and Environmental Research appreciates the vision and leadership exhibited by Bob Cottingham and Brian Davison (both from Oak Ridge National Laboratory) over the past year to conceptualize and guide the effort to create the DOE Systems Biology Knowledgebase Implementation Plan. Furthermore, we are grateful for the valuable contributions from about 300 members of the scientific community to organize, participate in, and provide the intellectual output of 5 workshops, which culminated with the implementation plan. The plan was rendered into its current form by the efforts of the Biological and Environmental Research Information System (Oak Ridge National Laboratory).

The report is available via

- **www.genomicscience.energy.gov/compbio/**
- **www.science.doe.gov/ober/BER_workshops.html**
- **www.systemsbiologyknowledgebase.org**

Suggested citation for entire report: U.S. DOE. 2010. *DOE Systems Biology Knowledgebase Implementation Plan*. U.S. Department of Energy Office of Science (www.genomicscience.energy.gov/compbio/).

DOE Systems Biology Knowledgebase Implementation Plan

September 30, 2010

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

The document is available via genomicscience.energy.gov/compbio/.

Table of Contents

Executive Summary v

1. Introduction 1

2. Near-Term Microbial Science Needs Supported by Kbase 10

3. Near-Term Plant Science Needs Supported by Kbase 33

4. Near-Term Metacommunity Science Needs Supported by Kbase 6/

5. Mid-Term Science and Leveraged Annotation Needs 92

6. Kbase Relationships with Existing or New Resources 95

7. System Architecture 102

8. Kbase Infrastructure Tasks and Timeline 111

9. Governance ..... 13 .

10. Project Management 136

Appendix A: Supporting Scientific Objective and Software Requirement Documents for Near-Term Microbial Science Needs ..... 15 .

Appendix B: Supporting Scientific Objective and Software Requirement Documents for Near-Term Plant Science Needs ..... 19 .

Appendix C: Supporting Scientific Objective and Software Requirement Documents for Near-Term Metacommunity Science Needs 213

Appendix D: Individual Reports from the 2009–2010 DOE Systems Biology Knowledgebase Workshops 217

Appendix E: References 39/

Appendix F: Acronyms 390

Appendix G: Contributors and Observers 175

DOE Systems Biology Knowledgebase Workshops and Organizers

- **Using Clouds for Parallel Computations in Systems Biology. Nov. 16, 2009, at the Supercomputing conference in Portland, Oregon.**
  [Co-organizers: Folker Meyer, Argonne National Laboratory (ANL); Susan Gregurick, U.S. Department of Energy (DOE); Peg Folta, Lawrence Livermore National Laboratory; Bob Cottingham, Oak Ridge National Laboratory (ORNL); and Elizabeth Glass, ANL]
- **Plant Genomics Knowledgebase Workshop. Convened jointly by the U.S. Department of Agriculture (USDA) and the U.S. Department of Energy (DOE) on Jan. 8, 2010, at the Plant and Animal Genome conference in San Diego.**
  [Co-organizers: Catherine Ronning, DOE; Susan Gregurick, DOE; Ed Kaleikau, USDA; Gera Jochum, USDA; and Bob Cottingham, ORNL]
- **DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop. Feb. 9–10, 2010, at the Genomic Science Awardee Workshop VIII and Knowledgebase Workshop in Crystal City, Virginia.**
  [Co-organizers: Susan Gregurick, DOE, and Bob Cottingham, ORNL. Co-chairs: Adam Arkin, Lawrence Berkeley National Laboratory (LBNL), and Robert Kelly, North Carolina State University]
- **DOE Systems Biology Knowledgebase Workshop at the 5th Annual DOE Joint Genome Institute (JGI) User Meeting. March 23, 2010, in Walnut Creek, California.**
  [Co-organizers: Susan Gregurick, DOE, and Bob Cottingham, ORNL. Co-chairs: Victor Markowitz, DOE JGI and LBNL, and Jill Banfield, University of California, Berkeley]
- **Knowledgebase System Development Workshop. June 1–3, 2010, in Crystal City, Virginia.**
  [Co-organizers: Susan Gregurick, DOE; Bob Cottingham, ORNL; and Brian Davison, ORNL]

These reports are available in Appendix D and at www.systemsbiologyknowledgebase.org.

Executive Summary

A knowledgebase is a cyberinfrastructure consisting of a collection of data, organizational methods, standards, analysis tools, and interfaces representing a body of knowledge. Driven by the ever-increasing wealth of data resulting from new generations of genomics-based technologies, systems biology is demanding a computational environment for comparing and integrating large, heterogeneous datasets and using this information to develop predictive models. As a leader in systems biology research, the Genomic Science program of the Office of Biological and Environmental Research (BER), within the DOE Office of Science, supports scientific research that seeks to achieve a predictive understanding of microbial and plant systems relevant to DOE missions (genomicscience.energy.gov). By revealing the genetic blueprints and fundamental principles that control the biological functions of these systems, the Genomic Science program advances the foundational knowledge underlying biological approaches to producing biofuels, sequestering carbon in terrestrial ecosystems, and cleaning up contaminated environments. To serve the research community and address the Genomic Science program's data-intensive computing needs, this document outlines the initial plan for creating a knowledgebase for systems biology.

As an open, computational environment for sharing and integrating diverse biological data types, accessing and developing software for data analysis, and providing resources for modeling and simulation, the DOE Systems Biology Knowledgebase (also called Kbase) will support a cultural change in biology from a focus on individual project-based efforts to open community science. The Knowledgebase would differ from current informatics efforts by bringing together the research products from many different projects and laboratories to create a comprehensive cyberinfrastructure focused on DOE scientific objectives in microbial, plant, and metacommunity (complex communities of organisms) research.

By democratizing access to data and computational resources, the Knowledgebase will enable any laboratory or project, regardless of size, to participate in a transformative community-wide effort for advancing systems biology and accelerating the pace toward predictive biology (see Fig. ES.1, below). Thus, the Knowledgebase will facilitate building a broader scientific community that will contribute to the fundamental science underlying DOE missions.

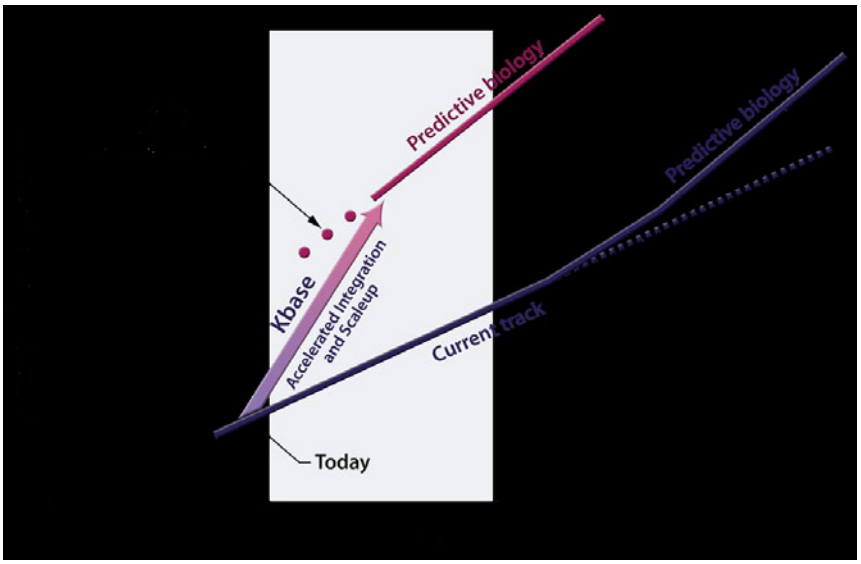


Fig. ES.1. A Faster Track to Predictive Biology. Knowledgebase-enabled integration of experimental data with models will accelerate the scientific advancements needed to improve inferences and achieve predictive biology. Building on the wealth of data being generated across many laboratories, the Knowledgebase will put biology on a new trajectory within the next decade. Scientists will hone their knowledge as they obtain answers to entirely new and more difficult generations of questions.

Systems Biology Knowledgebase Vision and Principles

The vision and justification for the Systems Biology Knowledgebase were defined in a May 2008 workshop report.[1] A key outcome for the many capabilities envisioned for the Knowledgebase (see sidebar, this page) is attaining more accurate models of dynamic cellular systems for microbes and plants. This requires a computational environment designed to support the iterative cycling of experimental design, analysis and integration of high-volume data, and modeling and simulation. As models of these cellular systems improve, they will address a progression of increasingly complex problems to help us understand and predict how these systems behave within a community of cells and organisms interacting with their environment. Ultimately, the Knowledgebase will allow users to perturb a biological system *in silico* (using "virtual experiments" on computer systems) and observe a predicted result.

By facilitating the efficient sharing of data, knowledge, best practices, and tools for rapidly developing and deploying applications for systems biology, the Knowledgebase will reduce the duplicative effort of individually establishing and maintaining similar resources for hundreds of laboratories and databases. Thus, researchers could direct more effort to scientific discovery. This open sharing and leveraging of the products from publicly funded scientific work will catalyze multidisciplinary collaborations and maximize the use and benefit of experimental results, analytical software, and modeling tools generated throughout the entire research community.

To provide the diverse capabilities envisioned for the Knowledgebase, infrastructural components will be distributed across many locations. Knowledgebase coordination, however, will be centralized and based on the following principles guiding development and operation:

- Provide open access to data, open contribution, and open-source software development—to the greatest extent possible—while simultaneously respecting a reasonable level of protection and temporary embargoes to allow publication and career development.
- Engage key stakeholders in developing the Knowledgebase, defining metrics for success, and assessing Knowledgebase performance in meeting the needs of the communities it serves.
- Support high-level policies (e.g., establishing standards for usability, interoperability, and contribution) recommended by a community-based Governance Board. This Governance Board combines features of a user advisory board and a scientific advisory board. Executive decisions (such as specifics on implementation) should be made by Project Management working closely with DOE management and the stakeholder community.

Capabilities Envisioned for the DOE Systems Biology Knowledgebase

- Curation of data, models, and representations of scientific concepts.
- Analysis (including method comparison) and inventory of results.
- Simulations and model modifications and improvements.
- Prediction-based simulation and analysis to form new hypotheses.
- Experimental design and comparison between predictions and results.

Community-Developed Plan for Knowledgebase Implementation

Building on the vision defined in the May 2008 workshop, the cumulative output of community participants in a series of five DOE-sponsored workshops[2] established the conceptual design, workflows, scope, and science to be addressed by the initial implementation of the DOE Systems Biology

[1] The May 2008 workshop report, *Systems Biology Knowledgebase for a New Era in Biology*, is available online (genomicscience.energy.gov/compbio/).

[2] Workshops are listed on p. iv of this report.

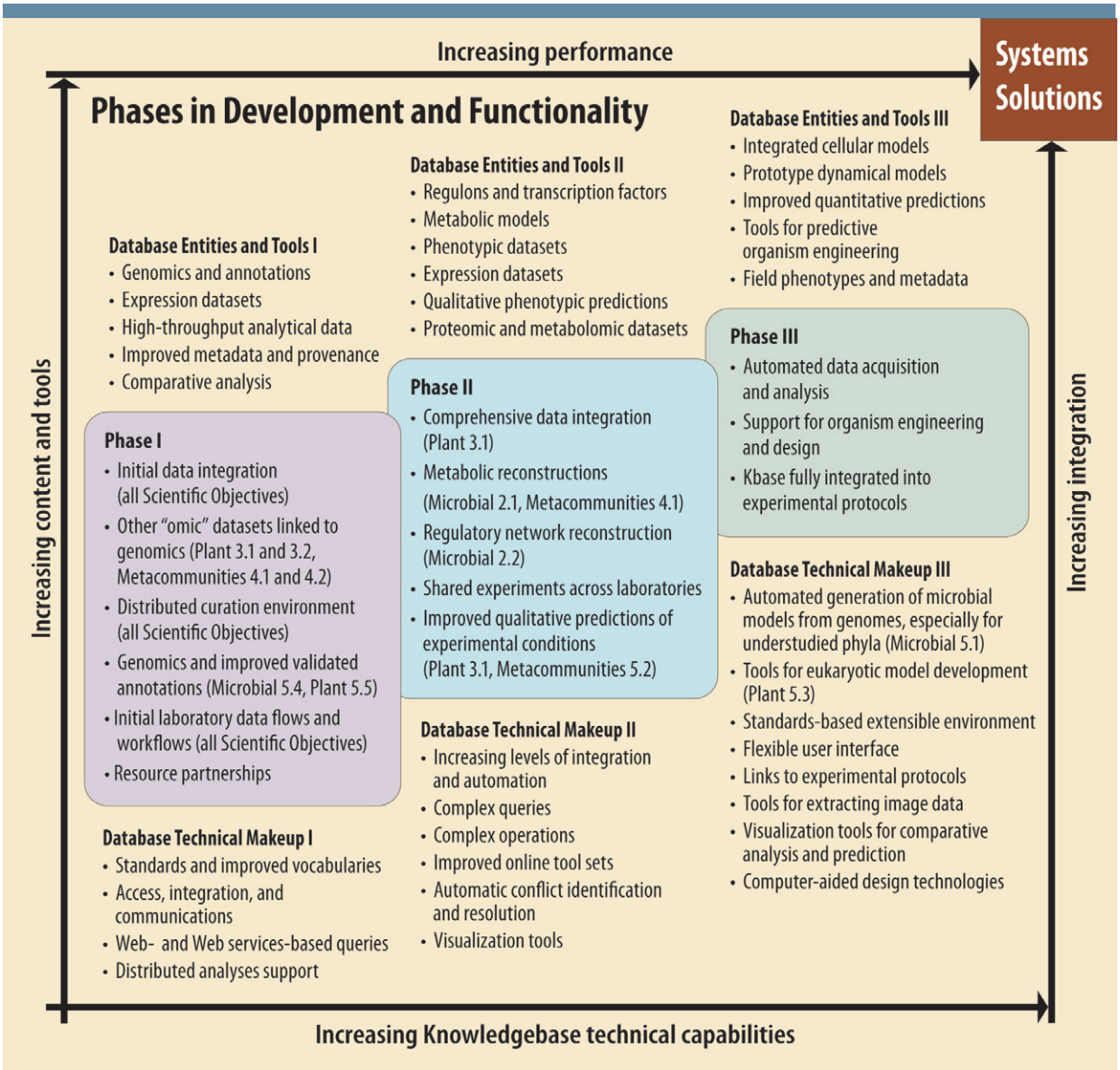


Fig. ES.2. Phases in Development and Functionality in the DOE Systems Biology Knowledgebase. This table shows three phases of technology development from less to more mature (lower left to upper right, respectively). The state of technology development for the biological systems (microbes, plants, and metacommunities) addressed in this report is in different stages of maturity. The notations in parentheses (e.g., Microbial 2.1) refer to the Science Area listed in Table ES.1, p. x. Technologies for microbial research and analysis currently are well into Phase I. Upon implementing this plan, the Microbial Scientific Objectives will move fully into Phase II. Though technologies are less mature for plant and metacommunities, deploying the implementation plan will result in substantial progress in Phase I and Phase II. [Updated from page vii in the 2008 workshop report, *Systems Biology Knowledgebase for a New Era in Biology*.]

Knowledgebase. Technologies in computer science, bioinformatics, and data management are available now to begin the transformation of the Knowledgebase vision into reality by creating an adaptable computational environment designed for expansion and modification over the coming decade. Therefore, the Knowledgebase will be implemented in phases characterized by progressively increasing functionality (see Fig. ES.2, this page). The 3 to 5 years covered in the implementation plan will move the community from Phase I into Phase II.

One clear consensus among workshop participants is that the Knowledgebase initially should target and achieve success in specific, focused scientific objectives that were identified, developed, and prioritized

as near-, mid-, and long-term needs at the workshops. Near-term priorities were described in the greatest detail, with progressively fewer details given for the other objectives. To define the core scientific objectives, workshop participants discussed and identified the key research goals that need to be solved for three science areas relevant to DOE systems biology: microbes, plants, and metacommunities. For the six near-term scientific objectives that were identified as priorities for the Knowledgebase, representatives from the biological and computational communities worked together to translate the objectives into experiment workflows, computing system requirements, and detailed implementation plans specifying the tasks, outcomes, and integrating infrastructure needed to accomplish the objectives (see Fig. ES.3, this page). Additional objectives describe mid-term science and leveraged annotation needs that will be addressed over the coming decade.

Community involvement is critical for the success of this effort. Many consider achieving community "buy-in" for the Knowledgebase as important as overcoming the technical challenges faced in developing a community infrastructure. This powerful commitment to engaging the community is reflected in the valuable contributions from about 300 scientists who participated in the workshops culminating in the DOE Systems Biology Implementation Plan. This level of community involvement will need to continue as planning for the Knowledgebase transitions to its implementation,

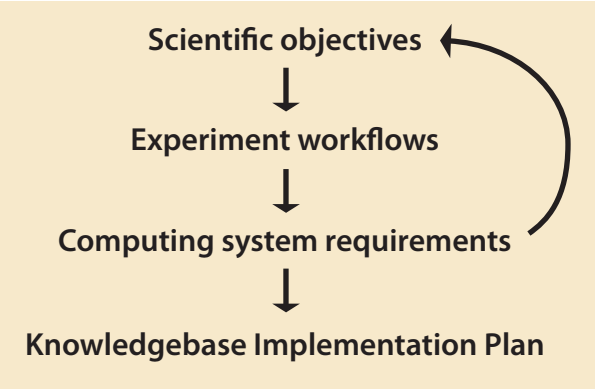


Fig. ES.3. Process for Community Development of Knowledgebase Implementation Plan.

which will require the expertise and skills from many different groups within the scientific community. Broadly, these groups represent plant and microbial researchers who design experiments and generate data; computational biologists and bioinformaticians who will develop the analysis methods and simulations that help interpret the data; and computer scientists, database developers, and software engineers who will develop the Knowledgebase infrastructure (see Fig. ES.4, p. ix).

DOE has a proven capability for linking strengths in biology and computational sciences in coordinated projects and programs. Genomic Science program collaborations involving experimental scientists, technology developers, and computational biologists have resulted in a deep understanding of specific microbes and microbial communities. In addition to advancing these ongoing efforts, the Knowledgebase will provide a unified framework for linking these different collaborations so that insights, workflows, and analytical programs resulting from these studies are more readily applied to investigations of more complex plant systems and metacommunities.

Knowledgebase Priorities and Scientific Objectives

The microbial, plant, and metacommunity science needs and objectives that will drive Knowledgebase development are listed in Table ES.1, p. x and described briefly in the following pages. The near- and mid-term science needs define the initial and immediate plans for the DOE Systems Biology Knowledgebase. Additional goals were identified for longer-term activities but were not further developed in detail for this implementation plan.

Microbial Sciences

In the microbial science area, the first objective is to improve the utility of metabolic network models, especially for microbes involved in biofuel production and bioremediation, so that metabolic engineering produces more predictable

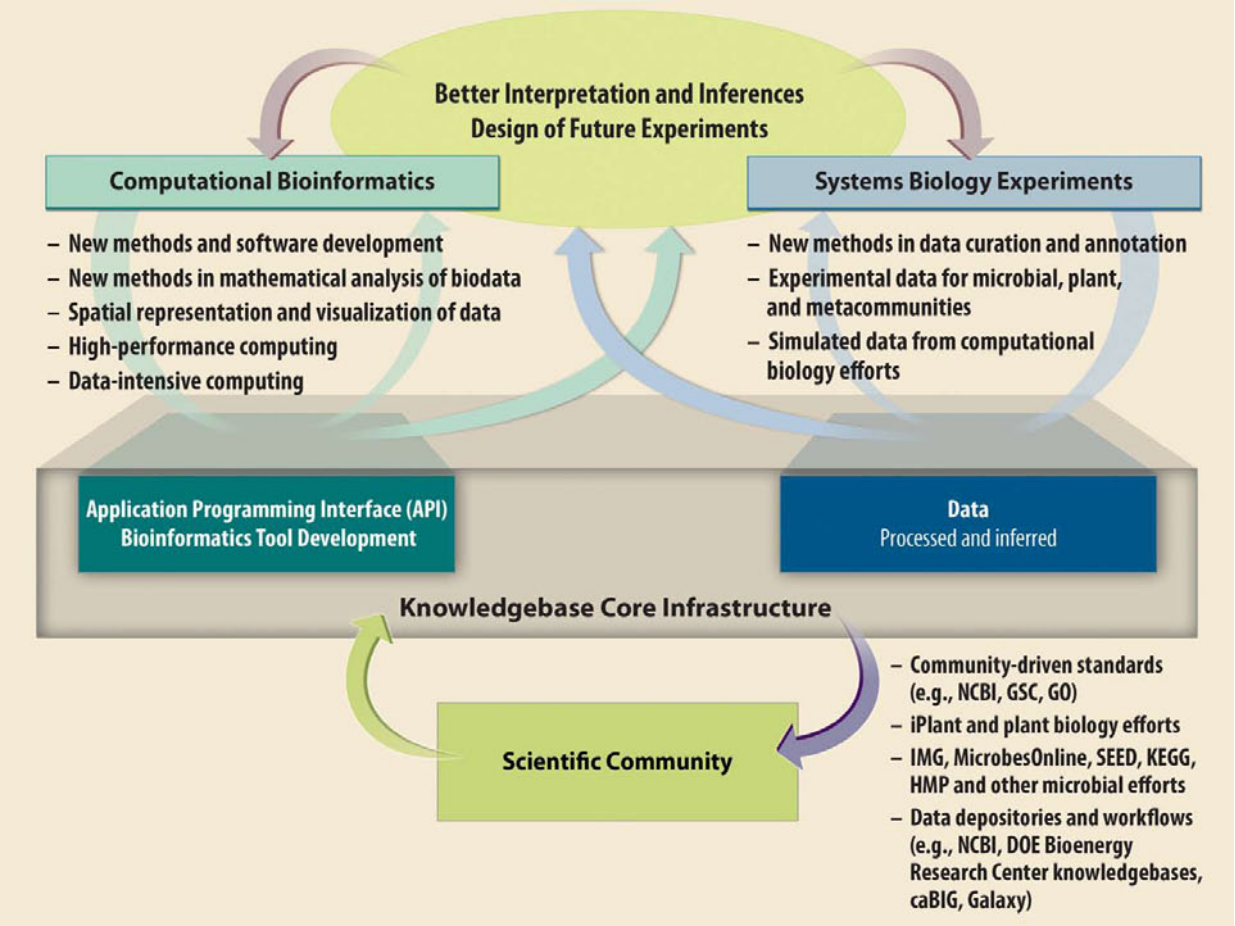


Fig. ES.4. Relationship Between the DOE Systems Biology Knowledgebase and the Larger Scientific Community.

results. The second objective is to enable automated inference of gene regulatory networks based on gene expression profiling data and then to validate inferred networks to improve prediction of cellular behavior and fitness.

Microbial Scientific Objective 1

Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function

The scientific community seeks to understand and manipulate the metabolic potential of organisms in order to understand growth and phenotypes. More specifically, this objective involves reconstructing metabolic networks, predicting the growth of organisms from their metabolic networks, understanding organisms' metabolic potential, providing scientists with software tools to interrogate and visualize metabolic networks, and enabling engineers to quickly determine the strategies necessary to remodel metabolism for specific purposes. Objective 1 will increase the speed and automation of metabolic network reconstruction and comparison and improve the accuracy of metabolic network predictions. This knowledge will lead to the informed modification of specific enzymes or the introduction of entirely new pathways, allowing researchers to determine better strategies for manipulating mass or energy flow in microorganisms. Achieving this capability will require integrating new experimental data with existing data and models of metabolic pathways, as well as developing methods to automatically create new metabolic reconstructions from newly sequenced organisms.

Current research and development in metabolic networks primarily involve two approaches:

Table ES.1. Near-Term, Mid-Term, and Leveraged Annotation Needs Supported by the DOE Systems Biology Knowledgebase

(Section numbers in first column refer to main report)

Section	Science Area	Scientific Objective
		Near-Term Science Needs
2.1	Microbial	Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function
2.2	Microbial	Define Microbial Gene Expression Regulatory Networks
3.1	Plant	Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype
3.2	Plant	Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Analysis, Comparisons, and Modeling
4.1	Metacommunities	Model Metabolic Processes Within Microbial Communities
4.2	Metacommunities	Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses About Their Function
		Mid-Term Science Needs
5.1	Microbial	Analyze Understudied Microbial Phyla
5.2	Metacommunities	Interpret Metagenomic Data to Identify Conditions Required for Growth by Key Microbial Communities Relevant to DOE Missions
5.3	Plant	Construct, Simulate, and Validate Plant Life Models
		Leveraged Annotation Needs
5.4	Microbial	Integrate Descriptions and Annotations of Microbial Genomic Features
5.5	Plant	Improve Plant Genome Annotation Datasets and Make Them More Accessible

(1) evaluating novel microbes to identify and improve desired metabolic phenotypes (e.g., recent work on *Clostridium phytofermentans*) and (2) manipulating the metabolic pathways of well-characterized microbes to enable novel functionality (e.g., initiatives to engineer cyanobacteria for photosynthetic production of alkanes and isoprenoids and recent achievements in hydrocarbon production from *Escherichia coli*). Given DOE's interest in metabolic engineering for biofuel production and bioremediation, the development of sophisticated metabolic modeling methods and experimental data for a select set of DOE-relevant organisms is a high-priority, near-term objective.

Microbial Scientific Objective 2

Define Microbial Gene Expression Regulatory Networks

In response to dynamic and competitive environments, microbes must deploy the products of diverse gene sets to survive and prosper. Expression

of the correct sets of genes at the correct levels could confer the best competitive advantage given the organism's genetic complement and the current environment. The mechanisms within cells that sense the environment and determine which gene sets should be deployed at what levels, thereby coordinating different stages of the microbe's growth and development, are collectively called the gene regulatory network. Knowledge of this network is the foundation for predicting, controlling, and designing the behaviors of microbes and their community.

The first component of this objective is to enable automated inference of gene regulatory networks, relying principally on expression profiling data. The second is to extend these inferred networks to include additional data types, both to refine network predictions and to test them. Prioritization should be given to those organisms that are key to DOE missions, with a focus on regulatory paradigms of greatest relevance to the microbe in question.

This high-priority objective can achieve near-term goals, but completion may take 2 to 10 years. The advent of genomic technology and the availability of many microbial genomes enable the development of capabilities providing data and tools from which regulatory networks and their behaviors may be inferred rather than directly measured. The regulation of networks of interactions within and among microbes defines their ability to remediate environments, improve energy crop growth, process biomass into fuels, and sequester carbon, among other things.

Plant Sciences

The first objective in the plant science area is to establish the capability to predict changes in plant biomass properties caused by genetic or environmental changes. This predictive capability is based on the mining of data that reflects the complex relationships among the physical properties of plants, their genetic makeup, and the environment in which they are growing. The second objective is to develop the capability to organize and analyze data from regulatory omics (e.g., transcriptomics, proteomics, and other large-scale molecular analyses) to improve understanding of how plants regulate gene expression in key plant species relevant to DOE missions. This capability will be critical for understanding genes, their functions, and regulation and then using this understanding to engineer plant growth and development and, in particular, biomass accumulation.

Plant Scientific Objective 1

Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype

The Knowledgebase will provide computational infrastructure to support and contextualize experimental plant phenotype data to an extent that enables researchers to predict changes in the physical properties of biomass resulting from environmental change, genetic diversity, or manipulation. Achieving this goal depends on the creation of a robust semantic infrastructure for collecting, annotating, and storing diverse phenotypic and environmental datasets. These data include measurements such as photographic images and analytical spectra that capture visible phenotypes and chemotypes related to yield, physiological performance, and sustainability.

Subsequently, the Knowledgebase will be used for data mining and analysis to understand the genetics underpinning desirable plant biomass properties relevant to DOE missions (e.g., biomass yield, conversion efficiencies to biofuels, and the ability to sequester soil carbon or contaminants). Specifically, it will serve as a basis for software applications that extract, quantify, and catalog phenotypic features from the diverse datasets and relevant metadata (data describing the primary data generated from experiments or other analyses) for data mining and further analysis. Development of a robust, semantic infrastructure for plant phenotyping research is a high-level, mid-term objective that could be carried out in 3 to 5 years and in synergy with the ongoing efforts of the National Science Foundation (NSF) iPlant Collaborative. By providing the community with a comprehensive

collection of experimental and phenotypic data for plant feedstocks important to DOE, this objective will accelerate the development and redesign of feedstocks with plant architectures, cell-wall characteristics, and other properties that improve biofuel production and carbon biosequestration.

Plant Scientific Objective 2

Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Analysis, Comparisons, and Modeling

Plant regulation is known to control key aspects of plant carbon allocation and partitioning, which are critical to biomass composition and soil carbon accumulation. Regulation is also a critical distinguishing characteristic between annuals and perennials and other aspects related to sustainability. To date, we have limited understanding of how plants regulate gene expression and how this is manifested in the cell.

Assembling regulatory omics data from plant biology into common platforms is essential to DOE's systems biology mission. This objective seeks to collect several key types of regulatory omics data and associated quality metadata for six target plant species: *Brachypodium*, *Chlamydomonas*, poplar, sorghum, switchgrass, and *Miscanthus*. The assembly begins with genomic and RNA expression data (from arrays or RNA-Seq) along with small RNA and target RNA information, differential RNA processing and decay information, epigenetic markers such as DNA methylation and histone modifications, as well as available proteomic data. In the near term (1 to 3 years), classical transcriptomic data (microarrays and mRNA-Seq) as well as small RNA and basic proteomic data will be assembled. These internal and external data will be publicly accessible with user-friendly web interfaces and downloadable for power users.

Metacommunities Science

The first objective in the metacommunities science area is to determine the metabolic role of each organism residing in a community and understand which community features provide adaptive robustness to environmental change. This information will lead to improved characterizations of microbial community physiology and ecology, which are necessary for designing strategies to accelerate or ameliorate microbial activity for environmental remediation or carbon sequestration. The second objective allows scientists to study microbial communities to discover novel functions and genes within these communities. Data generated in large-scale metagenomics projects can provide the information necessary to better understand the function of poorly characterized genes. The resulting data provide actionable hypotheses about the function of many genes that have yet to be studied in detail. Additionally, scientific efforts associated with this objective will lead to the discovery of new genes that perform useful biological functions of relevance to DOE priority areas such as energy production, carbon cycling and biosequestration, and environmental remediation.

Metacommunities Scientific Objective 1

Model Metabolic Processes Within Microbial Communities

An overarching need for systems biology is to determine the metabolic role of each organism or key species residing in a community. This objective focuses specifically on modeling the metabolic processes within a microbial community, which requires developing metagenomics workflows and systems biology tools. In the near term, the Knowledgebase will develop workflows to analyze metagenomes and other data from microbial communities and leverage existing data and tools to create descriptive community metabolic models. The data and metadata will include the full range of current systems biology tools. Both top-down (metagenomics) and bottom-up (multispecies models) approaches were formulated for near- and mid-term goals. Eventually, these models will allow us to not only predict, but actively drive changes in the community in desired directions (e.g., accelerate environmental processes relevant to DOE missions, including environmental remediation, cellulose degradation, or carbon biosequestration).

The integration of different types of experimental measurements relating to metabolic activity is necessary for (1) generating hypotheses about the nature of interactions among community members and interactions between the community and the local environment, (2) generating hypotheses about the organisms and pathways responsible for the community's metabolic activities, and (3) predicting how the community will respond to environmental changes or to the introduction of new microorganisms. These proposed objectives will lead to improved characterizations of microbial community physiology; such characterizations are necessary for designing strategies to either accelerate biotransformation activity (e.g., uranium bioremediation) or ameliorate the outcome (e.g., acid mine drainage). Developments in microbial community understanding also will have direct benefits in understanding ecosystems and direct improvements in carbon cycling (and biosequestration in soils), as well as in biofeedstock production (via plant-associated microbial communities)—an area of immediate interest to DOE and the U.S. Department of Agriculture (USDA).

Metacommunities Scientific Objective 2

Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses About Their Function

One reason to study microbial communities is to determine the novel functions and genes of organisms within these communities. Data generated in large-scale metagenomics projects can provide the information necessary to better understand the function of poorly characterized genes. As metagenomic (along with metatranscriptomic and metaproteomic) data are rapidly coming online, a critical scientific objective is the development of approaches for mining the data to identify previously unknown genes and for leveraging the wealth of metadata associated with metagenomic datasets. Information about gene-organism co-occurrence can help identify testable hypotheses about the function of newly identified or poorly characterized genes. Additionally, scientific efforts associated with this objective will lead to the discovery of new genes that perform useful biological functions of relevance to DOE priority research areas. Improvements in identifying unknown genes and their function will help to reduce potential error propagation in gene-calling databases.

Reliable functional annotations are critical prerequisites of a successful research program in systems biology. This objective will accelerate efforts aimed at characterizing the function of currently understudied genes. Additionally, the tools developed as part of this project will be valuable assets to scientists generating new datasets by allowing them to leverage Knowledgebase-associated datasets in the analysis process and to generate actionable hypotheses. A key element is to handle diverse types of associated metadata.

Further Scientific Goals: Mid-Term Science and Leveraged Annotation Needs

The community identified other desirable and achievable scientific goals to improve functionality of the Systems Biology Knowledgebase. Several feasible medium- and high-priority needs were considered important for the Knowledgebase. One of three mid-term scientific needs is to analyze understudied microbial phyla. The goal of this scientific objective is to understand the role of unclassifiable members of a microbial community in terms of genetic and phenotypic comparison. To achieve this objective, physiologic and metabolic datasets must be linked to metagenomic annotations to provide context and evidence. Another mid-term objective is metagenomic interpretation to identify conditions required for growth by key microbial communities relevant to DOE missions. This would improve our ability to cultivate (and isolate) target species from these communities. The third science need is to construct, simulate, and validate plant life models to enable semiautomated inference, construction, simulation, validation, and query of complex, multilevel (i.e., gene, protein, metabolite, small RNA, organelle, cell, and tissue) datasets. These plant life models would be used to

integrate and explore experimental data types collected during studies of plant feedstocks that impact bioenergy production and carbon cycling.

Two of the identified near-term science needs are for improving annotation of both microbes and plants—high-priority objectives that would be immediately leveraged by the Knowledgebase project. In addition, the increasing number of large and complex metagenomic sequence data (hundreds of gigabases for soil, for example) requires advances in algorithms for assembly. The DOE Joint Genome Institute (JGI) is the lead organization in primary sequencing and annotation for organisms of DOE and community interest. These organisms include microbes, plants, fungi, and microbial communities. The DOE JGI is pursuing and developing plans to improve its approaches for incorporating ongoing technology advancements. Programmatically, the DOE JGI would have the primary mission to develop and carry out implementation of improved annotation pipelines. The DOE JGI and the Knowledgebase will closely collaborate to reach these mutual goals.

Infrastructure and Architecture

The DOE Systems Biology Knowledgebase will be a large-scale system that:

- Makes massive amounts of biological data freely available to the scientific community, through hosted services and as links to external resources.
- Provides high-performance and scalable computational resources.
- Supports a large user community with tools and services to enable researchers to use the Knowledgebase.

To meet these requirements, the Knowledgebase must be designed with a highly elastic architecture that enables computer scalability on demand to meet the ever-changing computational requirements of scientific users. This elastic architecture must be supported by continual expansion and scaling to accommodate new data, computational platforms, and software innovations. The overall goal for the architecture is to support the creation of a broad-based, scalable Knowledgebase that provides a set of services to underlying data and computational resources. Decisions about the design and implementation of the architecture are critically important to the efficient and low-cost sustainability of the Knowledgebase. These decisions will be based on the following core set of architectural principles defined in the plan:

- **Open.** Provide the community with a published set of open-source application programming interfaces (APIs) to access Knowledgebase resources in an automatic fashion using software.
- **Extensible**. Enable the community to use the APIs to extend the capabilities of core Knowledgebase resources.
- **Federated.** Provide users with transparent access to a federation of physically distributed heterogeneous computational and data resources.
- **Integrated.** Create mechanisms to integrate existing databases and tools essential for the DOE systems biology community.
- **Exploit data locality.** Implement mechanisms for transparently moving requested analyses to execution sites that can best exploit data locality and provide maximum performance.
- **Modular.** Promote modular, component-based design for codes that can be readily connected to build pipelines for executing complex, multi-step analyses.
- **Scalable.** Expand Knowledgebase system architecture to accommodate increased use and functionality by transparently incorporating additional computational and storage resources.

The Knowledgebase infrastructure must be a rich collection of services and hardware. The problems faced by scientists require a variety of computing and data platforms and applications that do not fit nicely into a system based on a single hardware or software platform. Knowledgebase hardware and services include data repositories, data storage or data warehouses, data centers at multiple

locations, virtualization, data parallel processing on commodity hardware, cluster computing, and high-performance computing (HPC). Data and metadata representation and registries are other key aspects. This collection also will enable semantics-based searches of metadata (such as ancillary experimental data, ontologies, controlled vocabularies, and data models). With the inclusion of ESnet and Internet2 as the underlying network backbone, the Knowledgebase infrastructure is a cloud-based system providing a unique and valuable resource for biologists and offering these capabilities:

Platform as a Service. The Knowledgebase will provide a software platform for users to share, use, develop, and deploy bioinformatics applications that serve DOE systems biology. The platform will support users in exploiting the computational and data resources in the Knowledgebase cloud. By providing facilities that support the complete life cycle, from building to using Knowledgebase-enabled software, users can receive the full benefits of the Knowledgebase infrastructure.

Infrastructure as a Service. This will allow users to leverage the Knowledgebase hardware, thereby reducing local operational costs associated with purchasing, installing, and maintaining hardware as well as reducing the burden on local facilities to house the hardware. Advances in hardware virtualization now make it possible for users to create images of their local system that can be shared via the Knowledgebase with other users, enabling the replication of scientific results and the sharing of analysis environments.

Data as a Service. This will provide community data curation services and allow users to store, access, share, and curate heterogeneous data in the Knowledgebase, reducing the need to buy additional storage and to scale their existing infrastructure. Providing data services to the biological research community in a time when data accumulation rates are increasing exponentially will enable research scientists to spend more resources focusing on biological problems.

The primary architecture recommendation is for a layered architecture blueprint (see Fig. ES.5, p. xvi). The four layers will consist of a user access layer, an infrastructure layer, a federation layer, and a layer of federated hardware resources. The foremost task for the Knowledgebase platform is to provide the scientific user with access to the underlying Knowledgebase-associated data, shielding the user from how that access is achieved (e.g., federated versus centralized, cloud-based versus central server). It also should provide the user with elementary analysis and visualization tools to apply to that data, mechanisms for storing intermediate results, standards for exchanging data between tools, and ways to connect analysis tools for creating *ad hoc* workflows. In addition, the platform should provide a low-threshold infrastructure for tool development, reuse, and dissemination.

The implementation plan recommends that the Knowledgebase initially consist of up to seven data centers on ESnet, upgraded to interconnections at 100 gigabytes (GB), each with petabyte (PB) storage. The storage is expected to double every 2 years. Each scientific data center would be associated with one of the six scientific objectives, and one data center would focus on coordinating the Knowledgebase core infrastructure development. Coordination of infrastructure development would be replicated as required to the other locations. Co-locating and sharing computational resources at some data centers, if possible, also are recommended. Compute clusters for virtualization could be co-located with compute clusters that support data parallel applications. For example, the cluster to support virtualization is expected to have 1000 to 3000 nodes running standard Linux with 1 to 2 PB of scratch storage. It is expected that HPC resources will be provided by existing DOE Office of Advanced Scientific Computing Research (ASCR) facilities. Facilities must have space and infrastructure to expand.

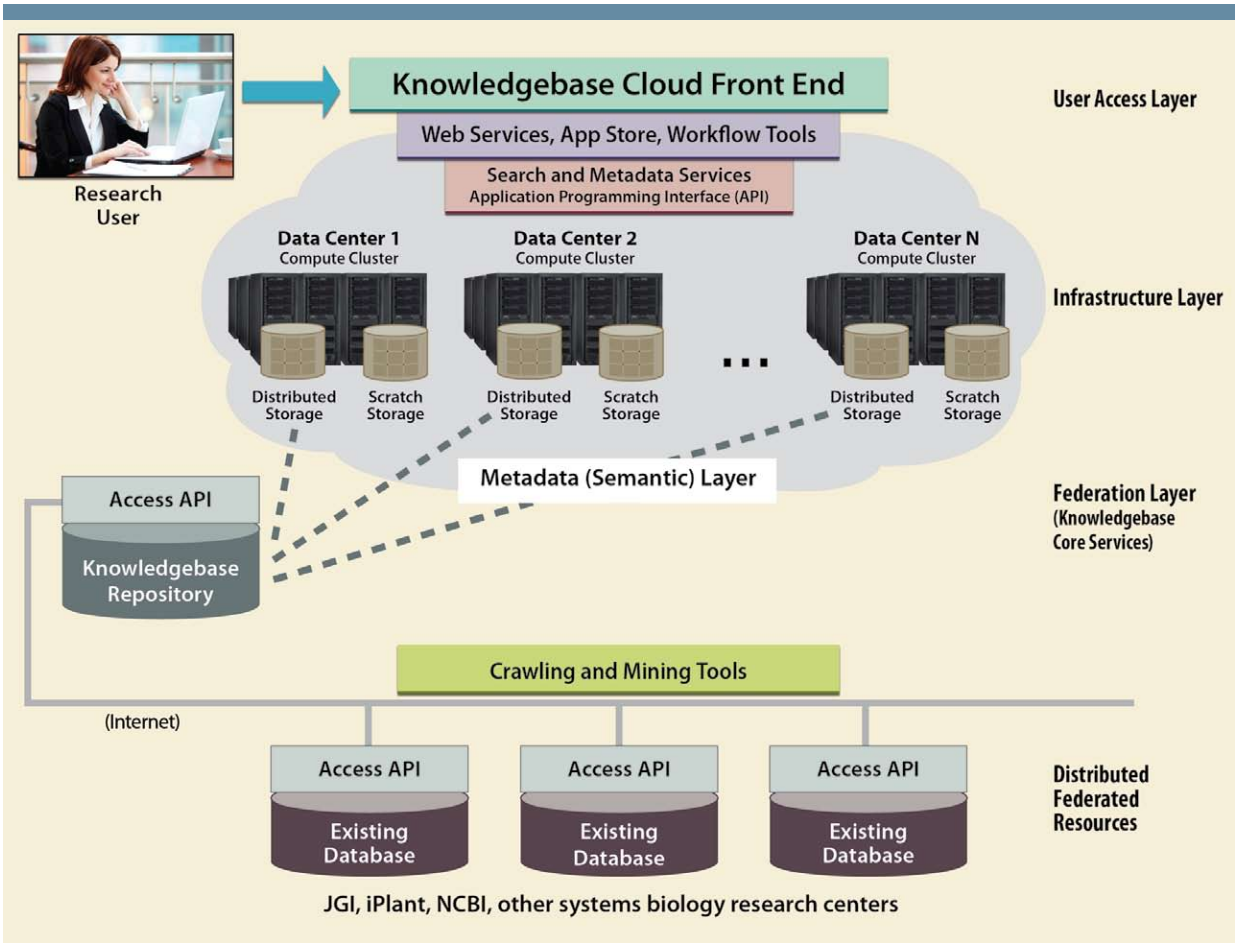


Fig. ES.5. Knowledgebase Architecture Overview. The architecture comprises four layers. The higher layers will span all systems within the Knowledgebase. The federated layers will reside within each of the specific data centers. Not every federated resource may be in every data center. The purpose of each layer and component is described in Chapter 7.

Key Systems Biology Knowledgebase Partnerships

The Knowledgebase is providing a unique impetus toward support and acceleration of the DOE systems biology research community. However, this effort is not operating in isolation of other synergistic efforts. There are critical partnerships that will be leveraged and are included in this implementation plan. These include the DOE JGI, DOE ASCR, National Center for Biotechnology Information (NCBI), and the NSF iPlant Collaborative.

The Knowledgebase will work with the DOE JGI to ensure that analysis tools developed are cross-compatible and that sequencing data and experimental data are shared to support a robust annotation system.

The Knowledgebase effort may employ several ways to leverage the exascale computing capability being developed in ASCR. Most of the scientific targets for the Knowledgebase are data driven and at a scale that, for the foreseeable future, likely could be met by a more moderate sized community cluster. However, opportunities exist for "co-design" partnerships between ASCR and the Knowledgebase to address unique problems such as the combinatorial analysis of biological networks. These partnerships may lead to better solutions for the potential "all versus all" comparisons in data-rich biological problems.

Another key partnership would be with NCBI, the major repository of primary sequence data. In addition to archiving bibliographic information (e.g., PubMed), NCBI has begun collecting more comprehensive biological information other than sequence data. NCBI recognizes the value of the Knowledgebase as an infrastructure that can fill the gap in the analysis and understanding of biological systems by providing users with a single portal to a variety of tools, resources, and multiple data types. The Knowledgebase will work with NCBI to share experimental data, cross-reference analysis resources, develop community-supported standards for new types of data, and develop tools that are cross-compatible for data analysis and data visualization. A Knowledgebase-NCBI working group will be formed and will meet on a regular basis to facilitate this collaboration.

Finally, the NSF iPlant Collaborative is another key partner that focuses on connecting plant biologists with plant breeders and supporting computational and analysis resources for these user communities. iPlant also is developing hardware and software tools for phenotyping plants in the field. In collaboration with the Integrated Breeding Platform, iPlant will support seed storage, phenotyping databases, pedigree support, and portable software and hardware tools useful for field biologists, and these resources can be leveraged for the Knowledgebase plant objectives. The Knowledgebase also will work with the iPlant community to establish common data standards and cross-compatible analysis tools.

Knowledgebase Development Timeline

The DOE Systems Biology Knowledgebase Implementation Plan describes the tasks needed to provide the research community with a comprehensive cyberinfrastructure to advance systems biology over the next several years. The basic timeline of the project is shown in Fig. ES.6, p. xviii. Details of the specific tasks and needed expertise are presented within the report.

The Implementation Plan outlines additional work to continue and expand initial Knowledgebase efforts over the next decade. By providing open access to data and tools that can address biological problems in various application areas, the Knowledgebase will have impacts beyond the scope of the specific targeted objectives, and it will directly impact the pace of biological research throughout the broader scientific community. Ultimately, the Knowledgebase will provide access to data, simulations, and tools to continue to move biology from a descriptive to a predictive science. The ability to make inferences based on broad community-derived datasets will help answer current research questions and will allow new (currently unanswerable) questions to be posed and tested.

The success of the Knowledgebase will be determined not only by demonstrating clear progress toward accomplishing the focused scientific objectives outlined in this report, but also by how effectively the research community can use and benefit from Knowledgebase resources and services. To be effective, the Knowledgebase must identify and address the needs of stakeholder communities and coordinate with other synergistic efforts. As a resource that is accessible to all, the Knowledgebase will catalyze new collaborations across disciplines and provide the community with a computational environment for testing hypotheses and investigating biological systems at a scale and scope not possible today. The Knowledgebase has the potential to open a new paradigm of biological science, truly engaging a systems approach.

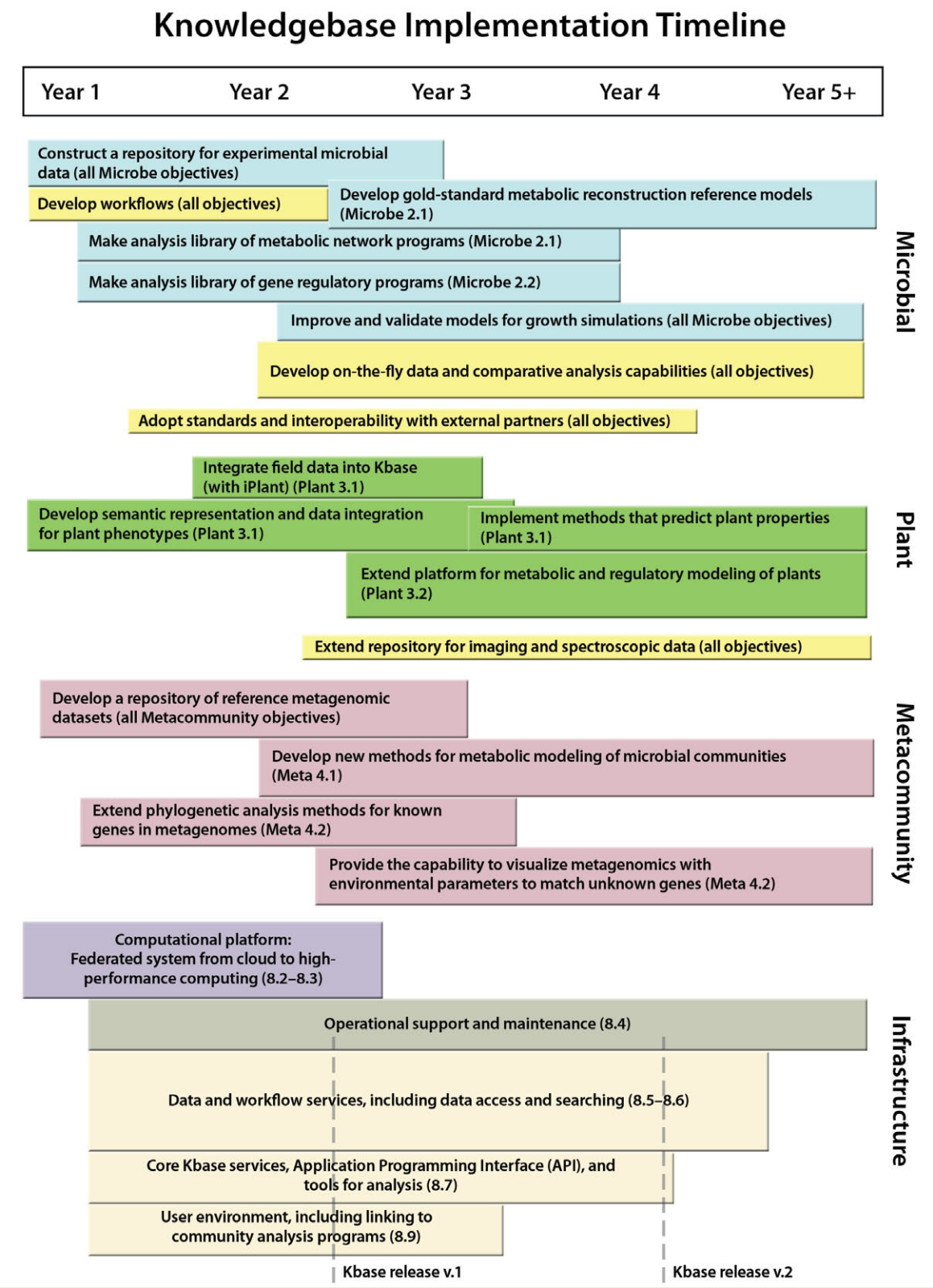


Fig. ES.6. High-Level Timeline for the Knowledgebase Implementation Plan. Light blue boxes represent tasks in the Microbial Science Area. Green boxes represent tasks in the Plant Science Area. Pink boxes represent tasks in the Metacommunity Science Area. Yellow boxes are tasks that apply to all objectives across all science areas. Infrastructure includes hardware (purple), operations (gray), and software (beige). Box height indicates relative level of effort for each task.

1. Introduction

The Department of Energy (DOE) Genomic Science program within the Office of Biological and Environmental Research (BER) supports science that seeks to achieve a predictive understanding of biological systems (genomicscience.energy.gov). By revealing the genetic blueprint and fundamental principles that control plant and microbial systems relevant to DOE missions, the Genomic Science program is providing the foundational knowledge that underlies biological approaches to producing biofuels, sequestering carbon in terrestrial ecosystems, and cleaning up contaminated environments.

1.1 Knowledgebase Purpose and Vision

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. To manage and effectively use this ever-increasing volume and diversity of data, the Genomic Science program is developing the DOE Systems Biology Knowledgebase (Kbase)—an open, community-driven cyberinfrastructure for sharing and integrating data, analytical software, and computational modeling tools.

Ultimately, a fully functional Kbase cyberinfrastructure is envisioned not only to include storage, retrieval, and management of systems biology data and information, but also to enable new knowledge acquisition and management through free and open access to data, analysis tools, and information for the scientific research community (see sidebar, A Fully Functional Systems Biology Knowledgebase, this page). Knowledgebase capabilities would include:

- Curation of data, models, and representations of scientific concepts.
- Analysis (including method comparison) and inventory of results.
- Simulations and model modifications and improvements.
- Prediction-based simulation and analysis to form new hypotheses.
- Experimental design and comparison between predictions and results.

A Fully Functional Systems Biology Knowledgebase

- Kbase will provide a computational environment for researchers to contribute data and analysis methods to model dynamic cellular systems of plants and microbes at a high level of accuracy. Such modeling will include many of these systems within a cell and a community of cells and organisms interacting with their environment. Ultimately, Kbase will allow users to perturb a system *in silico* and observe a predicted result.
- Kbase will serve as a productive cyberinfrastructure environment for storing, retrieving, managing, and analyzing systems biology data, thereby avoiding duplication of these efforts in hundreds of laboratories and databases.
- Kbase will maximize the use and benefit of research products by leveraging community-wide capabilities, experimental results, and modeling efforts.

Although numerous data repositories and databases have been developed throughout the systems biology community, many have varying amounts and quality of data, and some can be challenging to use by segments of the research community outside the narrow field of experts for whom these resources were designed. Bioinformatics efforts typically have been developed within smaller research groups. The broader research community is limited in its ability to take advantage of these tools. The current range of resources is scattered, difficult to access and search collectively, and often disconnected from related resources with important information. An integrated, community-oriented data and informatics resource such as Kbase would provide a broader and more powerful interface for conducting systems biology research relevant to BER's complex, multidisciplinary challenges in energy and environmental science.

1.2 Community-Developed Implementation Plan

The basis for developing the DOE Systems Biology Knowledgebase Implementation Plan was to engage the DOE biological research community to define core scientific objectives in key areas such as microbial, plant, and metacommunity (complex communities of organisms) research. The scientific objectives must answer the question, "What is the scientific or research goal that needs to be solved?" The related "requirements" establish workflows and provide details for accomplishing these objectives.

This report documents the conceptual design and outlines the initial plan for creating the DOE Systems Biology Knowledgebase to serve the systems biology scientific community and support DOE missions in the biological sciences. Successfully building such a system depends on sufficiently detailed science-driven objectives and their associated requirements and tasks, as articulated within this document. Based on community input from a series of five workshops, this document represents the cumulative output of these workshops and establishes the scope and plans necessary to begin the Kbase effort. One clear consensus among research community members involved in this effort is that Kbase initially should target and achieve success in specific, focused scientific objectives. Once these objectives were identified and developed at the first four workshops, they were prioritized as near-, mid-, or long-term needs at the final workshop in June 2010. Near-term priorities were described in the greatest detail, with progressively fewer details given for the other objectives.

Although workshop participants described more than 10 scientific objectives that could be accomplished over the next decade, six scientific objectives were selected as the highest priority. This prioritization was based on the overall impact and feasibility of the goals in the next few years. Two objectives were chosen from each of the three science mission areas: microbes, plants, and metacommunities. These six objectives were then developed into implementation plans that outline the tasks and workflows necessary to accomplish the defined research goals. An implementation plan also was developed for the Kbase infrastructure and architecture.

1.3 Knowledgebase Roles and Attributes

A knowledgebase is a computerized collection of data, organizational methods, standards, analysis tools, and interfaces representing a body of knowledge. For the DOE Systems Biology Knowledgebase, these interoperable components will be contributed from the research community and integrated into the system over time, resulting in an increasingly advanced and comprehensive resource. Key elements of the Kbase vision are defined in a May 2008 DOE workshop report, *Systems Biology Knowledgebase for a New Era in Biology*, (genomicscience.energy.gov/compbio/). Incorporating insights and recommendations from researchers with many different areas of expertise, ranging from environmental science to bioenergy, this 2008 workshop report highlights several roles Kbase will need to serve, including:

- An adaptable repository of data and results from high-throughput experiments.
- A collection of tools to derive new insights through data synthesis, analysis, and comparison.
- A framework to test scientific understanding.
- A heuristic capability to improve the value and sophistication of further inquiry.
- A foundation for prediction, design, manipulation, and, ultimately, engineering of biological systems.

Kbase will differ from current informatics efforts by integrating data and information across projects and laboratories. This integration requires Kbase to be an open community-wide effort (see Fig. 1.1 DOE Systems Biology Knowledgebase: Establishing a Systems Biology Framework, p. 4) rather than a monolithic project overseen and contributed to by only a few people. Kbase also will need to be more standardized than today's informatics resources. Although standardized components may not be cutting edge, they will be more interoperable, enabling comparisons among different laboratories and thus yielding important new insights. Standardization will involve not only data but also experimental protocols. As described in a recent *Science* article (Bell et al. 2009), biology—as with other areas of science—is demanding data-intensive computing. For systems biology, the computation is less numerical processing and more the mining and comparison of large datasets.

Another fundamental feature is that Kbase development will have a more mature software engineering approach. In the past, biologists not necessarily trained in state-of-the-art computational technologies were responsible for selecting and applying the tools needed to meet the computing needs of their individual laboratories. However, the exponential increase in the amount of DNA sequence and other data being generated requires the support of a more robust and integrative computational infrastructure. This infrastructure will allow analyses to be shared and distributed within a community and will enable researchers to quickly adapt new analytical methods developed by the entire research community. In this way, Kbase will encourage research and development based on the latest computational technologies.

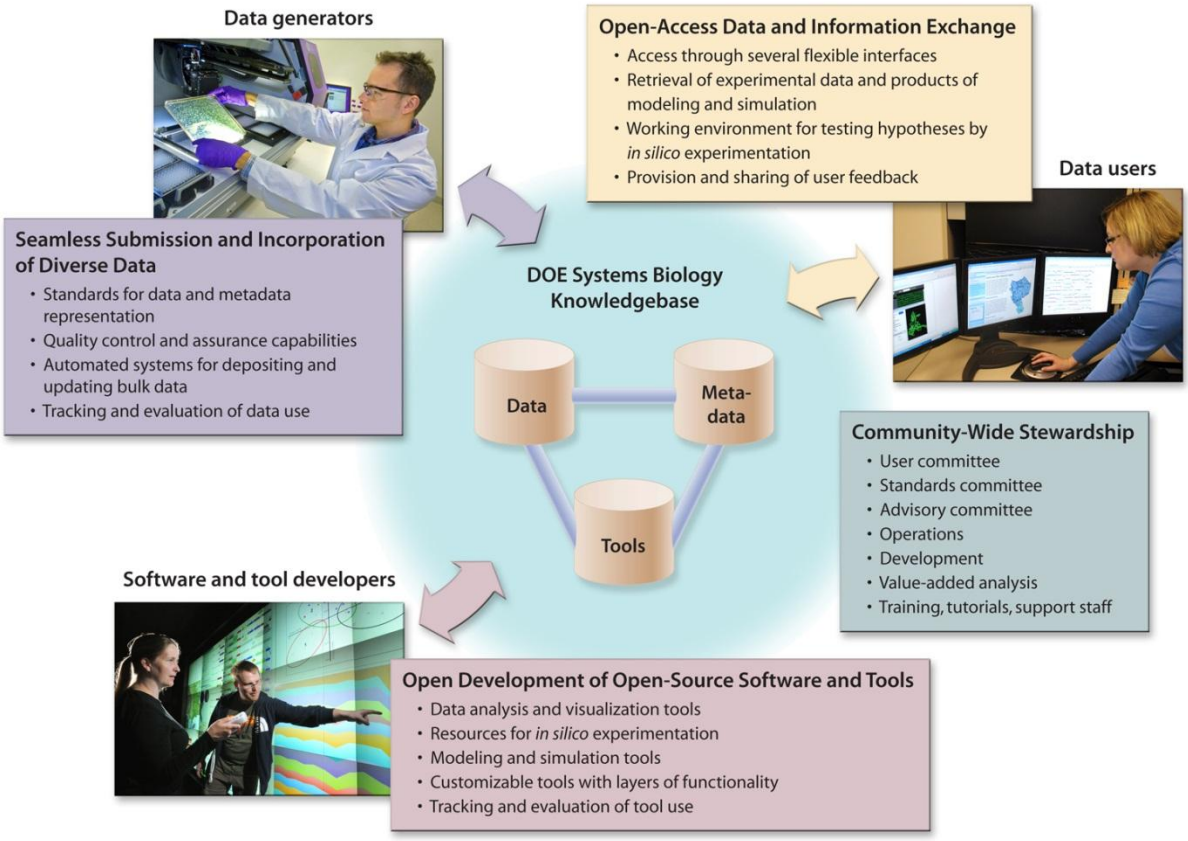


Fig. 1.1. DOE Systems Biology Knowledgebase: Establishing a Systems Biology Framework. The desired attributes and communities needed for a successful Systems Biology Knowledgebase are shown.

To establish Kbase as a community effort, several basic principles need to be considered. One is *open access*—the concept that data and methods contributed to the system will be available for anyone to use. Another is *open source* or open contribution, meaning that source code is managed in an open environment and is freely available to access, modify, and redistribute under the same terms. Perhaps the most important concept is *open development*, which would allow anyone to contribute to Kbase development under organizational guidelines. Analogous to submitting a publication, this would involve a review process by an authoritative group that would determine if a particular contribution meets established criteria. In such an environment, different groups would work together on a common piece of software to meet common needs. The review process would facilitate integration into Kbase and quality control, resulting in a product better than what an individual alone could create.

Several existing systems and applications can serve as reference models for thinking about Kbase development. Exemplifying the concept of an open-source environment for development is the computer operating system Linux, which is being built by a community of software developers working collaboratively to create a sophisticated and fairly successful system. Other

familiar examples include iPhone or Google apps that enable users to choose the kinds of features and capabilities they want and then easily integrate these functions into a phone or other device. Learning from user interfaces that show layering of data from Google Maps and Google Earth annotations (e.g., locations of landmarks and restaurants), similar interfaces could be developed for designing experiments and annotating data and research results.

Another example of open source and open development is Wikipedia, which allows individuals or groups to contribute content. Wikipedia has an editorial model, and the quality of its content improves over time. The open-development environment envisioned for Kbase would engage the community and enable everyone, not just computing experts, to play a more active role in Kbase development and evolution.

Standards for usability, understandability, discovery, and contribution also are important to Kbase. The design of Kbase should be intuitive so that researchers can use it with minimal training, and the system's components need to be understandable to users. Understandability implies that there is a good foundational basis for knowing that a result returned to a user is based on robust scientific assumptions and that these assumptions are clear. If results are not understandable, the system should allow the user to drill down to acquire additional information about how results were obtained. Kbase also should promote an environment of discovery, leading to new rounds of experiments or lines of research. Finally, engaging the entire research community in Kbase is critical because not all researchers today have comprehensive access to major computational capabilities. Democratizing access to data, analytical software, and modeling tools via Kbase would accelerate scientific discovery and lead to important innovations in energy and environmental research. Any system being used by scientists ultimately should be measured by how well it demonstrates these concepts of openness and usability, advances research, and supports the scientific method.

Within Kbase, the needs of the community should be balanced with the needs of individual researchers. Thus, some level of individual or team research privacy is required and could be achieved with user accounts. Prior to data release upon publication, data and code could be held in private and analyses conducted in a nonpublic environment. Kbase also will need to allow users to assess data quality, archive experimental protocols, and track version history and provenance so that new analyses can be usefully compared against previous work.

Throughout the development of this implementation plan, a clear consensus was to design achievable Kbase objectives and show scientific and technical success in the near term rather than trying to design and build the ultimate system to serve every research need. In contrast to past bioinformatic efforts, Kbase will continuously expand and adapt to meet the evolving needs of its core objectives while integrating and adding value to the information and tools resulting from this research. This concept supports the goals for open and modular design. Kbase will be a software engineering effort unlike any other project undertaken for the systems biology community. As such, it demands engaging the stakeholders to identify requirements and define success. Such engagement is evident in community discussions of Kbase scientific objectives and endpoints that could be achieved in the near-, mid-, and long-term. Success for Kbase will be as much about scientific accomplishment and community engagement as technological achievement.

1.4 Community Interactions and Input

Developing a successful open-informatics endeavor for DOE systems biology will require key input and skills from several groups within the scientific community. Broadly these groups represent plant and microbial researchers who design experiments and generate data; computational biologists and bioinformaticians who will interpret and simulate data; and computer scientists, database developers, and software engineers who will develop Kbase infrastructure. Representatives from these communities participated in the five Kbase workshops. In addition to contributing to this implementation plan, workshop participants also addressed the cultural transition the informatics community will need to make from individual project-based efforts toward research community-based informatics.

The workshops and the targeted communities were:

Using Clouds for Parallel Computations in Systems Biology. Held at the Supercomputing (SC09) conference on November 16, 2009, this Kbase workshop focused on applications of cloud computing. It brought together researchers in the computing, systems biology, bioinformatics, and computational biology fields. Modern genomics studies use many high-throughput instruments that generate prodigious amounts of data. For example, a single run on a current sequencing instrument generates 30–40 gigabytes of sequence data. The situation is complicated further by the democratization of sequencing; many small centers now can independently create large sequence datasets. Moreover, the immense amount and variety of omics data that must be integrated with genomics data to model and study organisms at a systems level create unique opportunities in computational biology. Consequently, the rate of sequence and related data production is growing faster than our ability to analyze these data. Cloud computing provides an appealing possibility for on-demand access to computing resources. Many computations can be considered embarrassingly parallel and should be ideally suited for cloud computing. However, challenging issues remain, including data transfer and local data availability on the cloud nodes. In discussing the feasibility of using cloud computing for Kbase, clear needs included flexible architecture and input/output (I/O), high-quality reference data and standards, and prioritized workflows.

Plant Genomics Knowledgebase Workshop. Held January 8, 2010, in conjunction with the Plant and Animal Genome XVIII conference in San Diego, California, this workshop was jointly convened by DOE BER and the U.S. Department of Agriculture National Institute of Food and Agriculture. It brought together 100 plant scientists, geneticists, breeders, and bioinformatic specialists to discuss current issues facing plant breeders in light of ever-increasing amounts of genomic data. The workshop featured lectures by leaders in the plant breeding, genomics, and bioinformatics communities. These presentations set the stage for afternoon breakout discussions by addressing the data needs of more-applied breeding programs and describing resources emanating from more-fundamental plant genomics and bioinformatics research. The overarching question was, “How can we best design the Knowledgebase to have the flexibility to grow with and adapt to new data and information challenges in the future?” A key objective was to specifically identify the requirements for effectively developing data capabilities for systems biology as applied to plants, particularly the research and development of plant feedstocks for biofuels. The current state of plant informatics is represented by many disparate

databases primarily focusing on specific taxonomic groups or processes. To enable a systems biology approach to plant research, integrating all types of data (including molecular, morphological, and omics) for bioenergy-relevant plant species is important. Thus, a challenge for Kbase will be to develop uniformity of data format and database architectures to effectively integrate diverse data types and enable user-friendly acquisition and analysis.

DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop. This meeting, held February 9–10, 2010, was part of the DOE Office of Science 2010 Genomic Science Awardee Workshop VIII and Knowledgebase Workshop in Crystal City, Virginia. Workshop participants discussed the current, near-, and long-term prospects for microbial systems biology research in the context of the Knowledgebase. The rapidity with which new genome sequence information appears in public databases is presenting a growing challenge for the data storage, analysis, and utilization necessary to foster scientific and technological advances. The systems biology framework has arisen in response to this challenge, but new computing strategies are needed to take advantage of this new context for examining microbial biology. The "monoculture" paradigm has been quite productive and will continue to be at the heart of microbiology. However, monocultures are not representative of how microbial systems exist in nature. To this end, metagenomics has provided a means for examining microbial complexity, but complementary functional information is still needed to understand the "metaphenotype." In biology, a grand challenge is to predict phenotype from genotype. This challenge is complicated in microbes because a significant fraction of microbial genomes interacts with other organisms and not all genes are continuously expressed. The scientific community is relatively well developed in terms of measuring various types of omics data, but challenges remain for highly complex environments, such as soil and sediments. In the long term, Kbase will be faced with capturing and interrelating data about all these processes at scales from molecules to meters. Several workflows were initiated at this workshop that have been further refined and incorporated in this implementation plan. These include Microbial Scientific Objective 1: Reconstruct and Predict Metabolic Network to Manipulate Microbial Function and Microbial Scientific Objective 2: Define Microbial Gene Expression Regulatory Networks.

DOE Systems Biology Knowledgebase Workshop at the 5th Annual DOE Joint Genome Institute (JGI) User Meeting. The focus of this Kbase workshop, held March 23, 2010, was to discuss scientific objectives and challenges for data handling and knowledge integration specific to the study of microbial communities or metagenomes. Some topics also were pertinent to all development and initial implementation of knowledgebases for the broader biological community. A main workshop theme was to discuss Kbase as a project that would build on existing systems for managing and analyzing omics data while achieving a higher level of support for the scientific community. Several objectives and workflows were initiated at this meeting.

Knowledgebase System Development Workshop. The final of five workshops, this meeting was held June 1–3, 2010, in Crystal City, Virginia. To define detailed requirements for initial priorities, a robust design, and implementation plans to create Kbase, this workshop involved 80 participants representing university, national laboratory, and international scientists, as well as key stakeholders (plant and microbial genomic researchers, bioinformaticians, computer scientists, database developers, and software engineers). Workshop participants also included

representatives from the DOE JGI; DOE's Bioenergy Research Centers; the National Science Foundation's (NSF) iPlant; and the National Institutes of Health's (NIH) National Cancer Institute and National Center for Biotechnology Information. Emphasis was placed on prioritizing clear scientific objectives and specifying the associated tasks and requirements for achieving these objectives. Participants were charged with developing and prioritizing three to five scientific objectives in three areas: microbial, metacommunity, and plant research. Extensive pre-meeting conference calls helped lay the groundwork for workshop participants to develop scientific requirements, time frames, and the level of effort expected for Kbase support of each objective. Once finalized, the requirements were translated into implementation plans for each objective. Workshop discussions also addressed system architecture and governance for the initial system, however, participants were not charged with defining funding or contractual structures. A consensus among participants was that initial Kbase efforts cannot be all things for all users. Showing strong success in a few areas is better than making minimal progress in many areas. Workshop participants also expressed continued support for Kbase principles identified at previous workshops: (1) science drives Kbase development; (2) the project should be a community effort; (3) Kbase should support open access and open contribution; and (4) Kbase resources and capabilities should be distributed. In addition to defining scientific objectives, the systems biology community also articulated the need to define research workflows that enable scientists to compare and contrast different methods. This was deemed a necessary component of the implementation plan, because workflows will form a basis for researcher interactions within the Kbase. The following section describes how the use of workflows helps define the scientific objectives.

1.5 Workflows: Bridging Scientific Objectives from Bench to Computer

In research, a scientific objective is satisfied by creating hypotheses and conducting one or more experiments depending on the scope of the objective. For every experiment, there are rationales, protocols to be executed, a number of data inputs (data sources) and outputs (results), and analysis tools. Workflows describe this information. They are sequential procedures that describe the envisioned steps to answer questions. Workflows are the bioinformatic equivalent of an experimental protocol. Detailed workflows form the bridge between experimental research and computing communities and thus are key to translating research objectives into computing requirements that will most effectively advance the science.

Six near-term, high-priority scientific objectives were selected at the June 1–3, 2010, workshop for this implementation plan. Workflows were developed for these and several other longer-term objectives. Once the initial phase of Kbase is complete, the longer-term objectives and workflows will be developed more completely for implementation (see Chapter 5, Mid-Term Science and Leveraged Annotation Needs and the individual workshop reports in Appendix D). From these workflows and the underlying objectives, the requirements could be defined that lead to the specification of an implementation plan with tasks and scope to achieve these scientific and technical goals. These workflows represent diverse problem-solving methodologies representative of the broad scientific community (see Fig. 1.2 Knowledgebase R&D Project, next page).

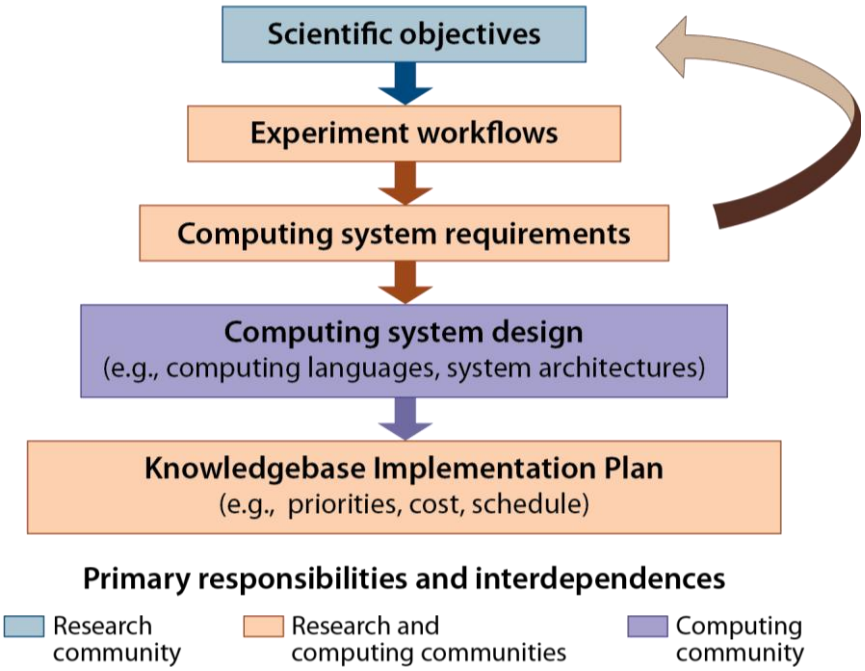


Fig. 1.2. Knowledgebase R&D Project: Scientific Objectives and Collaborations Critical to a Successful Knowledgebase Implementation Plan. The final product of this Knowledgebase R&D Project, the Knowledgebase Implementation Plan, specifies the components and functionality necessary for the systems biology research community to meet their defined scientific objectives. To do this, the research and computing communities must work closely together to define—realistically and at a significant level of detail—the scientific objectives and experimental workflows (protocols) necessary for defining computing system requirements and design and for completing the implementation plan for a robust, durable Knowledgebase.

Workflows provide important details for Kbase design, both in terms of the underlying data as well as the experimental or analytical objective. Kbase architecture will have layers including data repositories, workflow management, and output visualization, all of which relate to workflows developed by the scientific community participating in this Kbase development process. Workflows are essentially communication mechanisms that exchange ideas and information between the researchers and those who actually build the computing system.

Developing an executable Knowledgebase Implementation Plan must be a community effort—from both the experimental and computing research communities—where we integrate across projects and research laboratories. Fully developed, robust workflows will foster this integration and lead to a more standardized approach.

1.6 Report Structure

The DOE Systems Biology Knowledgebase Implementation Plan is the culmination of a year-long effort to engage the scientific community and to develop a collaborative effort between experimental biologists, computational biologists, and computer scientists. The resulting document outlines this effort to develop scientific objectives, prioritize these objectives, and generate an implementation plan, both for the science as well as for the necessary infrastructure. Each chapter contributes to this plan in a unique way, starting with the scientific objectives, the architecture, and the interface between hardware and science. The end of the implementation plan outlines a plan for governance and project management.

In Chapters 2–4, summaries are given of six high-priority scientific objectives and related requirements for the three science areas: Microbes, Plants, and Metacommunities (see Table 1.1). Each chapter corresponds to one of the science areas and addresses two scientific objectives. Each summary is followed by its implementation plan that lists the development and deployment tasks necessary to create, adapt, and test the objective as a part of the growing and integrated Knowledgebase. These detailed implementation plans also describe the duration of tasks and hardware needed. For each objective, near-term tasks, subtasks, and associated staffing resources are summarized in tables. The types of effort are estimated in the broad categories of computational biology research, software engineering, data management, information technology, data curation, and experimentalist advising. More detailed versions of the objectives and requirements in these chapters are provided in Appendices A–C, which describe the objectives' goal, purpose, background, benefits, data sources, inputs, outputs, user interactions, and workflows, along with other relevant information.

Chapter 5, Mid-Term Science and Leveraged Annotation Needs, provides summaries of five additional prioritized scientific objectives and requirements that can begin to be implemented within the next 5 years. These topics were not developed into full implementation plans. Some of the objectives will be leveraged through annotation efforts coordinated at the DOE JGI. Implementation plans for additional objectives can be developed at a later time.

Chapter 6 discusses Kbase relationships with existing or new resources and entities, including extreme-scale computing efforts within DOE Office of Science, the DOE JGI, iPlant, and NCBI.

Chapter 7, System Architecture, describes architectural attributes—such as integration and interoperability—within the planned Kbase. This section also identifies existing hardware and software that could support Kbase deployment and gives recommendations for the project's initial architectural and hardware requirements. Because a federated architecture is recommended for Kbase, the system will need to include computing capabilities and data that incorporate both external resources and those owned by Kbase (also potentially federated).

Chapter 8, Kbase Infrastructure Tasks and Timelines, describes the tasks, timelines, milestones, deliverables, and plan for implementing the underlying infrastructure for Kbase. This plan—essential for building Kbase beyond the six initial projects—provides the structure for adding future projects and tools. Key elements include interfaces, hardware, design, and operational requirements associated with Kbase infrastructure and maintenance. This effort will deploy

Table 1.1. Six Near-Term Science Needs Supported by Kbase			
Section	**Science Area**	**Scientific Objective**	**Priority**
2.1	Microbial	Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function	High
2.2	Microbial	Define Microbial Gene Expression Regulatory Networks	High
3.1	Plant	Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype	High
3.2	Plant	Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Analysis, Comparisons, and Modeling	High
4.1	Metacommunities	Model Metabolic Processes within Microbial Communities	High
4.2	Metacommunities	Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses About Their Function	High

application programming interfaces (APIs), along with data and tool registries, and will support multiple programming tools and web-based protocols.

Based on workshop consensus, Chapter 9 describes the underlying governance principles recommended for Kbase. This chapter also calls for the formation of a governance body to function as a representative of the scientific community in developing policies and standards and providing advice and feedback to DOE. The underlying principles will be drawn from the community consensus and the ongoing articulation of policies and standards. This will be driven by the governance principles of open access, open source, and federation. In these recommended approaches, the individual tools, datasets, and objectives must be designed from the start with the ultimate goals of consolidation and incorporation in mind. Several initial areas requiring establishment of policies are described, such as data release and embargoes.

Chapter 10, Project Management, provides a brief recommendation on structuring the organization of this federated project.

To succeed, Kbase must be valued by the research community and driven by focused scientific objectives with targeted goals for assessing progress and accomplishments. Although it is easy to build technology for its own sake, focusing on community-defined objectives ensures strong community "buy-in." This implementation plan was developed based on interactions among the experimental systems biology, bioinformatics, computational biology, and computer science communities working together to determine the goals for defining success. An ongoing outreach activity for the project will be providing incentives for continued community participation in developing and improving Kbase. In some ways, establishing a community cyberinfrastructure such as Kbase represents a cultural change needed to transition biology from a focus on individual project-based efforts to an open community science.

2. Near-Term Microbial Science Needs Supported by Kbase

In the microbial science area, the first objective is to improve the accuracy of metabolic network models, especially for microbes important in biofuel production and environmental remediation, so metabolic engineering produces more predictable results. The second objective is to enable automated inference of gene regulatory networks based on data from gene expression profiling. Predicted networks then would be validated to determine their accuracy and refined to improve prediction of cellular behavior and fitness. Both objectives have tasks in developing data repositories and workflows that link into the Kbase infrastructure.

Microbial Scientific Objective 1

2.1 Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function

Summary of Objective and its Requirements

Relevance

The scientific community seeks to understand and manipulate the metabolic potential of organisms using validated metabolic models. More specifically, this effort involves reconstructing metabolic networks, predicting organisms' growth phenotypes from their metabolic networks, understanding organisms' metabolic potential, providing scientists with software tools to interrogate and interactively visualize metabolic networks, and enabling engineers to quickly determine the strategies necessary to remodel metabolism for specific purposes. The goals are to move beyond the current state of the art to increase the speed and automation with which metabolic networks can be reconstructed and to improve the accuracy of metabolic network predictions. This knowledge will lead to the informed modification of one or more specific enzymes or the introduction of entirely new enzymes and pathways, allowing the scientific community to determine better strategies for manipulating mass or energy flow in microorganisms.

Objective

Microbial Scientific Objective 1 is to accurately evaluate an organism's metabolic potential; predict the phenotypic outcome of specific metabolic or environmental interventions or perturbations; and establish metabolic kinetics, capabilities, and fluxes for short-term dynamic responses. Achieving this objective requires integrating new experimental data with existing data and models on metabolic pathways and developing methods to automatically create new metabolic reconstructions from newly sequenced organisms. This objective is a high priority when applied to a select set of organisms relevant to DOE's current research efforts; for many other microbes, it is a medium priority.

The DOE Systems Biology Knowledgebase (Kbase) should provide access to a variety of data. Such data include metabolic maps (both stoichiometric and regulatory); enzyme concentration and activity levels; qualitative data on enzyme regulation and known substrate, product, and cofactor dependencies; enzyme kinetic data (if available); suggested kinetic rate laws or

reasonable approximations; metabolic flux maps (predicted or measured) and metabolite levels; sensitivity data such as rate limitations and control coefficients (if available); time-course data on changes in metabolites or enzyme concentrations; and relevant thermodynamic data (computed or measured) on individual metabolic reactions. This objective requires linking known metabolic models with experimental data and databases such as Chemical Entities of Biological Interest (ChEBI), Universal Protein Resource (UniProt), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Ontology (GO) as well as user-generated data.

Potential Benefits

Metabolism is the end point for many biological applications of interest to the U.S. Department of Energy (DOE). DOE researchers must have access to reliable and comprehensive tools to evaluate data and predict phenotypes. Given DOE's interest in metabolic engineering for biofuel production and environmental remediation, which requires detailed knowledge of metabolic dynamics, this objective is a high priority. Current research and development in metabolic networks primarily involve two approaches. The first is evaluating novel microbes to identify and improve desired metabolic phenotypes (e.g., recent work on *Clostridium phytofermentans* or *Caldicellulosiruptor*). The second is manipulating the metabolic pathways of well-characterized microbes to enable novel functionality (e.g., initiatives to engineer cyanobacteria for photosynthetic production of alkanes and isoprenoids and recent achievements in hydrocarbon production from *Escherichia coli* or cellulose expression in *Saccharomyces cerevesiae*). This objective benefits both approaches.

Synergies with Other Projects and Funding Agencies

This scientific objective will build on the three main sources of online metabolic data: Encyclopedia of Metabolic Pathways (MetaCyc; www.metacyc.org), Kyoto Encyclopedia of Genes and Genomes (KEGG; www.genome.jp/kegg/), and Braunschweig Enzyme Database (BRENDA; www.brenda-enzymes.org). The current range of data sources is scattered, not always easy to use, and lacks important information. Repositories such as MetaCyc could be modified and new, third-party tools developed to enable more seamless access to data. This effort also should build on current genome-based, curated metabolic reconstructions. Kbase could leverage other DOE-relevant metabolic databases including the *Shewanella* Knowledgebase (from the *Shewanella* Federation), BeoCyc (a database of 33 bioenergy-related organisms from the DOE BioEnergy Science Center), PlantCyc (metabolic database for *Arabidopsis* and poplar from the Carnegie Institution), FungiCyc (from the Broad Institute), and YeastCyc (from Stanford). Although many of these data are of much higher quality, they too are scattered and stored in a number of different, conflicting, and sometimes undocumented formats. The development of agreed-upon standards for storing flux-balance information will be required.

No concerted effort has been made to collect and curate quantitative data, enzyme levels, and time-course data. Kbase support of Microbial Scientific Objective 1 would have very little to no overlap with existing projects such as iPlant (www.iplantcollaborative.org), GenBank (www.ncbi.nlm.nih.gov/genbank/), or other efforts by the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov). Experimental projects within the Office of Biological

and Environmental Research (BER) that seek to alter metabolic pathways for various DOE missions would be leveraged as "first adopters" and serve as beta testers for this Kbase objective. Since validation is critical to testing and developing tools and data sources, these linkages are mentioned in workflows described below and in Appendix A, Supporting Scientific Objective and Software Requirement Documents for Near-Term Microbial Science Needs. Likewise, this scientific objective has clear linkages to others identified for Kbase, including Define Microbial Gene Expression Regulatory Networks (see Section 2.2) and Model Metabolic Processes within Microbial Communities (see Section 4.1).

Illustrative Workflow

This objective has a number of workflows with various intermediate goals and timelines described in Appendix A. One workflow example, which is illustrated in Fig. 2.1 on the next page, includes:

- Generation of automatic genomic annotations for *automated* inference of a draft metabolic network.
- A reconstruction and simulation engine that *automatically* generates a list of gaps (e.g., missing enzymes or transporters) and inconsistencies (e.g, functions without context or "dangling" compounds). Such a list by itself is of huge scientific value because it points scientists to open research problems, missing knowledge, and important experiments.
- Existing and newly developed software tools that attempt to fill in gaps and impose consistency on annotations (e.g., negate "weak" functional assignments not supported by the functional context).
- A modified set of annotations, as well as additional assumptions about boundary conditions and pathways (e.g., based on experimental physiological data), used to guide hypotheses and experimental designs.
- Incorporation of experimental data to validate or disprove parts of the metabolic network.

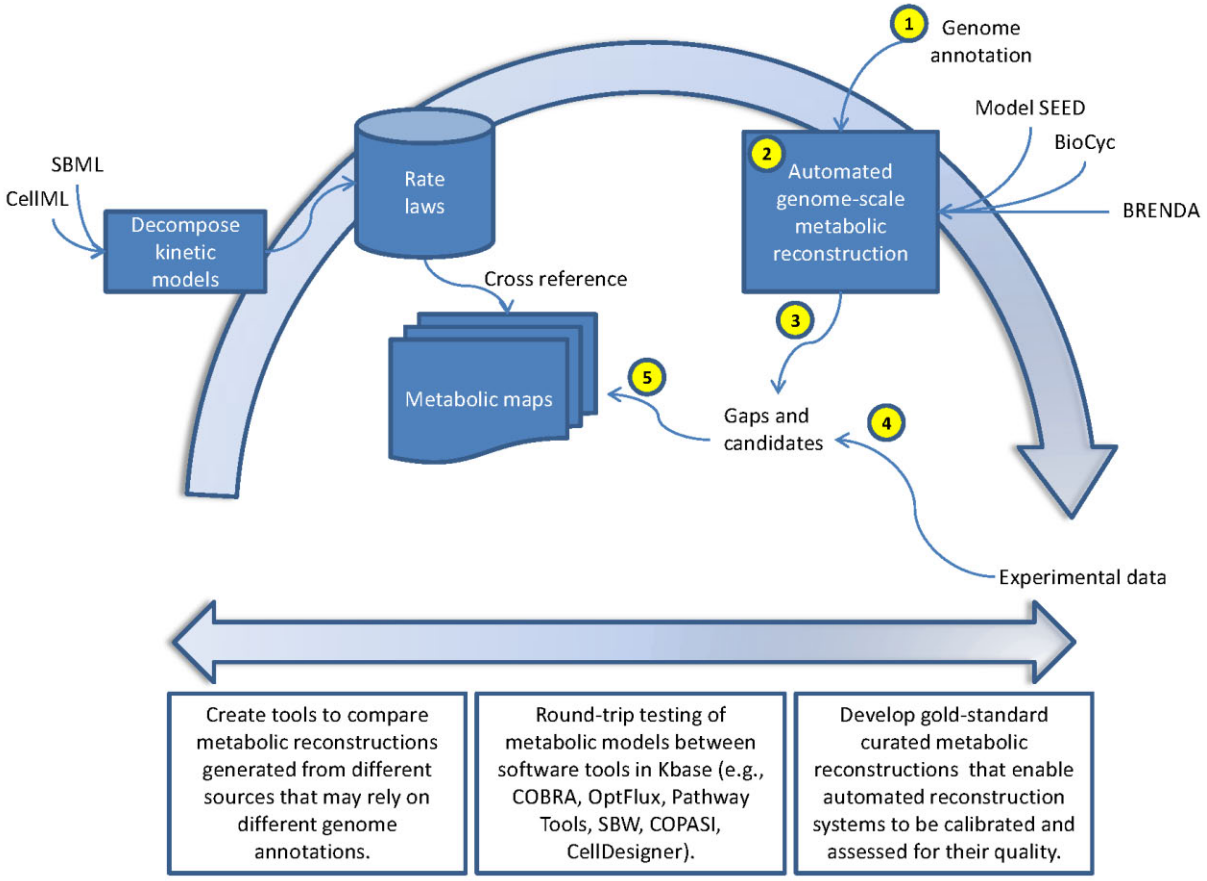


Fig. 2.1. Workflow for Reconstructing and Predicting Metabolic Networks. Based on a microbial genome annotation (1), generate a metabolic reconstruction (2) and an associated list of gaps and inconsistencies (3) that are then checked experimentally (4) and verified and corrected (5) to improve the database.

Implementation Plan for Reconstructing and Predicting Metabolic Networks to Manipulate Microbial Function

System Capabilities

The envisioned Kbase system will involve a number of interoperating metabolic databases and software tools for manipulating descriptions of metabolic networks in pursuit of scientific objectives. These objectives include evaluating the metabolic potential of an organism; predicting the phenotypic outcome of specific metabolic or environmental interventions; and developing quantitative, validated metabolic models.

Kbase must provide tools for capturing and updating such models, for reconstructing models rapidly from genomic data, and for performing a variety of analyses and comparisons with the models. Therefore, exchange of metabolic data among multiple databases and software platforms is essential.

Leveraging and harmonizing software and databases extant in this area are important. These include the constraint-based reconstruction and analysis (COBRA) toolbox, Pathway Tools, and other resources listed under Task 4: Interoperations and Standards.

Tasks

Implementation of the following components is needed to realize the preceding capabilities.

Task 1. Databases.

1A. *Create a repository of growth data for organisms of importance to DOE in validating growth-prediction algorithms*.

Accurate metabolic modeling depends on a standard set of experimental data on key DOE-relevant organisms. This task involves identifying such organisms and establishing a standard set of experimental data and metadata (e.g., media composition, temperature, pH) that properly account for the experimental design and parameters of the model.

Based on this data design, the associated Kbase data infrastructure would be built and computer methods for uploading and linking data sources to Kbase for the associated experimental data and metadata would be established. Beyond establishing the data representation, this task would also involve the curatorial activity of compiling existing data into the standard representation.

1B. *Create a repository of metabolic flux data*.

This task includes identifying first adopter experimentalists and establishing collaborative relationships so that their laboratories provide data and design advice and are beta testers. These would be chosen from separately funded relevant projects.

1C. *Develop gold-standard, manually curated metabolic reconstructions for approximately 20 organisms important to the DOE mission.*

These reconstructions will serve as important resources and will enable automated reconstruction systems to be calibrated and assessed for their quality. In many cases, existing efforts funded by other organizations [e.g., the National Institutes of Health (NIH) and the National Science Foundation (NSF)] should be leveraged.

The resources required for each reconstruction will vary depending on the complexity of the organism and the amount of information available in the literature.

Task 2. Software.

2A. *Improve fully automated metabolic reconstruction systems.*

- Improve the speed of these systems.
- Improve the accuracy of these systems.
- Improve the comprehensiveness of these systems by allowing them to automatically generate aspects they currently cannot, such as flux-balance models that are close to operational.

2B. *Develop methods to integrate metabolic and regulatory models and automate their refinement.*

Using the previously developed gold-standard metabolic reconstructions, develop integrated metabolic and regulatory models, which will leverage regulatory network reconstructions arising from other Kbase efforts.

2C. *Evaluate existing tools and methods for automated design of pathways for metabolic engineering.*

Adopt one or a few of these that are consistent with or could be extended to allow a graphical, workbench-style user interface for design and that also follow data standards needed for Kbase interoperability.

2D. *Create tools for comparing metabolic models with simulation results and with experimentally determined fluxes*.

Create tools to compare metabolic reconstructions generated from different sources that may rely on different genome annotations. For example, do BioCyc and Model SEED agree or disagree on the presence of reactions and enzymes in a given organism?

2E. *Create tools for predicting rate-limiting steps within metabolic networks.*

For example, examine existing software for carbon 13 isotopic flux prediction (e.g., FiatFlux) and improve it to enable better predictions for fluxes through all pathways in the cell, not just central metabolic fluxes. These software tools require metabolic network reconstructions, atom mappings between substrates and products, and experimental measurements (^{13}C labeling distributions on metabolites, biomass composition, and cellular uptake and secretion rates). The tools should provide estimates for intracellular fluxes (net and exchange fluxes) and confidence intervals for these estimates.

Develop methods for determining metabolic fluxes and their confidence intervals based on time-dependent carbon 13 isotope measurements as a function of time after carbon 13 addition (before an isotopic steady-state has been reached).

Computationally (theoretically) predict the distribution of the *degree* of rate-limitingness in metabolic pathways under different conditions and in relation to the activity of negative and positive feedback loops. This would complement the flux, enzyme activity, and kinetic data and also be related to the validation procedures outlined in 3F below (see Level 4 validation).

Task 3. Applications.

3A. *Convert into Systems Biology Markup Language (SBML) all flux balance models currently unavailable in this format.*

Automatically generate genome-scale metabolic reconstructions for DOE-relevant organisms and make them available in SBML format. This could involve a combination of existing reconstructions from BioCyc (670 models to date), Model SEED (130 reconstructions to date), and others generated from various methods (Palsson) and could include aspects of model generation not currently automated.

3B. *Convert stoichiometric maps into SBML format.*

Convert and store many constraint-based models and stoichiometric maps into standard formats such as annotated SBML and the SBML Flux Balance Analysis (FBA) extension (Bergmann and Olivier 2010). Since many tools already read SBML, it would be a natural format to use. Conversion to other formats (e.g., Matlab, COBRA, and OptFlux) can be easily achieved. It is already possible, for example, to convert COBRA format to SBML (using the Python Simulator for Cellular Systems, or PySCeS). An agreed and clear definition of "stoichiometric map" is needed.

3C. *Decompose the hundreds of existing microbial SBML and CellML kinetic models into individual reaction steps and rate laws.*

This will provide a database of published rate laws that could be used in future models. The data should be cross-referenced to metabolic maps. For example, by selecting a reaction on a metabolic map, a user will be provided with all published rate laws associated with that reaction step.

3D. *Provide better access to an online metabolic regulatory map.*

These data would include all modifiers that affect enzymes, both activators and inhibitors. At the simplest level, modifiers for each enzyme could be listed and the data expanded later to include mechanisms (e.g., allosteric or covalent modification) and possibly information on K_is, Hill coefficients, and proposed rate laws.

3E. *Integrate gene functional annotations and genome-scale metabolic reconstruction and simulation capabilities within the Kbase environment.*

This will enable iterative improvement of both layers of information. An example workflow is described in the Illustrative Workflow section and Fig. 2.1 above.

3F. *Validate metabolic models at five successively harder levels.*

Five suggested levels for validating a model are proposed, with each level more demanding than the previous. The first level of validation and possibly the easiest to achieve involves comparing growth or no-growth phenotypes for wildtype and mutant strains. Related to this is the comparison of flux balance analysis predictions with isotopic flux measurements to further validate the flux balance models. In the next level, predicted steady-state flux and metabolite levels are compared against experimentally measured fluxes and metabolites. Level 4 validation will test the ability of the model to predict the effect of "small" perturbations in enzyme activity levels and environmental conditions. Finally, the most demanding validation test in this sequence involves comparing time-course changes that arise from major environmental changes, such as shifts in nutrients or O_2. Kbase will need to leverage experimental biology efforts to perform the collaborative validation experiments. These leveraged experimental efforts likely will be the first adopters and selected from appropriate BER-funded research to work closely with the Kbase project.

Validation levels:

1. At the level of growth or no-growth predictions.
2. Compare flux balance predictions against isotopic flux measurements.
3. Compare predicted steady-state metabolite concentrations and fluxes to experimentally measured values.
4. Perturb enzyme levels by specified amounts and recompute the resulting fluxes and metabolite changes.
5. Time-course validation.

Task 4. Interoperation and standards.

4A. *Exchange and align metabolic models.*

Fostering the exchange of metabolic models between platforms (e.g., Pathway Tools, Palsson, KEGG, and Model SEED) is desirable to facilitate comparison and application of models developed under different platforms. Here is an example of what could be done.

- Build SBML importer for Pathway Tools.
- Build SBML importer for Palsson platform.
- Build Pathway Tools module to align Palsson model with Pathway Tools model.

- Build module within Palsson platform to align Pathway Tools model with Palsson model.

4B. *Establish round-trip testing of metabolic models between different platforms and software tools*.

Examples include COBRA, OptFlux, Pathway Tools, Systems Biology Workbench (SBW), Complex Pathway Simulator (COPASI), and CellDesigner. This would involve a bioinformaticist working across multiple interacting groups.

Resources

Microbial 1: Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function

Table 2.1 Hardware Resources for Microbial 1

Hardware Purpose	Type	Size
Data management	Storage	Terabytes
Data analysis	Processing	Large (more than 1000 cores)

Microbial 1: Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function

Table 2.2 Staffing Resources for Microbial 1

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
1. Databases		
1A. Create a repository of growth data for organisms of importance to DOE in validating growth-prediction algorithms.	B, Bfx	1–36
1B. Create a repository of metabolic flux data.	B, SE	1–36
1C. Develop gold-standard, manually curated metabolic reconstructions for approximately 20 organisms important to the DOE mission.	B, Bfx	12–60
2. Software		
2A. Improve fully automated metabolic reconstruction systems.	SE, Bfx	1–48

Microbial 1: Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function

Table 2.2 Staffing Resources for Microbial 1

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
2B. Develop methods to integrate metabolic and regulatory models and automate their refinement.	SE, Bfx	12–48
2C. Evaluate existing tools and methods for automated design of pathways for metabolic engineering.	SE, Bfx	1–36
2D. Create tools for comparing metabolic models with simulation results and with experimentally determined fluxes.	Bfx	1–24
2E. Create tools for predicting rate-limiting steps within metabolic networks.	Bfx	1–48
3. Applications		
3A. Convert into Systems Biology Markup Language (SBML) all flux balance models currently unavailable in this format.	Bfx	1–12
3B. Convert stoichiometric maps into SBML format.	Bfx	1–12
3C. Decompose the hundreds of existing microbial SBML and CellML kinetic models into individual reaction steps and rate laws.	Bfx	1–36
3D. Provide better access to an online metabolic regulatory map.	SE, Bfx	24–48
3E. Integrate gene functional annotations and genome-scale metabolic reconstruction and simulation capabilities within the Kbase environment.	SE, Bfx	24–60
3F. Validate metabolic models at five successively harder levels. (Leverage separate experimental efforts.)	Bfx	1–60
4. Interoperation and standards		
4A. Exchange and align metabolic models.	SE, Bfx	12–24
4B. Establish round-trip testing of metabolic models between different platforms and software tools.	Bfx	36–48

System Releases

Release 1 (Year 2). Well-curated metabolic reconstructions exist for 10 additional DOE mission–critical organisms. The reconstructions can be exchanged seamlessly among a variety of software tools within Kbase and can be compared in detail, along with their quantitative predictions. Growth predictions have achieved 85% accuracy. New metabolic reconstructions can be generated and updated for 100 to 1,000 sequenced bacteria in a short period of time.

Release 2 (Year 4). Integrated metabolic and regulatory network models can produce simulations and flux predictions of significantly increased accuracy. Growth predictions have achieved 90% accuracy from manually curated models and 70% accuracy from automatically generated models. Computer-designed metabolic pathways implemented through synthetic biology have exhibited significant flux rates.

Microbial Scientific Objective 2

2.2 Define Microbial Gene Expression Regulatory Networks

Summary of Objective and its Requirements

Relevance

In response to varying and competitive environments, microbes must deploy the products of diverse gene sets to survive and prosper. Expression of the correct sets of genes at the correct levels could confer the best competitive advantage, given the organism's genetic complement and the current environment. The alternative is to starve, be destroyed by the environment, or be outgrown or directly killed by other microbes. The networks of interactions within and among microbes in a given community define the capabilities for more or less stable or inducible biotransformation of the environment. These interactions also determine microbes' ability to remediate environments, improve growth of energy crops, process biomass into fuels, and sequester carbon, among other things. The mechanisms within cells that sense the environment and compute which gene sets should be deployed at what levels, thereby coordinating different stages of the microbe's growth and development, are collectively called the gene regulatory network. Knowledge of this network is the foundation for predicting, controlling, and designing the behaviors of microbes and their community.

Objective

This scientific objective can be divided into two broad components. The first is to enable automated inference of gene expression regulatory networks, relying principally on expression profiling data. The second is to extend these inferred networks to include additional data types to refine network predictions and test them. The availability and evolution of genome-scale expression data and its rapid extension into new data types (e.g., proteomics and transcriptomics) make defining microbial gene expression regulatory networks an attractive goal of the Kbase project. In the near term, the preliminary inference of regulatory networks from just genome sequences and expression profiles under varied cellular conditions will be possible and of general use to researchers in constructing and understanding cellular processes such as carbon and nitrogen cycling. Interconnecting regulatory networks with metabolic reconstructions and multidimensional annotations (two other high-priority objectives identified by the Kbase microbial group and described in Sections 2.1 and 5.4, respectively) would greatly facilitate development of microbial systems biology (Koide et al. 2009).

A variety of phylogenetically diverse microbes would be selected for initial efforts. These should range from well-characterized microbes for which extensive data exist, enabling the most informed analyses [e.g., *E. coli*, *Shewanella oneidensis*, *Geobacter sulfurreducens*, *Halobacterium salinarum*, *Synechococcus* (a cyanobacterium), and *Dracunculus vulgaris*] to those less well characterized (e.g., *Zymomonas mobilis* or *Clostridium thermocellum*) to those for which little information exists. Priority should be given to organisms key to DOE missions, with a focus on regulatory paradigms of greatest relevance to the microbe in question. Understanding O_2 and carbon regulation was identified as one important initial focus.

Potential Benefits

Some near-term goals can be achieved by pursuing this high-priority objective, but completing various valuable stages of this effort may take 2 to 10 years.

The advent of genomic technology and the availability of many microbial genomes have permitted the development of technologies to accelerate these careful studies and provide data from which regulatory networks and their behaviors may be *inferred* rather than directly measured. Comparing regulatory network models against gold-standard determination methods will result in model validation and refinement in the longer term. This, along with an increasing amount of various functional data types, will allow robust correlation of regulatory network predictions to genome features and cellular behavior and fitness.

O_2 regulation of carbon metabolism is a central issue for engineering biofuel-producing microbes. A complete understanding of the regulatory networks that mediate this regulation will allow researchers to specify the patterns and extent to which the expression of different genes turns on as cells are shifted from aerobic to anaerobic growth conditions. Furthermore, gaining complete control over gene regulation during anaerobiosis is essential for optimizing the conversion of reducing equivalents into biofuels. This also may allow efficient production of advanced biofuels like isopentanol or alkanes in anaerobic conditions where loss of reducing equivalents to O_2 can be avoided. (Currently, only fermentation products such as ethanol or butanol can be produced anaerobically with significant yields.) Finally, elucidating the regulatory network by which O_2 influences carbon metabolism is important for the general advancement of science. Until we know the roles and interactions of the different regulatory modalities involved (e.g., repression, activation, small RNAs, and attenuation) and how these networks have evolved among microbial lineages, we will lack understanding of the fundamental components in the evolution of life on Earth. Methods developed to increase our knowledge of a few regulatory factors are expected to be reusable in applications to understand a myriad of other regulatory factors such as temperature, light, salt, and moisture availability.

Synergies with Other Projects and Funding Agencies

This objective could work synergistically with NIH Pathway Tools, EcoCyc, and DOE efforts such as MicrobesOnline and the Joint Genome Institute (JGI). Much of the experimental work would come from DOE's Bioenergy Research Centers (BRCs) and the larger DOE science-focused work on microbial systems. A number of ongoing experimental campaigns were identified that could provide the required data and are listed in Appendix A, along with more details. Given the scale of the problem, these overlaps are more likely to generate synergies than conflicts, provided adequate attention is given to coordinating efforts.

Illustrative Workflow

In generating a regulatory network by inference (a "bottom-up" approach), it is assumed that for the organism of interest, the genome has been completely sequenced and fully annotated. Also assumed is that RNA-Seq or tiling array data are available for a minimum of 10 growth curves with 6 time points and 3 biological replicates on biological conditions relevant to the regulation of O_2 and carbon use.

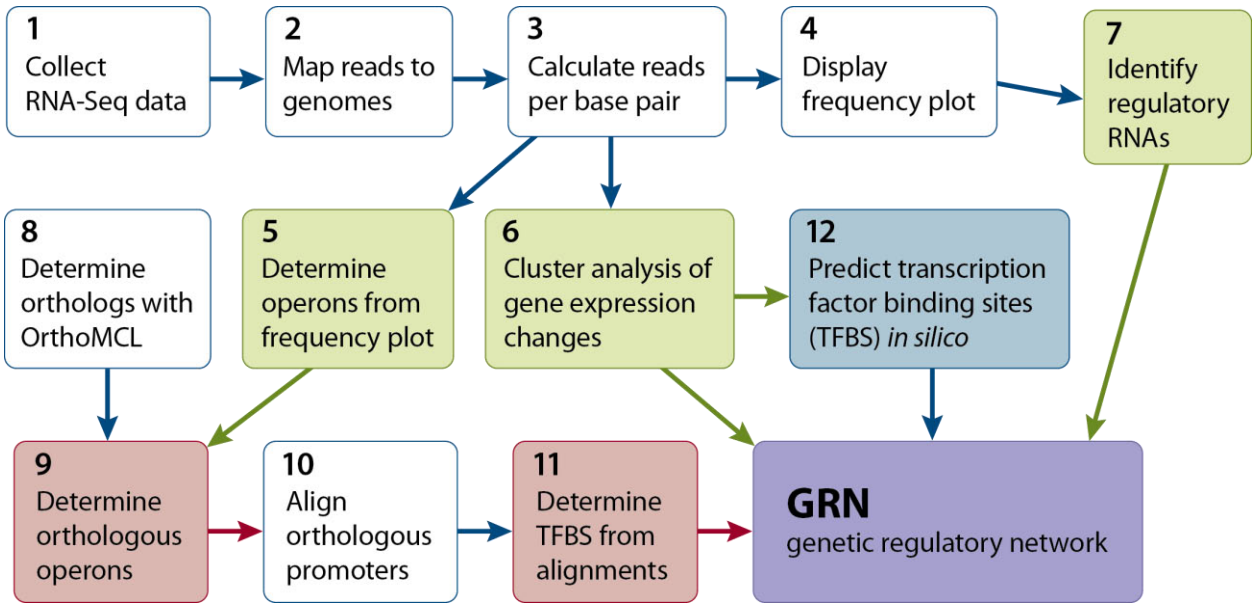


Fig. 2.2. Transcriptome Analysis Pipeline for Gene Regulatory Network Prediction. White boxes are procedures we already know how to do. Green boxes are procedures that have not been determined but are expected to be fairly easy to construct (year 1). Red boxes are procedures that will be more difficult to construct (year 2). The blue box depicts a technique that is optional but would increase analysis accuracy. The purple box is the final product (year 2).

An example of a transcriptome analysis pipeline is shown in Fig. 2.2. Once RNA-Seq data (short sequences) are collected from a particular growth state for a specific species (step 1), preferably keeping the strand information by synthesizing only single-strand cDNA, the short sequences will be mapped back to their associated genome sequence (step 2) and the reads/bp (reads per base pair) will be calculated as a measure of each gene's or operon's expression level (step 3). The reads/bp will be displayed in conjunction with the genome sequence (step 4) using the latest version of Artemis, which already has this capability. Rules will be generated to define operons (step 5) based solely on these data. The output of this analysis will be a list of operons and their expression level for each growth state of every species analyzed. Using OrthoMCL to help define orthologous genes (step 8), orthologous operons will be identified in related genomes (step 9) and used to identify as many orthologous promoters as possible (step 10). Next, the transcription factor binding sites (TFBS) for these promoters will be predicted using two separate techniques. One will involve multiple sequence alignment of the orthologous promoters in an attempt to define the TFBS (step 11) based on their conservation. This technique depends upon the number of sequenced, related genomes and the total genetic distance between all the organisms in each alignment. The average nucleotide identity (ANI) thus will be used to estimate if there will be sufficient sequence divergence in an alignment. If the orthologous operons can be identified in more distant relatives, attempts will be made to expand the alignments. The second technique will use more traditional TFBS prediction algorithms (step 12) such as (Liu et al. 2008) and (Conlan et al. 2005). Results from both techniques will be compared for consistency. Next, cluster analysis will be performed on the differences in gene (operon) expression identified in the RNA-Seq data (step 6). Finally, small

regulatory RNAs will be identified from the frequency plot (step 7), as previously described (Passalacqua et al. 2009; Yoder-Himes et al. 2009). Although not shown, valuable information could be added by sequencing the 5' end of mRNAs via rapid amplification of 5' cDNA ends (5' RACE) and determining TFBS using microfluidic or other assay systems. All of this information will be combined to generate genetic regulatory networks (GRNs) for the studied organisms. Currently, GRNs have been created for only two organisms: *E. coli* (Cho et al. 2009) and *H. salinarum* NRC-1 (Bonneau et al. 2007).

Data and the metadata on experimental design would be automatically parsed from public data, or users could be prompted to upload this information. The user interface should provide options to choose algorithms based on the amount and type of available data. Users also should have access to published citations for the algorithms and basic information on their workings in nontechnical, jargon-free language. Storing a session with default or user-edited settings should be possible so that the entire analysis can be recreated. Advanced users should have privileges to change or override default settings by changing, for example, the source of information or threshold of significance. (For additional workflow details, see Appendix A.) The end result is that users should be able to select an organism; upload, broadcast, or import expression data from public repositories or their own data; and submit a request for network inference.

Implementation Plan for Defining Microbial Gene Expression Regulatory Networks

System Capabilities

This scientific objective can be broadly divided into two major components. The first is to enable automated inference of gene expression regulatory networks relying, principally on molecular expression profiling data and comparative sequence analysis. The second is to extend these inferred networks to include additional data types to refine the network analysis tools.

Kbase would serve as a repository and data integration resource for microbial expression profiles and associated experimental data and metadata. These capabilities will require organizing genome-scale datasets for TFBS distributions, RNA expression profiles and potentially quantitative regulator binding assays, mutant studies, proteomics, and metabolomics. Collecting and integrating these data will drive development of tools for data manipulation, analysis, and visualization that aid microbial systems biology research, both cutting-edge studies and everyday activities in microbiology laboratories.

This effort will coordinate and synergize with tool development projects such as NIH Pathway Tools, EcoCyc, and EcoliHub as well as DOE efforts like MicrobesOnline, JGI Integrated Microbial Genomes (IMG) system, BRCs, and the agency's larger science-focused work on microbial systems.

Tasks

Task 1. Enable automated inference of gene regulatory networks (short term).

1A. *Finalize the definition of regulatory network reconstruction workflow*.

The initial objective concentrates on O_2 and C regulation as an illustrative example. However, providing a broadly applicable tool for generating gene regulatory networks from RNA expression data is the priority. Selecting specific microbes and networks is beyond the scope of Kbase and this plan, but the potential selections are assumed to be high-priority, high-value DOE mission projects with multiple related sequenced genomes. Here we describe the Kbase capabilities that would be applicable to any microbial network.

The assumption is that (1) a complete annotation of the finished genome sequence exists and that (2) the analysis is based on strand-specific transcript profiles with, for example, high-density tiling or RNA-Seq data from multiple growth conditions, such as varying O_2 tension with different sugar carbon sources.

Several network reconstruction approaches are described in the Software Requirements document in Section A.4 in Appendix A. Given a specific set of planned experiments, the initial implementation would be based on combining the best of these approaches as applicable. Current algorithmic approaches to network inference generally rely on a well-documented, quality-controlled compendium of expression data (usually RNA expression from various microarray, high-throughput qPCR–like methods or sequencing methods such as RNA-Seq). Many of these algorithms can use or require (1) known interactions measured through direct means such as ChIP-chip or gel-shift, (2) known or sequence-analysis-predicted cis-regulatory sequences, and (3) other information such as gene-neighbor scores or common functional class annotations. To implement current best-practice workflows, Kbase will have to handle these data types.

The initial approach will encourage the use of high-density, strand-specific tiling arrays or high-coverage RNA-Seq data, but integrating traditional expression array (low-density) data will also be necessary because substantial amounts of this data type exist and are still being collected. Also, algorithms will need or can use input from a variety of additional data types, including experimentally determined or *in silico*–predicted TFBS, mutation analysis, gene-neighbor scores, or common functional class annotations.

Workflows would include approaches for assembling, visualizing, and quality-assessing these various datasets; visualizing and comparing results of different algorithms; and, ultimately, validating inferences against direct measurement of network structure and behavior. The workflow is assumed to be modifiable and subject to periodic re-evaluation to update new understanding and capabilities.

The method for implementing workflows will be developed as part of the Kbase Infrastructure (Section 8.6, Workflow Services).

1B. *Identify specific network inference algorithms*.

The ultimate objective is to create a computational environment that provides network inference in an integrated way. In this task, available tools would be evaluated and selected. These would include methods for determining operons and regulons if suitable datasets are available and for clustering genes into putative regulatory modules whose transcription is correlated over a set of conditions. From these clusters, the goal is to assign the common regulators that are the causal antecedents to this observed clustering and to then infer the networks of interaction (chain of regulators) that underlie the overall observed behavior. This constitutes the inference of the static network.

There are many methods for data reduction (e.g., clustering, generalized singular value decomposition, self-organized maps for which there are standard open-source libraries) and for static network inference (e.g., variants on correlation networks, regression-based approaches, Bayesian networks, and parameter inference for biochemical-like network representations). The initial approach will be to find a workable set of proven algorithms that cover the data and prediction types mentioned above.

This approach would provide a starting set of algorithms implemented in Kbase but does not exclude other contributions in an open-community environment. Part of this task also involves collecting a set of gold-standard network datasets with the best information on the *direct* measurement of transcriptional network structure and dynamics in a number of organisms. Existing synthetic datasets will be identified or otherwise constructed and then evaluated for inclusion in Kbase as part of the datasets used for testing algorithmic inferences.

This task involves identifying and testing algorithms and organizing network data. The actual implementation of these algorithms in a workflow would be conducted under Task 1D.

1C. *Collate existing expression data for microbes of interest or those available*.

The two most basic data types used by inference algorithms are sequence and transcript data. Kbase will start with these and later expand to include protein, metabolite, and mutant phenotypic data, among others. Sequence data handling is mature and expected to be easily managed. However, despite great progress in technologies for measuring gene expression, the rigor lags in annotating experimental designs and in assessing the quality of these datasets. Since different algorithms require different experimental designs for collecting data (e.g., time-series, deletions, or replicate point-measurements compared to control over a large number of well-chosen conditions), this task requires establishing methods for uploading or linking expression data sources to Kbase.

Associated experimental data and metadata, which properly account for experimental designs, also should be included. This task will require linking expression data to sequence data and prior predictions of operon or regulon structure. When appropriate, Kbase will link to existing repositories, such as GEO and ArrayExpress.

This task includes identifying first adopter experimentalists and establishing collaborative relationships whereby their laboratories provide Kbase with data and design advice and serve as beta testers.

The total data storage required is based on coverage and number of replicates, conditions, and time steps and therefore would be a multiplicative factor of 4 gigabytes (180X minimum as proposed). For the first 1 to 3 years, 30–100 datasets are expected to be collected per year (each dataset corresponds to studies on one microbe) and then grow to 100–300 per year in the 3- to 5-year time frame when data will be coming from many laboratories.

Storage in the terabyte to petabyte range will be needed in the first 5 years. Data reduction will play a role in keeping storage resources manageable, and online backup capabilities are needed for disaster recovery and long-term archival. The necessary computational resources will be large (more than 1000 cores) and used for data management and integration as well as for network analysis.

1D. *Make available for general use a capability for inference of regulatory networks from expression data (e.g., RNA-Seq, tiling array, or possibly ORF-specific array data; if generalized as an n × m matrix, any technology that generates such data could serve as input)*.

This task involves integrating and deploying the first version of the workflow for general use based on the results of the previous three tasks. A library of the various inference algorithms will be created, along with methods for comparing the outputs of each algorithm to each other and to the gold-standard datasets. Workflows will be developed for organizing and performing quality control of data required for input to each algorithm, for running algorithms and collating their results, and for visualizing and assessing their predictions and quality compared to the gold-standard datasets.

1E. *Create and make available inferred regulatory networks from existing expression datasets*.

This task will use the capability from Subtask 1D to run the system on all available datasets relevant to DOE mission science. It will involve investigating all possible sources, collating the data, and running the system to produce the networks.

1F. *Create a controlled vocabulary for metainformation to capture experimental design, including perturbed environmental and genetic variables, media compositions, and growth conditions*.

The metadata could include optical density, substrate consumption, metabolites, temperature, and incubation condition (as comprehensive as possible). Although some could be manually collected, Kbase would need to have the ability to store these data in conjunction with RNA-Seq as an experimental project.

Kbase would work with GEO and ArrayExpress to capture additional information so that required controlled vocabularies are developed and adhered to in conjunction with the Genomic Standards Consortium and other interested groups and communities.

1G. *Provide a user interface for importing and displaying existing datasets, inferred transcriptional regulatory networks (TRNs), and predicted binding sites (e.g., Pathway Tools, MicrobeOnline, Cytoscape, BioTapestry)*.

The user would specify an organism and import (or broadcast) the various types of data. Many of these data are stored in existing databases such as GEO, MicrobesOnline, or ArrayExpress and can be loaded automatically through interoperability with these sources. Discussion with these groups will be necessary to plan the needed transition toward the much larger RNA-Seq datasets. This effort will not duplicate the existing data in Kbase but will make the systems interoperable. The only reason for such data to permanently reside in Kbase will be because of performance issues.

1H. *Standardize interfaces and application programming interfaces (APIs) for interoperation across selected data repositories, algorithms, and visualization software*.

Kbase will be a repository for algorithms and software tools with open and standardized APIs. This task will be a necessary joint effort with other repositories and services (noted above) to establish community architectural standards for interoperability (e.g., SOAP or REST and client side vs. server side). However, interoperability also is needed in regard to actual service and exchanged data and relates to the specifics of prior tasks described above. Developing interoperability often also involves developing standards, which historically has required many multi-year efforts.

This task would be performed in conjunction with the Kbase Infrastructure team's effort, which is not estimated here. This task, however, is expected to be an ongoing activity that may expand further depending on the number of different activities involving interoperation and standards development.

1I. *Generate standards for regulatory network representations*.

This task is specifically about description of the network. Current technologies for transmitting network hypotheses, such as SBML, CellML, and BioPAX, will be evaluated to determine if a new format is necessary.

1J. *Incorporate other data types into regulatory network models [e.g., transcription start sites (TSS), ChIP-Seq, proteomic, and genome-anchored or unbiased determinations of regulator binding site specificity] for a bottom-up definition of regulatory networks*.

Meeting the mid- to long-term objective will require expanding the data model to incorporate additional types of experimental data, both for improving predictions and analyzing the results of experimental validation. We need to have methods for capturing experimental evidence and quality and then to use these types of data in the analysis and improved predictions.

In addition, there will be an ongoing need to more precisely define and represent phenotype and associated confidence depending on how it is measured.

Task 2. Extend and test inferred gene expression regulatory networks (mid to long term).

The network modeling capability should be extended to additional data types, both to refine the models and test their predictions against experimentally validated identification of transcription units, promoters, regulator binding sites, regulator binding specificity, protein-protein interactions, genetic interactions, metabolomics, and metabolic flux measurements.

2A. *Validate and refine models using various functional data types to allow robust correlation of regulatory networks to genome features* (5 to 10 years).

As they become available, new and especially high throughput data types that can improve models need to be incorporated into Kbase. This will require Kbase to evaluate data and identify and establish collaborative relationships with experimentalists so that their laboratories provide data, offer advice on design and methodology, and serve as beta testers. This task builds and expands from the collaborative experimental relationships in Task 1C.

2B. *Archive in a standardized manner a collection of diverse systems biology data (e.g., transcript profiles, protein interactions, precise transcriptome structures, regulator binding sites, regulator binding specificity, small-molecule concentrations) collected using best practices and accompanied by metainformation on how the experiments were conducted* (5 to 10 years).

Although certainly worthy, this task seems to be beyond the scope of this scientific objective. Nonetheless, Kbase would need to be prepared for and engaged in this effort. Presumably the early Kbase tasks are building toward having this capability. Therefore, no effort has been estimated for this.

2C. *Extend regulatory networks to enough organisms to build a Knowledgebase of the evolution of selected regulatory networks and network motifs through comparative network analysis capabilities such as multiple network alignment* (5 to 10 years).

Once regulatory network inference in a broad range of organisms has been implemented, the next phase of this objective is to analyze and compare their topological structure and attempt to reconstruct their evolutionary history. The workflow for this phase of the project involves the following.

Implement Kbase software tools allowing users to analyze and visualize the genome-wide architecture of a regulatory network. In particular, these tools would allow one to

- Calculate the distribution of regulon sizes and the number of regulatory inputs.
- Perform the hierarchical layout of TRNs using a variety of algorithms (e.g., breadth-first, depth-first, and minimization of the number of bottom-up links).
- Use this layout for network visualization.
- Identify feed-forward network motifs of different types depending on the combination of signs of regulatory interactions (activation or repression).
- Identify and characterize cross-talk and regulatory overlap between different functional pathways.
- Develop tools for comparing regulatory networks in different species.
- Align regulatory networks in different organisms using information about orthologous proteins.
- Trace and visualize phylogenetic profiles for network topological properties in a group of genomes selected by the user.
- Incorporate into this workflow methods for determining transcription factor binding sites.

2D. *Develop a capability for coupled regulatory network models, metabolic network models, and annotation so that information is updated and exchanged* (5 to 10 years).

Interconnecting regulatory networks with metabolic reconstructions and multidimensional annotations (two other high-priority objectives identified by the Kbase microbial group) would greatly facilitate development of microbial systems biology (Koide et al. 2009).

This task involves careful coordination with the other objectives and associated repositories and requires computational services that enable seamless and current interoperation of these capabilities, leading to a more holistic representation of microbial systems.

Resources

Microbial 2: Define Microbial Gene Expression Regulatory Networks

Table 2.3 Hardware Resources for Microbial 2

Hardware Purpose	Type	Size
Data management	Storage	Tens of terabytes to 1 petabyte
Data analysis	Processing	Large (more than 1000 cores)

Microbial 2: Define Microbial Gene Expression Regulatory Networks

Table 2.4 Staffing Resources for Microbial 2

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
1. Enable automated inference of gene regulatory networks (short term)		
1A. Finalize the definition of regulatory network reconstruction workflow.	2 Bfx	0–12
1B. Identify specific network inference algorithms.	2 Bfx	1–6
1C. Collate existing expression data for microbes of interest or those available.	3 Bfx 3 B	0–6 0–12
1D. Make available for general use a capability for inference of regulatory networks from expression data (e.g., RNA-Seq, tiling array, or possibly ORF-specific array data; if generalized as an n × m matrix, any technology that generates such data could serve as input).	4 Bfx	6–12
1E. Create and make available inferred regulatory networks from existing expression datasets.	Bfx	6–12
1F. Create a controlled vocabulary for metainformation to capture experimental design, including perturbed environmental and genetic variables, media compositions, and growth conditions.	Bfx	0–12

Microbial 2: *Define Microbial Gene Expression Regulatory Networks*

Table 2.4 Staffing Resources for Microbial 2

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
1G. Provide a user interface for importing and displaying existing datasets, inferred transcriptional regulatory networks (TRNs), and predicted binding sites (e.g., Pathway Tools, MicrobeOnline, Cytoscape, BioTapestry).	SE	0–12
1H. Standardize interfaces and application programming interfaces (APIs) for interoperation across selected data repositories, algorithms, and visualization software.	2 Bfx	0–36
1I. Generate standards for regulatory network representations.	Bfx	37–60
1J. Incorporate other data types into regulatory network models [e.g., transcription start sites (TSS), ChIP-Seq, proteomic, and genome-anchored or unbiased determinations of regulator-binding site specificity] for a bottom-up definition of regulatory networks.	4 Bfx	37–60

System Releases

Release 1 (Year 1). Integrate and deploy the first version of the general use capability for inference of regulatory networks from expression data.

Release 2 (Year 2). Port the capability to the full Kbase infrastructure.

Release 3 (Year 3). Standardize interfaces and APIs for interoperation across selected data repositories, algorithms, and visualization software.

Release 4 (Year 5). Incorporate additional types of experimental data to improve predictions and to analyze results of experimental validation.

Release 5 (Year 10). Develop a capability for coupled regulatory network models, metabolic network models, and annotation so that information is updated and exchanged.

3. Near-Term Plant Science Needs Supported by Kbase

The first objective in the plant science area is to establish the capability to predict alterations in plant biomass properties caused by genetic or environmental changes. This capability would be based on the mining of data that reflect the complex relationships among the physical properties of plants, their genetic makeup, and the environment in which they grow. The second objective is to develop the ability to organize and analyze regulatory "omics" data to improve understanding of how plants (particularly species relevant to DOE missions) regulate gene expression. This capability will be critical for understanding genes, their action, and regulation—knowledge required to engineer plant growth and development and, in particular, biomass accumulation.

Plant Scientific Objective 1

3.1 Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype

Summary of Objective and its Requirements

Relevance

Gaining an understanding of the genetics underpinning desirable plant biomass properties relevant to DOE missions (e.g., biomass yield, conversion efficiencies to biofuels, and ability to sequester soil carbon or contaminants) depends on the ability to conduct co-relational assessments between molecular and phenotypic data. Identifying the genes underlying the expression of desired phenotypes depends on the association of multiple genotypes contributing to a trait of interest (forward genetics) or, if a candidate gene is being investigated, an understanding of the gene product's gain- or loss-of-function impact on an extended phenotype (reverse genetics). In most cases, the complexity and plasticity of plant growth and development make predicting a perturbation's impact in one specific gene difficult because this phenotypic impact is rarely confined to the pathway in which the gene product operates. Providing a platform for integrating information on genotype, extended phenotypes, and the metadata associated with field and greenhouse growth conditions is key to understanding these genotype-phenotype relationships.

Objective

Computational infrastructure improvements are required to support and contextualize experimental plant phenotypic data to an extent that enables researchers to predict changes in the physical properties of biomass that occur as a result of environmental change, genetic diversity, or manipulation. Achieving this goal depends on creation of a robust semantic infrastructure for collecting, annotating, and storing diverse phenotypic and environmental datasets. These data include measurements such as photographic images and analytical spectra that capture visible phenotypes and chemotypes fundamentally related to yield, physiological performance, and sustainability. Specifically, this infrastructure will serve as a basis for software applications that extract, quantify, and catalog phenotypic features from the data for data mining and further analysis. This involves combining the data with relevant metadata to enable

querying, modeling, clustering, and comparing data from diverse datasets generated by different platforms.

In the short-term, computational tools to aid researchers in designing experimental protocols that provide semantically contextualized data and metadata are required. Implementation of these experimental designs will be facilitated by software applications that support the collection of the semantically contextualized data using mobile devices such as smart phones (e.g., iPhone) or laptop computers. The long-term goal is formal representation of community knowledge regarding the relationships between phenotype, genotype, and environment as a basis for inferring the logical implications of diverse experimental datasets.

Statistical methods are required to assess data consistency, identify correlations, and provide metrics describing the confidence of any conclusions inferred from the data (e.g., genetic or environmental causality of a phenotypic variation). The general statistical framework for such analysis largely exists but is evolving. Currently, implementation of statistical methods that incorporate both phenotypic and genotypic data (e.g., for parent selection in plant breeding experiments) is extremely slow and cumbersome, and methods tailored for processing plant phenotypic data are needed.

A parallel effort in defining metadata, standards, and ontologies is a recognized need. Attaining the scientific objective will require appropriate vocabulary standards for a wide variety of data and metadata that describe phenotypes, chemotypes, genotypes, and the experiments designed to collect these data. Although several such standards and ontologies exist, they require additional expressiveness to achieve this objective. To share the relevant experimental data and ensure its completeness (in terms of associated metadata), a community-approved standard for the Minimum Information for A Plant Phenotyping Experiment (MIAPPHE) would be helpful. Such a standard does not currently exist. Development of all of these standards demands a long-term, committed collaboration between computer and plant scientists.

Appropriate standards for the semantic description and exchange of primary data (physical measurements, images, and spectroscopic data) are not available. Such standards are required to specify, for example, plant form, morphology, anatomy, coloration, development, and function. Developing these standards may involve extending existing standards after identifying their shortcomings. Because some measurements are species-specific, customizing the standards to the representative target plant species (e.g., *Brachypodium*, *Chlamydomonas*, poplar, sorghum, switchgrass, and *Miscanthus)* may be necessary in some cases.

Initial testing of data structures and semantic annotation protocols would be facilitated by phenotypic and genomic datasets that could be analyzed retrospectively, comparing the conclusions obtained via the newly developed Kbase infrastructure and tools to results previously acquired by manual methods.

There are no genomic databases for target species that support the specification of genetic diversity [e.g., single-nucleotide polymorphisms (SNPs)] within the germplasm of existing stocks. Such databases are necessary to identify useful correlations between genetic and phenotypic variations. Populating these genomic databases requires pipelines for calling SNPs *de novo* in the absence or presence of an annotated genome. Such pipelines exist but have not

been validated for the target species, leading to high false-positive rates and low validation rates.

Methods that detect and quantify defined features from complex data (such as photographic images or spectroscopic data) are required to facilitate data correlation within or among datasets. Comparison of the vast amounts of raw data that will be generated is not practical. Furthermore, conceptualizing the correlations embedded within diverse datasets will require representation of identifiable features rather than raw data patterns.

Potential Benefits

Development of a robust, semantic infrastructure for plant phenotyping research is a high-level, mid-term objective that could be carried out in 3 to 5 years. It will streamline the acquisition, annotation, archiving, retrieval, processing, and mining of data that reflect the complex relationships among plants' physical properties, their genetic makeup, and the environment in which they grow.

Developing and redesigning feedstock properties from the level of plant architecture and yield to biomass recalcitrance would benefit from having a unifying semantic infrastructure from which to draw inferences and organize diverse datasets. These benefits may take the form of mobile applications for high-level modeling and for acquisition of previously inaccessible data, experimental design tools, and statistical analyses. For bioenergy crops and model species, integration of data from both high- and low-throughput phenotyping experiments across species and with other omic datasets, although not a short-range goal, is nonetheless critical to refining gene function definitions, building high-level models, interpreting orthologies, and understanding the genetic architecture of traits. This goal depends on being able to relate diverse datasets in a broader biological context that then can be interpreted and used for inference.

Synergies with Other Projects and Funding Agencies

The underlying analytical software and modeling capability developed for Kbase will be generally applicable to all crops and of interest to other groups and government agencies that should be involved in this activity, including the National Science Foundation (NSF), Plant Genome Initiative, and the U.S. Department of Agriculture (USDA) Agriculture and Food Research Initiative. These other initiatives are oriented toward defining trait ontologies for individual crop groups and developing database models to handle phenotypic, genotypic, and provenance data. In most cases, these activities are synergistic with Kbase in that they already have laid much groundwork. Significant overlap existing within these initiatives needs to be resolved into individual contributions. No plant improvement program has any ongoing efforts to provide rapid analysis through the integration of phenotypic and genotypic data.

Illustrative Workflow

Scientists will enter relevant information describing the experimental setup directly into a Laboratory Information Management System (LIMS) or onto a PC application. This information then will be used to develop an experiment-specific data model to automatically configure an application implemented on a mobile device to acquire data in the field. Complementary data

for the same set of plants will be generated using a broad range of instruments, but the data will be integrated using semantic annotation and made conformant with community standards for representation and content (e.g., the proposed MIAPPHE standard). During data acquisition, the experimentalist will be able to eliminate artifactual data and impute missing data using automatic, semiautomatic, and manual methods implemented on the mobile device. These data then would be uploaded, along with metadata, using representations that reflect the relevant experimental data model. This data model will suggest certain types of analyses that could be run automatically or prompted via an interview process. Data processing may be as simple as performing analysis of variance. Complex experiments, however, might require comparing varieties or individuals whose phenotypes were recorded in different years, in different experimental groups, and in different locations, in combination with genotypic data in a genome-wide association study and archival environmental (e.g., weather) data. This will make it possible to evaluate temperature and moisture variations across years and locations as well as determine how they affect the identification of candidate quantitative trait loci (QTL) or estimated breeding value.

Kbase capabilities and support of such a workflow would provide several additional benefits. First the experimental design platform could help organize collaborative efforts, clarify thinking, assist with project management, and align the experiment to semantic relationships and ontologies. Second, the user interface will configure instruments required for data collection, making this process more accurate and efficient. This interface also will ensure data are uploaded through client software to Kbase and will allow GPS and other datasets to be collected in the background via satellite communication and networks of weather data. Kbase also offers the benefit of leveraging someone else's efforts in translating proprietary data formats into standardized ones. New methods developed by users would be recorded in Kbase for other researchers to use and potentially improve. Finally, Kbase would enable systems biology through its semantic architecture.

An example workflow, illustrated in Fig. 3.1 below, proceeds from upper left to lower right. A user designs an experiment employing a Kbase interface (upper left oval). The data to be collected (yellow box surrounded by data and metadata in green) is determined in part by instrumentation and also by the user based on the specific objective. Data would be checked for errors and for conformance to controlled vocabularies. One or more analysis modules shown as blue rectangles following the decision points (orange diamonds) would be selected. These modules themselves may be multicomponent pipelines for reducing data dimensionality, extracting features, genotyping, or other specific goals. Results would be incorporated into Kbase. As the database grows, potential for comparison across experiments would expand and further enable systems-based approaches (brown oval, lower right). For additional workflow details, see Section B.2 in Appendix B.

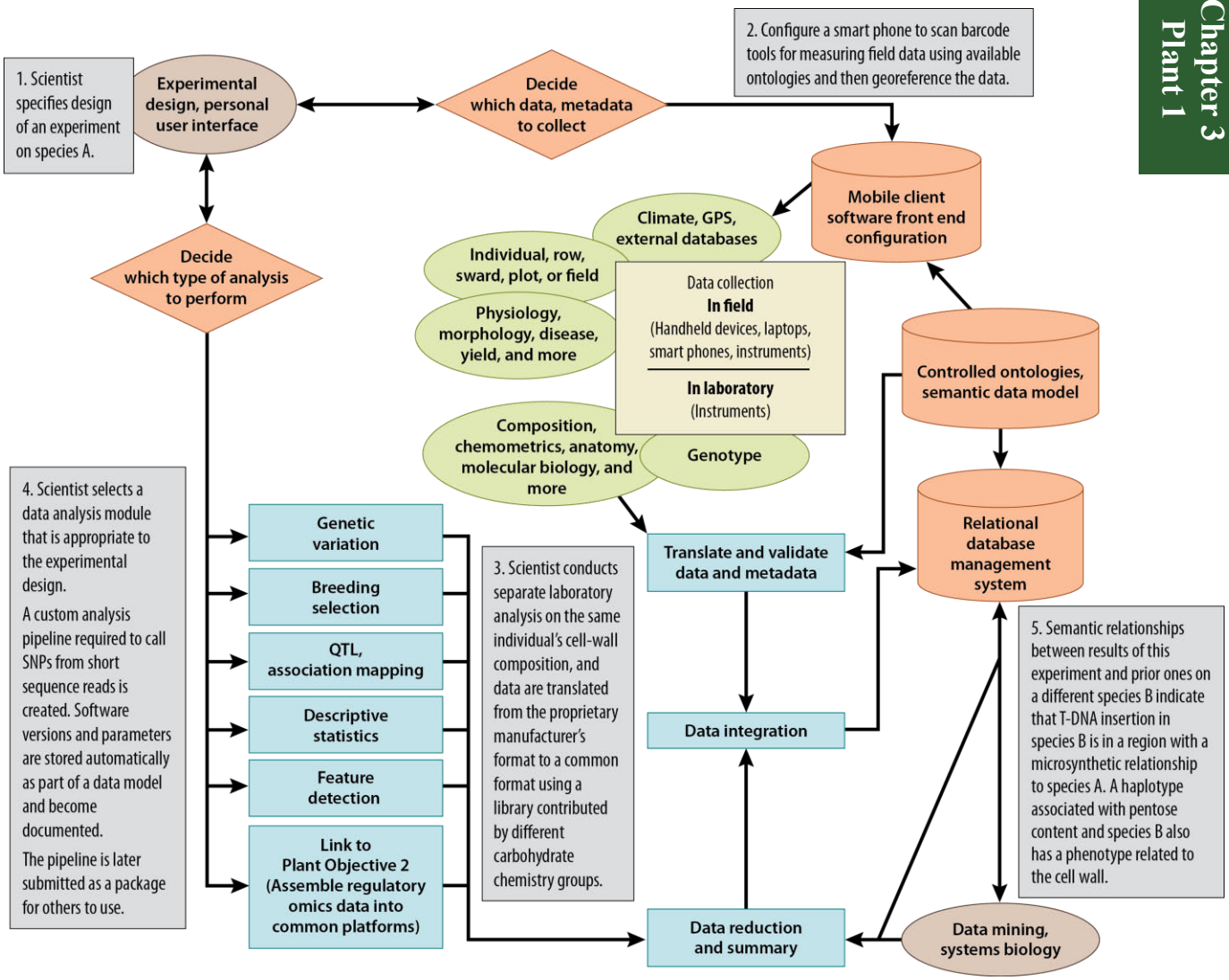


Fig. 3.1. Example Workflow for Integrating Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype.

Implementation Plan for Integrating Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype

System Capabilities

Broadly stated, Kbase should provide the capability to organize and use related phenotypic and molecular data for predicting changes in the physical properties of plant biomass arising from shifts in environmental conditions or from genetic manipulation. This capability should include methods enabling users to acquire and upload data in the correct semantic syntax even though they may not know the semantic principles involved or details of the experimental design. The releases should support data from a variety of sources but focus on phenotypic data, genotypic data, and associated metadata collected during the study of biomass characteristics such as agronomic traits and chemotypic and physiological properties. These data types range from multidimensional images and spectra to single values and are associated with extensive design information. To be valuable to the end user, Kbase must provide data analysis capabilities not otherwise available.

Three capstone capabilities define the platform:

- **Standards:** The ability to develop, use, and extend new and existing standards as they apply to common vocabularies, taxonomies, thesauri, and ultimately ontologies.
- **Semantic representations and linking:** The formalized relationship among what is measured, its environment, and properties.
- **Enabling software:** Tools to efficiently acquire and analyze phenotypic and molecular datasets.

Implementing the infrastructure and tools required to accomplish this scientific objective will require continuous interactions with developers of other computational and bioinformatics resources, including (but not limited to):

- The International Crop Information System (www.icis.cgiar.org/icis/index.php/ICIS/)
- Epicollect (www.spatialepidemiology.net)
- PhenoMap (www.appstorehq.com/phenomap-iphone-113872/app)
- The International Plant Genetic Resources Institute (IPGRI; www.bioversityinternational.org/scientific_information/themes/germplasm_documentation/overview.html)
- The Gramene Plant Ontologies (www.gramene.org/plant_ontology/)
- Gene Ontologies (www.geneontology.org)
- The Genomic Diversity and Phenotype Data Model (www.maizegenetics.net/gdpc/)

Maintaining a sufficient understanding of the capabilities and limitations of these resources is necessary to facilitate collaborative efforts (which are categorically required for developing standards), optimize their synergy with Kbase, and minimize functional overlap. Due to the complexity and diversity of these resources, maintaining this information and establishing

strong communication ties with resource developers are major undertakings that most likely will require the attention of a full-time scientist working as a liaison. Such an individual would need in-depth knowledge of the current status of plant semantics within Kbase and have sufficient authority to initiate collaborative projects and terminate or redirect Kbase projects redundant (with respect to function or information content) with external resources.

Tasks

Task 1. Develop a semantic infrastructure for representing concepts related to plant phenotype, chemotype, genotype, and growing environment.

1A. *Use and extend existing controlled vocabularies and develop new ones that apply to plant phenotype, chemotype, genotype, and growing environment.*

In addition, define the relationships between terms in controlled vocabularies. This task will require working with appropriate existing infrastructure such as the Gene Ontology (GO) project (www.geneontology.org), Protein Ontology (PRO) project (pir.georgetown.edu/pro/pro.shtml), and the Phenotypic Qualities Ontology (PATO; obofoundry.org/wiki/index.php/PATO:Main_Page) (GO/PO/PATO). Collaboration with these projects will help support curation of controlled vocabularies, identify gaps in them that prevent implementation of requirements, and extend the ontologies through relationship-building with existing plant ontology efforts. Existing software (both commercial and freeware) will be evaluated for the task of managing controlled vocabularies. Protégé or a similar tool often will suffice and has the extra benefit of being an open-source project with an active base. Also, the Protégé-OWL editor enables users to build ontologies for the semantic web, in particular in the World Wide Web Consortium's Web Ontology Language (OWL).

The creation of a semantic infrastructure will require an interdisciplinary effort that consists of staff with computer science and biology skills. The computer science–related skills would be in the area of semantic data representation, likely requiring someone with experience using extensible markup language (XML), Resource Description Framework (RDF), and OWL. The development of reasonable metamodels for plant phenotypes, chemotypes, genotypes, and growing environments will involve the effort of two full-time staff. Various XML-based standards already exist or are under development for a variety of data types listed here. The work involved in assessing these data models is going to be large. The duration of this task will depend on the effort required to get some community consensus on standardized vocabularies and the relationships among terms in those vocabularies. For this, a third part-time person is needed to solicit input from professional societies and experts, coordinate and plan meetings used to select or develop standards, and advertise the standards.

1B. *Translate semantic structures to a consistent schema for database design.*

Using semantic structure, a relational database management system (RDBMS) will be developed that is consistent with the structure and extends existing phenotypic, germplasm, and genetic data schemas. This work will require a computer scientist familiar with RDBMS and biological databases such as the Generic Model Organisms Database (GMOD), Gramene, and Genomic Diversity and Phenotype Data Model that provide some perspective for creating Kbase. The significant and ongoing effort needed for developing this resource will require coordination with activities described in Subtask 1A.

1C. *Provide necessary data services to register, store, query, and retrieve data from the data model.*

Well-developed data transfer protocols (e.g., FTP with support for XML and interconvertable formats like delimited data) can be implemented to support standards and semantics. User-initiated queries may be supported through software interfaces employing existing technologies such as SPARQL. Interconnectivity with other biological databases can be established via applications through SOAP protocols. These data services would be developed concurrently with Subtask 1A. This task would require a full-time computer scientist with experience in semantic web technologies and web service technologies.

1D. *Apply the metamodel developed in Subtask 1A to relevant existing phenotypic and physiological data.*

To evaluate the metamodel and the fit with existing phenotypic and physiological data collected from bioenergy species, the model will be applied to several existing datasets. This evaluation will verify the metamodel's validity and identify further gaps that need correcting for subsequent releases. This task will occur during the first year of the project and will require, along with several collaborating plant biologists, the part-time effort of a computer scientist.

1E. *Apply the metamodel developed in Subtask 1A to relevant existing image and multidimensional datasets.*

To evaluate the metamodel and the suitability of existing data to construct formal data models, we need to work with proprietary data formats provided by instrument manufacturers. Adopt open-source community standards where possible. For example, mass spectrometry data might be represented in a proprietary format, but a common format (mzXML) also has been developed (see Subtask 5B). Similarly, near-infrared (NIR) spectral data and calibration models sometimes are nonconformant or platform specific—a recognized impediment to progress—but Continuous Media Markup Language is an XML-based alternative for working around these difficulties. Existing image data and metadata models will be evaluated (e.g., the National Information Standards

Organization's Metadata for Images in XML, called NISO MIX). Integrating these data types into Kbase probably will be a problem encountered in other objectives. It thus should be dealt with at a high level.

Task 2. Develop software for data collection that utilizes the semantic infrastructure.

2A. *Develop software clients for collecting data in the field.*

New software that will run on portable hardware devices and conform to the relevant semantic metamodel is a high priority. Metamodels of plant phenotyping, genotyping, and environmental growth conditions must be utilized by data-collection software clients in a way that enables a person in the field or greenhouse to collect phenotypic data easily. This means supporting mobile hardware devices such as tablet and laptop computers as well as smaller hand-held devices. Developers may need to target devices equipped with appropriate hardware to enable acquisition of GPS data, barcode scanning, and tagged images. These devices would communicate with the server and would perform data validity checks. Accomplishing this subtask would require a software developer and biologist to work together and build off of other existing software mentioned above.

2B. *Develop server software that will accept, validate, and add data from a variety of clients.*

This task will enable communication with a variety of mobile devices, desktop computers, and tablets. Through wireless or other means of data transport, the software will receive data from the field-collection client software and then store and register it into the appropriate model. Because the range of data types and sizes varies significantly (from measurements like temperature to eukaryotic genome sequencing data), multiple data transfer protocols are required. NCBI's Sequence Read Archive has implemented Aspera Connect data transfer protocols that use a proprietary protocol on top of User Datagram Protocol to maintain reasonable transfer rates over wide-area networks for short-read sequence data. Although this is a commercial product, open-source variations are available that address the need to efficiently move large datasets over wide-area networks. Alternatively, moving data from small data-collection devices in the field or greenhouse over a wireless network will involve relatively small datasets.

2C. *Enable users to save and store routines or configurations used by client software for experimental data collection.*

Envisioned is an application that is flexible and configurable enough to be used in a variety of circumstances. This task will provide methods by which individual users can register devices and configure them in either offline or online modes (in cases where there is no wireless coverage or in remote locations) to gather data though stored "routines. It is related to Task 3, which also involves

interactive data manipulation. These methods would configure the application in different modes to prepare it for accepting new data of different types. This task would require a software developer familiar with AJAX or a software development kit similar to those for Android or iPhone smart phones.

2D. *Enable rapid deployment of barcoding systems within a field setting.*

A further driver for Kbase adoption would be incorporation of software functions to streamline barcode creation and printing that would be consistent with ontology terms and could be used independently, but ideally in conjunction with the mobile software. This is a relatively straightforward software development task and can be based on existing systems for maize and other crops. The principle activity will be creating documentation and user guides describing different potential applications.

Task 3. Implement interactive methods for manipulating, describing, and assessing the quality of data and metadata.

3A. *Develop server software features that enable interactions (e.g., additions or modifications) with data and metadata.*

This task is related to Subtasks 2B and 2C. It would occur through a web browser after uploading data from a variety of sources and enable manipulation of experimental details to more accurately describe data in terms of the semantic model. For example, automatic prompts for missing information and suggested additional descriptors could increase metadata value and completeness. Ideally, software should encourage conformance of data and metadata to a standard: MIAPPHE could be based loosely on the Minimum Information About A Microarray Experiment (MIAME) standard for microarray datasets. The interface should allow downloading of data in formats appropriate for local analysis (e.g., Tassel, Flapjack, Excel, and JoinMap).

3B. *Aggregate related datasets; identify outliers, duplicates, and irrational values; and summarize experimental metadata.*

This task will develop server-side software that will provide statistical summaries to individual users about the current dataset. Higher statistical functions may be accessed by programmatic calls to R statistical software, and graphics abilities through Matlab or coded directly. This task would be limited to simple methods, summaries, correlations, and counts of columnar data.

Task 4. Provide an infrastructure for data mining and analysis based on statistical procedures.

4A. *Evaluate the suitability of existing data models for genetic diversity and phenotype and develop or extend these systems to align with the semantic infrastructure.*

Existing database schemas of phenotypic and marker data may not be ideally suited for next-generation genotyping methods or optimally aligned with developing ontologies. Required activities include identifying and evaluating existing models and developing new models suitable for application of trait and genotype association methods in plants that may or may not have reference genomes. Creation of a robust and flexible database aligned with developing ontologies will require an interdisciplinary effort that consists of a staff with computer science and biology skills. The computer science–related skills will be in the area of database development, likely requiring someone with experience using RDBMS based on Structured Query Language (SQL). This database will be developed with assistance from a part-time consulting biologist or statistician with expertise in quantitative genetics or statistics as well as next-generation sequencing.

4B. *Implement a basic set of analyses for a genome-wide association study, QTL study, or for applying genome-wide selection.*

Predictive modeling techniques desired by plant scientists, including ridge regression, partial-least squares regression, and best-linear unbiased prediction, should be implemented to provide model-based genomic estimates of breeding value. Clustering of individuals based on genetic similarity will also be required. This task may best be implemented through contributed packages to R or a statistical genetics project. Some existing parallel open-source statistical computing and statistical genomics efforts are well advanced. These efforts would need to be identified and assessed, augmented if required, and integrated. This task needs to be evaluated in light of the complexity of some taxa, particularly with respect to polyploidy. Establishing how best to determine allelic variation at a single locus in polyploids through sequencing or other genotyping methods is an area that still requires better technology and more research effort. Simulation studies and empirical data are lacking.

Task 5. Provide feature recognition software for extracting and quantifying features in raw data (e.g., images and spectra).

5A. *Adopt and integrate existing software for detecting features in photographic images for bioenergy applications.*

This field is well advanced, and integration of existing feature extraction techniques should occur through collaboration with major research centers and should focus on bioenergy areas of application. This task overlaps largely with other Kbase groups, so work should be coordinated at a higher level. There are

many image acquisition and analysis applications—including specialized microscopy databases (www.cbi.cmu.edu) and biological image databases that enable storage and semantic query (e.g., the University of California at Santa Barbara's Vision Research Lab; www.bioimage.ucsb.edu). These databases can provide time series, video, and z-stacks that reconstruct plant cell fate and developmental pathways. Anticipating which techniques would find the greatest use is difficult. Highest priority would be image segmentation for automatically measuring area, color, volume, and length of irregular objects and the ability to apply image screening to parameterize images with minimal intervention. This task will require collaboration among Kbase computer scientists and researchers at other institutions to integrate key functionality and ensure semantic structures are rich enough to accommodate image data. To enable basic image features, one person skilled in programming and familiar with areas of bioimaging is needed.

5B. *Incorporate spectroscopic data and provide quality metrics.*

As with imaging, high-throughput analysis of experimental data can involve simultaneous measurement of 100 to well over 10,000 analytes on the order of ~100,000 or more samples. Instrument-neutral XML standards are still under development by the International Union of Pure and Applied Chemistry and ASTM (e.g., Analytical Information Markup Language; animl.sourceforge.net) as well as industry and user groups. Some prerelease data models will be tested with existing datasets and used as a basis for later releases of Kbase that will accept user data and set up infrastructure for analysis of proteomic and metabolomic datasets. Once these standards are identified, evaluated, and incorporated into Kbase (see Subtask 1E), implementing the ability to perform spectral quality analysis and provide feedback to users would be an initial valuable feature, particularly for some types of mass spectrometry. The focus initially should be on Subtask 1A and Subtask 1E, which will require a multifaceted approach to manage interactions with all different entities.

5C. *Implement methods to analyze datasets of correlated features to provide predictive ability (NIR, mass spectrometry, images).*

Both analytical and predictive applications of NIR, Fourier transform infrared, Raman, and mass spectra datasets are available, and as many as possible will need to be implemented in Kbase. NIR is used in many laboratories for analysis of biomass. Predictive approaches use NIR training and validation datasets along with wet-chemistry analysis to create calibrations. Methods using principal component analysis (PCA) and partial-least squares regression will be implemented in Kbase, probably through the efforts of a computer scientist or statistician through calls to R or Matlab. However, this ability is already available to most NIR users through proprietary software provided by instrument manufacturers. Lacking for most users is an efficient method to transfer

calibration models between instruments that are all somewhat different. Kbase's collaborative features and open data models along with statistical methods for NIR standardization will provide the user with real value by enabling laboriously developed calibrations to be used across more than one instrument. For example, a researcher could use the same calibration model for biomass in multilocation field trials over time. This specific task could be better formulated by someone with experience in a broad range of spectroscopic applications, whose expert opinion needs to be actively sought to identify additional opportunities. At least one such person should be tasked with actively seeking such opportunities and acting as a liaison with other efforts.

Resources

Plant 1: *Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype*

Table 3.1 Hardware Resources for Plant 1

Hardware Purpose	Type	Size
Data management	Storage	Terabytes
Data analysis	Processing	Small (less than 100 cores)

Plant 1: Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype

Table 3.2 Staffing Resources for Plant 1

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics; CH = Chemistry)

Task or Subtask	Expertise	Duration (Months)
Liaison with external community efforts (USDA, NSF)	B	1–60
1. Develop a semantic infrastructure for representing concepts related to plant phenotype, chemotype, genotype, and growing environment.		
1A. Use and extend existing controlled vocabularies and develop new ones that apply to plant phenotype, chemotype, genotype, and growing environment.	CS B	1–36
1B. Translate semantic structures to a consistent schema for database design.	Bfx	12–36
1C. Provide necessary data services to register, store, query, and retrieve data from the data model.	CS	24–36
1D. Apply the metamodel developed in Subtask 1A to relevant existing phenotypic and physiological data.	CS B	24–36
1E. Apply the metamodel developed in Subtask 1A to relevant existing image and multidimensional datasets.	CS	24–36
2. Develop software for data collection that utilizes the semantic infrastructure.		
2A. Develop software clients for collecting data in the field.	CS B	24–36
2B. Develop server software that will accept, validate, and add data from a variety of clients.	SE	24–36
2C. Enable users to save and store routines or configurations used by client software for experimental data collection.	SE	24–36
2D. Enable rapid deployment of barcoding systems within a field setting.	SE B	24–36
3. Implement interactive methods for manipulating, describing, and assessing the quality of data and metadata.		
3A. Develop server software features that enable interactions (e.g., additions or modifications) with data and metadata.	Bfx	36–48

Plant 1: Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype

Table 3.2 Staffing Resources for Plant 1

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics; CH = Chemistry)

Task or Subtask	Expertise	Duration (Months)
3B. Aggregate related datasets; identify outliers, duplicates, and irrational values; and summarize experimental metadata.	CS	36–48
4. Provide an infrastructure for data mining and analysis based on statistical procedures.		
4A. Evaluate the suitability of existing data models for genetic diversity and phenotype and develop or extend these systems to align with the semantic infrastructure.	CS B	36–48
4B. Implement a basic set of analyses for a genome-wide association study, QTL study, or for applying genome-wide selection.	CS S	48–60
5. Provide feature recognition software for extracting and quantifying features in raw data (e.g., images and spectra).		
5A. Adopt and integrate existing software for detecting features in photographic images for bioenergy applications.	CS	24–36
5B. Incorporate spectroscopic data and provide quality metrics.	CS, S	36–42
5C. Implement methods to analyze datasets of correlated features to provide predictive ability (NIR, mass spectrometry, images).	CS CH	42–48

System Releases

The three enabling capabilities will be delivered in three releases, such that each release will deliver a portion of every capability.

Release 1: Standardized data collection and description capability. The first release is anticipated in a 1- to 2-year time frame. It will involve establishing a basic semantic infrastructure that includes support for the development and maintenance of ontology-based domain metamodels as well as the first release of these models for plant phenotype, genotype, chemotype, and environmental growth conditions. Statistical capabilities for summarizing data also will be included. The initial release is primarily focused on the interconnected semantic infrastructure and mobile application. The primary focus would be to provide users of smart phones and other Kbase-enabled devices the means to reduce time, labor, and human error associated with data entry in environmental and field studies. This will simultaneously drive the

adoption of *de facto* data standards within target communities by promoting the use of these devices, which, along with barcoding systems, are becoming ubiquitous.

Release 2: Capabilities refinement and data models. This release is anticipated in the 2- to 3-year time frame. The user should be able to perform additional standard statistical analysis and inference. This release will include refinements to the existing metamodels. Included will be actual models containing biomass-related data of the relevant types (e.g., phenotypic, genotypic, and environmental) for sample taxa.

Release 3: Knowledge discovery. This release is anticipated in the 3- to 5-year time frame and will host a formal representation of community knowledge regarding the relationships among phenotype, genotype, and environment. The goal of this release is to enable a user to predict changes in the physical properties of biomass that result from environmental or genetic changes.

Plant Scientific Objective 2

3.2 Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Analysis, Comparisons, and Modeling

Summary of Objective and its Requirements

Relevance

Assembling regulatory omics data from plant biology into common platforms is essential to DOE's systems biology mission. Without key data, including dataset acquisition, coupled with analysis of their interactions, no informed predictions of biological systems can be attempted. Naive attempts at networks are certainly possible with co-expression data, but they are highly limited and represent neither the full spectrum of what can be accomplished with current technology nor what should be completed if the mission is to understand plant species on a systems level.

Objective

This scientific objective seeks to collect several key types of regulatory omics data and associated quality metadata for six target plant species: *Brachypodium*, *Chlamydomonas*, poplar, sorghum, switchgrass, and *Miscanthus*. Such information will support the other plant objectives, including annotation (see Section 5.5, Improve Plant Genome Annotation Datasets and Make Them More Accessible), comparison (see Section 3.1, Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype), and modeling (see Section 5.3, Construct, Simulate, and Validate Plant Life Models). RNA levels as measured by expression arrays or RNA-Seq are no longer sufficient to evaluate the mechanisms and networks that regulate plant transcriptomes. Kbase also must include available small RNA and target RNA information, differential RNA processing and decay information, and epigenetic marks such as DNA methylation and histone modifications. This information is important for data integration and for filling in important missing links in gene regulatory networks within a species and facilitating their comparison across two or more species.

In the near term (1 to 3 years), classical transcriptomic data (microarrays and mRNA-Seq) as well as small RNA and basic proteomic data will be assembled. Epigenetic data, small RNA target and RNA degradome data, other types of RNA processing data, and additional proteomic data will be assembled after the first year, beginning with the most developed genomes such as *Brachypodium*. The data will be made publicly accessible with user-friendly web interfaces and will be downloadable for power users.

Understanding which genes are regulated during growth and development and under various conditions is critical for elucidating gene function and regulatory networks. The massive amounts of genome-wide gene expression data accumulating for plant systems can be used to evaluate these controls at the transcriptional and post-transcriptional levels during development and in response to stimuli such as adverse environmental conditions. RNA abundance levels have been assayed routinely using microarrays and, more recently, using mRNA-Seq, which is the current state-of-the-art approach (Wang et al. 2009). Since 2005, small RNA data from deep sequencing also have been accumulating. These data report on miRNA and

siRNA abundances and gene silencing potential (reviewed in Chen 2010). Additional types of emerging data and data analyses are providing insight about miRNA targets and the RNA degradome (German et al. 2008; Addo-Quaye et al. 2008), as well as other aspects of RNA processing such as alternative and regulated splicing and polyadenylation (Licatalosi and Darnell 2010). Beyond RNA data, proteomic data from shotgun mass spectrometry are available for some species, allowing evaluation of protein levels to examine translational control. To effectively evaluate gene expression, all of these data are required. They also provide essential support for the other plant objectives.

A pipeline is required to provide access to omics datasets, genome sequences, and genome annotations from external sources. The acquired data will include sequences, quality information (e.g., Q values), and associated metadata. Sources will include the National Center for Biotechnology Information [e.g., GenBank, Gene Expression Omnibus (GEO), and Sequence Read Archive (SRA)], the DOE Joint Genome Institute (JGI), ArrayExpress, and the Plant Expression Database. Analysis of the data assembled by the pipeline will include genome mapping, normalization (across datasets and platforms), association to annotated genome features (e.g., genes, exons, and splice junctions), *de novo* assembly of applicable high-throughput screening (HTS) data, clustering of expression profiles, clustering and special analysis for small RNAs, and summarization for linkage to genome annotation pipelines.

Standards are well defined for some omics data (e.g., MIAME for microarrays) and for conventional expressed sequence tags and cDNA sequences. For other types of omics data, however, they are emerging, poorly defined, or nonexistent. NCBI's SRA and GEO standards may be acceptable surrogates for RNA-Seq and other HTS data.

Potential Benefits

Achieving the foundation of this high-level, near-term objective is feasible in a 1- to 3-year time frame. Methods exist for generating and analyzing large-scale regulatory omics data. However, these methods need to be applied to the target species, analyzed, and integrated. Although a portion of regulatory omics data has been generated on select target species, no comprehensive effort is under way to characterize complete sets of regulatory omics data.

Plant regulation is known to control key aspects of plant carbon allocation and partitioning, which are critical to biomass composition and soil carbon accumulation. Regulation is also a critical distinguishing characteristic between annuals and perennials and other aspects related to sustainability. To date, we have limited understanding of how plants regulate gene expression and how this is manifested in the cell. Essential to understanding and then engineering plant growth and development for DOE missions is an informed understanding of genes, their actions, and their regulation. Our early understanding of gene regulation was focused on upstream promoters and mRNA expression levels. We now are aware of entirely new pathways of regulation involving small RNAs, post-transcriptional control, the epigenome, and more. Deep research in understanding multiple types of regulation at the DNA, RNA, and protein level is occurring in plant, mammalian, yeast, *Caenorhabditis elegans*, and fly systems. Currently, *Arabidopsis* is the most well studied plant with respect to regulatory pathways affecting genes and their products.

Synergies with Other Projects and Funding Agencies

Systems biology is an immature field in plant biology (Coruzzi and Gutiérrez 2009). Certainly, large-scale datasets are being generated in an array of plant species. The focus of this objective on key species relevant to the DOE mission will deepen and expand these resources. Additional major advances relevant to this objective will arise from the genome technology field, such as improvements in cost and throughput in genomic sequencing. Algorithmic and computational advancements in network prediction and visualization are under way in model organisms and are made available to the greater research community via publications, open-source software, and collaborations including Kbase. The DOE JGI, through its work with plant sequencing and the Phytozome portal, will also provide a valuable resource and partner for this objective. Partnering with DOE microbial systems biology scientists who have experience in constructing regulatory networks would provide great synergy. This objective may overlap with iPlant (see Chapter 6) and other resources such as the Protein Data Bank (PDB), but the focus on bioenergy crops and models is unique to DOE and USDA.

Illustrative Workflow

Plant biologists want to access high-quality, well-documented omics datasets associated with relevant plant gene annotations. There are three main deliverables:

- Consolidated platform for access to omics datasets, genome sequences, and genome annotations acquired from external sources.
- Platform for pre-computed and on-the-fly analysis of plant omics datasets.
- Web-based interface that will enable users to mine plant omics datasets and associated annotations.

Plant biologists want to be able to access omics datasets in a single location (Kbase) and traverse between plant species, while being confident that the underlying data analysis and annotation methods are comparable and of consistent high quality. Additionally, they will want the capability of processing new or custom omics datasets with the same tools and pipelines used to analyze the data already summarized in Kbase. To achieve these goals, Kbase will need to feature a user-friendly interface for the general user, providing summaries of gene and protein expression profiles and clusters as well as links to functional genomics resources (e.g., genome browsers, descriptive annotations, and publications). Kbase also must make the analyzed and summarized data available to users as downloadable, genome-scale datasets and associated metadata. Workflows that enable analysis of user-supplied data in Kbase will be required. These workflows need to be easy to use and comprised of well-defined pipeline modules.

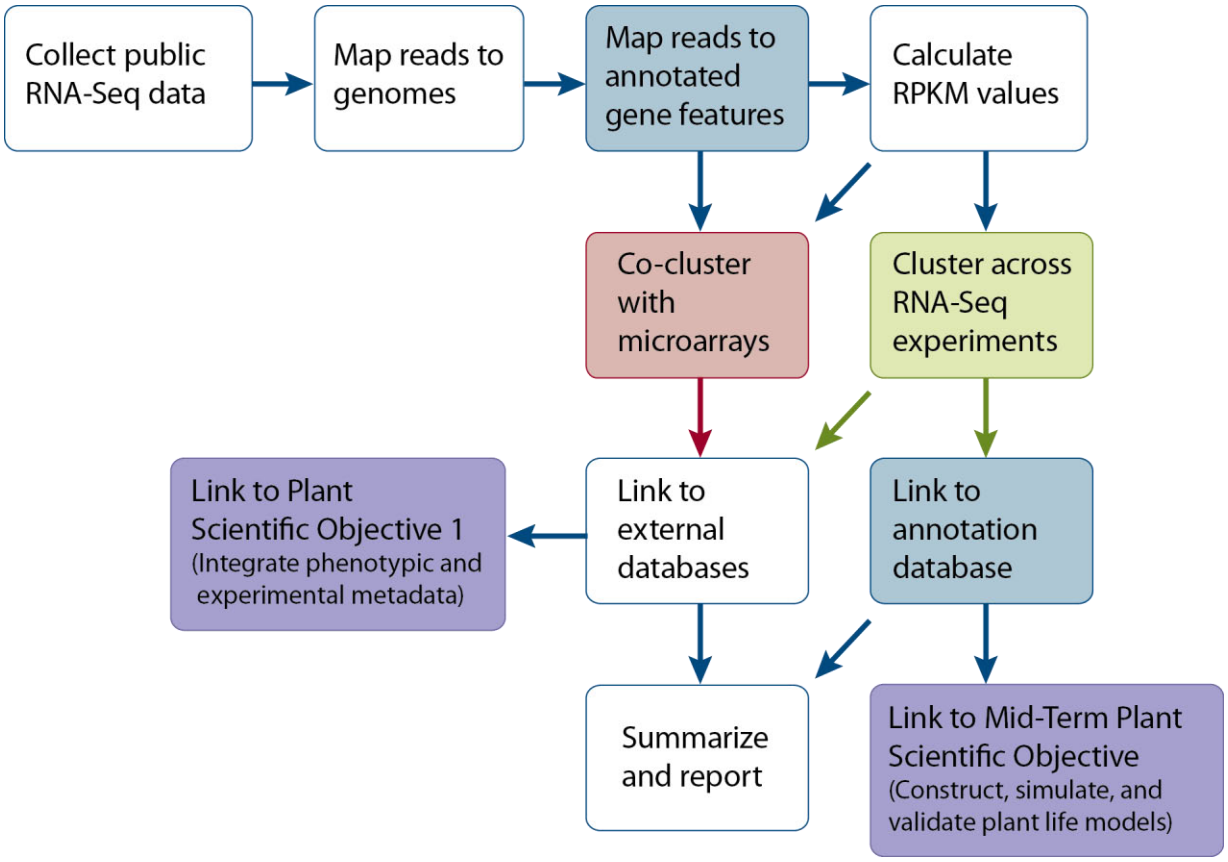


Fig. 3.2. Transcriptome Analysis Pipeline for RNA-Seq Data. White boxes are established procedures. The green box is a procedure that has not been developed but is expected to be fairly easy to construct. The red box is a procedure that will require research efforts. Blue boxes depict a linkage to existing and improved annotation sources, and purple boxes depict linkages to the other near- and mid-term plant objectives for Kbase (see Sections 3.1 and 5.3, respectively).

For additional workflow details, see Appendix B.

Implementation Plan for Assembling Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Analysis, Comparisons, and Modeling

System Capabilities

Kbase system capabilities will be critical to the understanding of genes, gene actions, and gene regulation required for engineering plant growth and development for DOE missions, particularly biomass accumulation. The system will have the capability to collect several key types of regulatory omics data and associated quality metadata and integrate such data with six target representative plant species: *Brachypodium*, *Chlamydomonas*, poplar, sorghum, switchgrass, and *Miscanthus*. These plant species would be developed and curated to high quality as a foundation for global plant studies and interpretation of omics data.

Tasks

Task 1. Establish a reference plant genome platform starting with six foundational genomes and with capabilities for visualizing and comparing genomes, recognizing orthologs and parologs, and automating curation of reference genomes.

1A. *Develop a platform and methods for better comparing plant genomes.*

This task will include developing new tools for small RNA and potentially other features beyond current annotation. Also required are comparison and interface tools to recognize orthologs and paralogs and view such relationships. Other requirements involve developing and deploying tools to (1) recognize and filter transposable elements; (2) perform repeat finding for plants (gene sequence), perhaps including semiautomated methods; and (3) retrieve or compare the contents of external plant reference databases in collaboration with the DOE JGI and iPlant.

This task involves improvement of automated or semiautomated methods for assessing a gene's function by better combining related informatics data and experimental data. Also needed is development of curated reference plant protein datasets that can be provided as reference data to the community.

1B. *Establish a curatorial process and third-party curation tools for continual improvement.*

This task deals with establishing a team, tools, and process for persistent data curation of genomes, genes, RNA, proteins, and function areas. It also involves building and deploying tools for automated and third-party curation for continually improving curatorial processes and results, as well as providing the necessary database models and procedures for integrating such omics data.

Task 2. Develop a platform for access to consolidated omics data.

The concept of allowing data to be hosted on a remote non-Kbase system will be governed by the stability of the host system, the programming interface enabling access to data, and the quality and stability of the metadata structure. In general, these criteria are difficult to achieve. With advancements in semantic web technologies and their application to RNA-based data, remote hosting of data will become more widespread. The long-term success of this plan and Kbase depends on leveraging experimental efforts across the research community. In the meantime, data will be aggregated within the Kbase system unless the criteria described here are met.

2A. *Develop standards and methods for locating, transporting, storing, and retrieving plant omics data.*

Develop methods to locate RNA sequence data. The primary initial source of RNA sequence data will be NCBI's SRA and GEO. These two resources contain considerable amounts of existing gene expression data. Using the Entrez server,

it is possible to identify new submissions and automatically download them from both SRA and GEO. Implementing these programs is needed and could be done fairly quickly.

Additional methods need to be developed that monitor sources of RNA sequencing data not in the NCBI repositories, such as the DOE JGI and other willing sources. These methods would likely range from using automated protocols for remote data synchronization to manual methods such as email. In the case of manual-based data location methods, a user interface should be supplied for registering new datasets with Kbase as well as for transferring the data when centralizing it within the Kbase system is necessary.

Develop methods for transporting RNA sequence data. These methods will vary depending on the source of data. For example, in SRA, data is transported primarily using a commercial client and server application sold by the company AsperaSoft. For the DOE JGI, data can be located and transferred using RESTful (Representational State Transfer) web services. Each data source is anticipated to have a unique infrastructure that will require specific methods to be written for data transport.

Develop storage resources needed for RNA sequence. Storage resources for RNA sequence data will be considerable, with estimates ranging from in the terabytes to petabytes. Current consideration of recently emerging file systems centers on Hadoop, a file system supported by a Kbase pilot effort focusing on a new architectural paradigm for large-scale computing based on the MapReduce architecture published by Google. Alternatives that can be provided by DOE's Office of Advanced Scientific Computing Research and the commercial cloud-computing industry should be utilized.

2B. *Develop appropriate semantic metamodels to apply to omics data.*

This will be an ongoing task developing and refining metamodels and involves collaboration between a biologist and bioinformaticist.

Task 3. Extend the platform to support the generation of pre-computed and on-the-fly analyses of plant omics datasets. (CPU medium, storage TB)

3A. *Develop a configurable pipeline(s) to analyze RNA sequencing reads.*

Map RNA sequencing reads to reference genomes. Cross-reference mapped reads to annotated genes, calculate coverage data per gene, and cluster expression profiles across experiments and platforms (both RNA sequencing and microarray platforms).

3B. *Develop appropriate semantic metamodels to apply to pre-computed analysis results and to the more stable on-the-fly analyses.*

Describe pipeline(s) using a formal process description language. Such languages have been developing in recent years and will be applied to formally describe pipelines created in the previous subtask. The new Hadoop Process Definition Language (hPDL) is a process workflow language used to build workflows subsequently executed on Hadoop-based computer resources. This language should be used when the analysis workflow is well suited to the new MapReduce computing paradigm.

3C. *Extend analysis pipelines to include proteomic, RNA degradome, and epigenetic datasets.*

3D. *Extend semantic metamodels to incorporate proteomic, RNA degradome, and epigenetic data.*

Task 4. Provide an easy-to-use user interface that supports both plant biologists and plant bioinformaticists.

4A. *Develop a graphical user interface access to the data.*

4B. *Develop an application programming interface to the data.*

A RESTful programming interface along with programming examples and documentation should be delivered and made available at a public website. Programming examples should cover a few of the popular programming languages in the bioinformatics community.

4C. *Provide a graphical user interface for constructing and executing on-the-fly analyses.*

4D. *Provide an application programming interface for constructing and executing on-the-fly analyses.*

Resources

Plant 2: Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Analysis, Comparisons, and Modeling

Table 3.3 Hardware Resources for Plant 2

Hardware Purpose	Type	Size
Data management	Storage	1 to 10 petabytes
Data analysis	Processing	Medium (100 to 1000 cores)

Plant 2: Assemble Regulatory Omics Data for Target Plant Species in Common Platforms To Enable Analysis, Comparisons, and Modeling

Table 3.4 Staffing Resources for Plant 2

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; IT = Information technology)

Task or Subtask	Expertise	Duration (Months)
1. Establish a reference plant genome platform starting with six foundational genomes and with capabilities for visualizing and comparing genomes, recognizing orthologs and parologs, and automating curation of reference genomes.		
1A. Develop a platform and methods for better comparing plant genomes.	Bfx B	0–36
1B. Establish a curatorial process and third-party curation tools for continual improvement.	Bfx	12–60
2. Develop a platform for access to consolidated data.		
2A. Develop standards and methods for locating, transporting, storing, and retrieving plant omics data.	Bfx IT	1-6, plus ongoing enhancements
2B. Develop appropriate semantic metamodels to apply to omics data.	B CS	1-60, ongoing activity

Plant 2: Assemble Regulatory Omics Data for Target Plant Species in Common Platforms To Enable Analysis, Comparisons, and Modeling

Table 3.4 Staffing Resources for Plant 2

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; IT = Information technology)

Task or Subtask	Expertise	Duration (Months)
3. Extend the platform to support the generation of pre-computed and on-the-fly analyses of plant omics datasets.		
3A. Develop a configurable pipeline(s) to analyze RNA sequencing reads.	Bfx SE CS	1–36
3B. Develop appropriate semantic metamodels to apply to pre-computed analysis results and to the more stable on-the-fly analyses.	B CS	1–36
3C. Extend analysis pipelines to include proteomic, RNA degradome, and epigenetic datasets.	Bfx SE CS	36–60
3D. Extend semantic metamodels to incorporate proteomic, RNA degradome, and epigenetic data.	B CS	36–60
4. Provide an easy-to-use user interface that supports both plant biologists and plant bioinformaticists.		
4A. Develop a graphical user interface access to the data.	SE	1–60
4B. Develop an application programming interface to the data.	SE	1–60
4C. Provide a graphical user interface for constructing and executing on-the-fly analyses.	SE	1–60
4D. Provide an application programming interface for constructing and executing on-the-fly analyses.	SE	1–60

System Releases

Release 1 (expected in the 1- to 2-year time frame). Classical transcriptomic data, small RNA, and basic proteomic data would be assembled

Release 2 (expected in the 2- to 4-year time frame). Epigenetic data, small RNA target and RNA degradome data, other types of RNA processing data, and additional proteomic data will be available with a user-friendly user interface and be downloadable for power users.

Release 3 (expected in the 3- to 5-year time frame). This period would include an API and associated toolkit that provides developers with a solid resource to program against.

4. Near-Term Metacommunity Science Needs Supported by Kbase

The first objective in the metacommunities science area is to determine the metabolic role of each organism residing in a community and understand which community features provide robustness to environmental change. This will lead to improved characterizations of microbial community physiology, which are necessary to design strategies to accelerate or ameliorate microbial activity for environmental remediation.

Another reason to study microbial communities is to discover novel functions and genes within them, which is the goal of the second objective. Data generated in large-scale metagenomics projects can provide the information necessary to better understand the function of poorly characterized genes. The resulting data provide actionable hypotheses about the function of many genes that have yet to be studied in detail. Additionally, scientific efforts associated with this objective will lead to the discovery of new genes that perform useful biological functions relevant to DOE priority areas such as energy production, carbon cycling and biosequestration, and environmental remediation.

Metacommunities Scientific Objective 1

4.1 Model Metabolic Processes within Microbial Communities

Summary of Scientific Objective and its Requirements

Relevance

An overarching need is to determine the metabolic role of each organism residing in a community to understand which community features provide robustness to environmental change. Community members can be highly abundant, rare, or hidden players, and determining which organisms are involved in which processes is part of this objective.

Scientists need to be able to integrate different types of experimental measurements relating to the metabolic activity of different microbial communities in microbiomes relevant to DOE missions in bioenergy production, environmental remediation, and carbon cycling. This information is necessary for (1) generating hypotheses about the nature of interactions among community members and interactions between the community and local environment, (2) generating hypotheses about the organisms or pathways responsible for the community's metabolic activities, and (3) predicting how the community will respond to environmental changes or the introduction of new microorganisms. The ability to understand and compare communities, including those that vary spatially and temporally, also will be essential to building community metabolic models and requires tools for comparative community analysis.

Objective

This objective focuses specifically on modeling the metabolic processes within a microbial community, which ties directly into developing metagenomics workflows and systems biology tools. This predictive understanding of communities will progress in three stages.

1. **Understanding.** Descriptive models that provide insight into the metabolic role of community members and their interactions.
2. **Prediction.** Predictive models that allow us to simulate a community's metabolic processes and the response of community activity or composition to environmental conditions.
3. **Manipulation.** Eventually, these models will allow us not only to predict, but actively drive changes in the community in desired directions (e.g., accelerate processes such as environmental remediation, cellulose degradation, or carbon sequestration).

As a first step, Kbase will need to develop workflows to analyze metagenomes and other data from microbial communities and leverage existing data to create community metabolic models.

This is a near- to mid-term objective that would require leveraging existing metagenomic databases (e.g., BioCyc and KEGG) and analysis tools [e.g., Integrated Microbial Genomes with Microbiome samples (IMG/M), Metagenome Rapid Annotation using Subsystem Technology (MG-RAST), and Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA)]. Development of new analysis tools also would be required. Clear and achievable near- and mid-term goals were formulated both for top-down (metagenomics) and bottom-up (multispecies models) approaches. A mockup integration of these two approaches can be achieved in the near term, but full integration into a single analysis workflow is a mid-term task. Extending this basic modeling paradigm to integrate additional data types and tackle spatial and temporal variation is a mid- to long-term goal. Fully leveraging the predictive understanding of these communities to guide and control them is a long-term goal. At all stages of this process, the availability of relevant simplified communities (whether artificial co-cultures, low-complexity natural communities, or enrichments) should significantly accelerate tool development and allow a gradual buildup to more complex communities.

Potential Benefits

Single microbial strains rarely, if ever, act alone, and it is the complex network of interactions among microbial populations that drives all of the major metabolic processes in the world around us. These proposed objectives will lead to improved characterizations of microbial community physiology and ecology; such characterizations are necessary to design strategies to either accelerate biotransformational activity (e.g., uranium bioremediation) or ameliorate the outcome (e.g., acid mine drainage). Understanding metabolic interactions and the substrate preferences of relevant organisms is anticipated to assist in developing design strategies to optimize biotransformational activity. If successful, this understanding could provide a framework for analyzing microbial physiology in any impacted environment and lead to lower treatment costs as well as accelerated removal strategies. Developments in microbial community understanding also will directly benefit the understanding of plants and their associated microbiota, an area of immediate interest to DOE and USDA.

Synergies with Other Projects and Funding Agencies

Existing metagenomic analysis tools such as IMG/M, MG-RAST, or CAMERA currently provide some of the initial preprocessing needed for the analysis presented here, including genome

assembly and functional annotation. However, none of them currently provides satisfactory phylogenetic binning tools, or more importantly, the powerful systems biology analysis tools necessary to take functional analyses to a higher level. Platforms such as Pathway Tools include inference engines to predict pathways from potentially incomplete data (Dale, Popescu and Karp 2010) or fill holes in predicted pathways (Green and Karp 2004), but they are not adapted to the noise and incompleteness inherent in metagenomic data. Several databases funded by federal grants (e.g., BioCyc and KEGG) have some of the components necessary for the metabolic modeling parts of this objective's workflow, but there is no clear integrated database and simulation effort. Leveraging existing databases would be useful in accelerating these development efforts. There may be potential overlap with some of the National Institutes of Health's (NIH) human microbiome projects (although probably more with metagenomics than with metabolic modeling), which will result in a large amount of data relating to the structure and activities of microbial communities that interact with their human host. Some of the computational and experimental methods being developed for those projects could be applicable to some of the datasets and analysis envisioned for Kbase.

Several existing BER experimental programs explore a wide variety of metagenomic studies in diverse environments (e.g., acid mine drainage, enhanced biological phosphorus removal, termite gut, rumen, compost, soils, permafrost, oceans, and sediments). In many of these processes, the biotransformational activity is related to the integrated phenotype of microbes present in the community. To enhance these biotransformational activities, it is important to characterize the metabolic pathways of constituent members and link the individual organisms to their substrate and product profiles. Such projects would be leveraged as first adopters or beta testers. These and other groups would be needed to help define the minimum feasible metadata for metagenomic samples.

Illustrative Workflow

Workflows for constructing metabolic models from an individual organism's genome sequences have been developed (Thiele and Palsson 2010). Although many of the steps for generating metabolic models for microbial communities may be similar, missing information (such as unsequenced genes) may be a more challenging problem when dealing with metagenomic datasets. Workflows were developed for the bottom-up microbial community modeling framework and for the top-down metagenomic analysis method. The inputs, outputs, and tools for each are provided in Appendix C, Supporting Scientific Objective and Software Requirement Documents for Near-Term Metacommunity Science Needs. Such workflows currently are scattered, and integrating them into a common framework and database would be required.

The first step in analyzing defined communities will involve incorporating all individual metabolic models into a common environmental model. This will require information on the substrates metabolized and secreted by community members as well as a common nomenclature for the exchanged metabolites (Zhuang et al. 2010). Additionally, the kinetics of substrate uptake and secretion as well as biomass yields will be critical to develop such community models. This first step is a key requirement before more complex communities can be studied.

Characterization of environmental microbial physiology can proceed through two broad approaches: (1) the bottom-up approach in which microbes are isolated and cultured in the laboratory and integrated, evaluated, and modeled in a defined community and (2) the top-down, metagenome-based approach in which DNA from environmental samples is directly sequenced for understanding the metabolic potential through bioinformatics and pathway reconstruction. See Fig. 4.1, below, for a simplified version of this workflow.

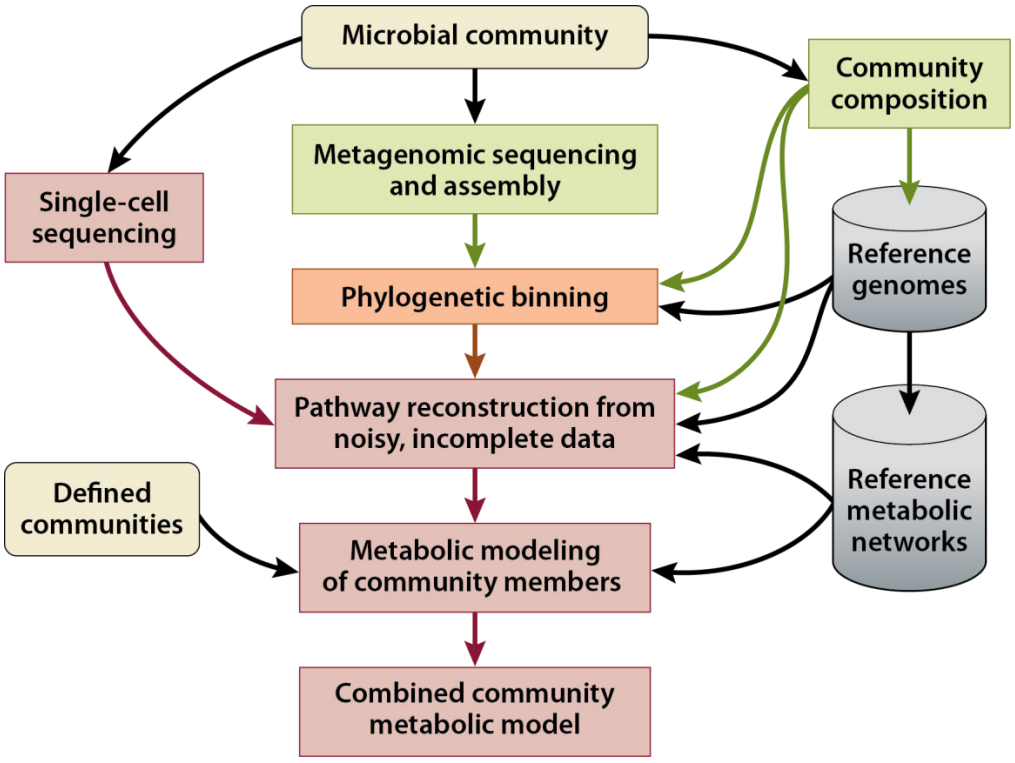


Fig. 4.1. Workflow for Reconstructing Metabolic Models from Metagenomic Data. The "bottom-up" approach involves data from individual species or members, and the "top-down" approach relies on whole-metacommunity analysis. In addition to available genomic sequencing technologies, Kbase must prepare to use other data in the near future. Green modules already are reasonably well established. The orange box shows an available method that needs more development. Red boxes show capabilities that still require significant new development.

The end result is that researchers will be able to access the data collected in this workflow and their own data via interfaces for data upload, integration, and visualization of metabolic pathways. They also would be able to perform simulations via web interfaces; use and develop tools to predict, visualize, and compare community responses to experimental data; perform queries for model and network comparison; and perform queries across metabolic models and pathways, such as reachability analysis from one metabolite to another across species boundaries. Other valuable advancements would be development of tools for simultaneously visualizing simulation results and experimental data and methods to flag conflicts between datasets. Comparisons among community models would include clustering representations (such as trees and PCA plots) and systems representations (such as metabolic maps). Additional

inferences would include the ability to visualize predicted fluxes through metabolic networks and compare them with genome-wide omics data.

Implementation Plan for Modeling Metabolic Processes within Microbial Communities

System Capabilities

The scientific objective is to be able to integrate different types of experimental measurements relating to the metabolic activity of different microbial communities in microbiomes relevant to bioenergy production, environmental remediation, and carbon cycling and biosequestration. This integration is necessary to understand the nature of interactions among community members and between the community and its environment, to understand the organisms and pathways responsible for the community's metabolic activities, and to predict how a community will respond to environmental changes.

Kbase would provide capabilities to discover, query, access, and integrate required experimental measurements. Such measurements include metagenome sequence data; environmental conditions (e.g., available nutrients, extracellular metabolite profiling, carbon, nitrogen, phosphorus, oxygen, pH, temperature, and light); temporal and spatial measurements; transcriptomic, proteomic, metabolomic, and microbial physiological data; and stable isotope probing—all held at Kbase-associated organizations and institutes or within Kbase itself. Furthermore, it would provide access to analysis and modeling tools, flexible workflows, computational and data storage facilities to enable metagenomic sequence assembly, and phylogenetic analysis and metabolic modeling of microbial communities, which are necessary to determine interactions and metabolic activity drivers and to predict responses to environmental changes.

The effort will leverage and integrate with existing resources, including:

- Pathway data repositories (e.g., BioCyc and KEGG).
- Pathway inference engines (e.g., Pathway Tools).
- Ontological data descriptions [e.g., GO, Ontology for Biomedical Investigations (OBI), and the National Center for Biomedical Ontology (NCBO)].
- Semantic search, query, and access (e.g., Bio2RDF).
- Metagenomic analysis tools (e.g., IMG/M, MG-RAST, and CAMERA).
- Community diversity analysis tools (e.g., DOTUR, Mothur, UniFrac, and Primer).
- Workflow development (e.g., Kepler and Taverna), sharing, and execution (e.g., MyExperiment, Galaxy, and CAMERA).

Tasks

Task 1. Providing a common platform.

The subtasks below will provide capabilities useful across all subsequent tasks in this implementation plan for Metacommunities 1. A range of these more generic and required tools also are envisaged to be useful for achieving other scientific objectives.

1A. *Identify essential resources and analysis tools.*

Identify which tools can be ported to or reimplemented in Kbase, which can be called programmatically from within Kbase, and which can only be integrated with the Kbase by providing data exchange routines.

- Tools for processing metagenomic sequence data, including assembly methods and quality filters.
- Phylogenetic binning methods.
- Community diversity analysis tools, such as ARB/Silva, greengenes, Ribosomal Database Project, DOTUR, Mothur, UniFrac, and Primer.
- Larger metagenome annotation tools and workflows, such as IMG/M, MG-RAST, and CAMERA.
- Metabolic inference tools.
- Flux balance analysis tools, including network "debuggers" and simulators such as COBRA, MetaFluxNet, and CellNet Analyzer.

1B. *Develop a repository of essential tools and workflows (or access to them).*

This repository would include descriptive information for all tools and workflows, the means to search and compare them, and some way to capture usage patterns and opinions from the Kbase user community. Needed metadata for these resources include their general purpose; main area of application; special requirements; integration into existing workflows; and quality mark indicating, for example, whether the tool or workflow has been tested, is widely used, or newly developed. This repository could be created by customizing widely used open-source packages (e.g., openwetware.org, github.com, or GForge.org). A similar strategy was used to develop an imaging tools repository for NIH: the Neuroimaging Informatics Tools and Resources Clearinghouse (nitrc.org).

1C. *Provide validation and characterization methods for tools used for assembly, binning, pathway reconstruction, and other metagenome analyses.*

Metagenomic datasets vary dramatically in terms of, for example, complexity, guanine-cytosine (GC) content, and sequencing read length. The focus therefore should not be to validate each tool on a small set of "typical" datasets, but rather to characterize the range of datasets on which it works

best. Develop the infrastructure to simplify cross-validation by restricting what each tool sees as its known reference dataset.

1D. *Develop an environment facilitating easy discovery, assessment, and access to key data sources.*

Essential for metabolic modeling are easy access to the best and most complete experimental measurements and subsequent data analysis that is relevant to the specific research work. To accomplish this, particularly in the context of more automated workflows, it will be necessary to develop:

- Common access mechanisms to data sources (API, query terms, semantic mapping and ontologies, protocols).
- Common descriptive metadata and annotation formats.
- Common data and analysis description (i.e., workflow) formats.
- Clearinghouse of data sources and their content.

1E. *Develop a workflow environment (repository, shared development, execution) and a common tool platform for* ad hoc *experimentation and workflow development.*

Researchers will need to be able to experiment with combinations of analysis tools. They also will need the ability to develop, store, share, and execute commonly used biology workflows utilizing a variety of workflow systems (including Kepler, Medici, and Taverna) built for the essential tools and data sources (linking to the tools repository and data clearinghouse). Workflows must be discoverable and sharable to allow the community to develop its understanding and the best possible analysis methods, as well as share knowledge of their best usage between different groups. Furthermore, scientists need to be able to modify and execute developed workflows. To provide researchers with complete information about workflows and a history of their utilization, a core element of this environment will be a repository for the provenance from executed workflows (i.e., the history of workflow runs). The solution should leverage the experiences of existing platforms, such as MyExperiment, CAMERA, openwetware.org, and Galaxy, for biological workflow sharing.

1F. *Provide computational and intermediate storage resources.*

There will be a growing demand for computational analysis tasks, many either in tightly coupled workflows or more loosely connected in research collaborative efforts. In either case, Kbase needs to provide access to computational resources and to intermediate storage space in the system to allow the sharing of workflows and interim results for further analysis. We estimate that 100+ terabytes of intermediate and long-term storage will be

required, based on experiences with biological modeling applications at worldwide high-performance computing (HPC) centers and with collaborative environments supporting scientific workflows in biology (e.g., the Biomedical Informatics Research Network, CAMERA or MyExperiment). Specific requirements will be driven by the uptake rate of the community. Based on prior experiences, requirements for computational resources are estimated to be on the order of 2000 cores. We currently do not expect to need high-speed interconnects, but rather systems that can cope with data-intensive applications. This infrastructure should be deployed and integrated with tools being developed under the other subtasks. Ongoing infrastructure development is critical to ensure proper operation of the environment. Several of these tasks for metabolic modeling of communities can be effectively parallelized. Hence, a computing module based on graphics processing unit (GPU) could be valuable.

1G. *Develop and maintain curated data repositories.*

Many research results are anticipated to be created within Kbase. Scientists will be able to contribute some of these results back to the community through existing repositories. However, there will also be a demand for publishing and sharing data. For example, when a scientist computationally predicts a novel metabolic pathway in a community metagenome, he or she would share with the research community these results and the analysis that led to them. This predicted pathway could be supported by experimental measurements but may also include holes. By maintaining a repository, the pathway could be added to or modified by the community, perhaps leading to validation and acceptance in existing metabolic pathway databases. Developing curated repositories within Kbase itself therefore would be very useful.

Resources

Metacommunities 1: *Model Metabolic Processes within Microbial Communities*

Table 4.1 Hardware Resources for Metacommunities 1

Hardware Purpose	Type	Size
Data management	Storage	100 terabytes to 1 petabyte
Data analysis	Processing	Large (2000 cores)

Metacommunities 1: *Model Metabolic Processes within Microbial Communities*

Table 4.2 Staffing Resources for Metacommunities 1:

Milestones Task 1

(SE = Software engineering; Bfx = Bioinformatics; IT = Information technology; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
1. Providing a common platform		
1A. Identify essential resources and analysis tools.	Bfx	0–6
1B. Develop a repository of essential tools and workflows (or access to them.) (Repository implemented as part of the Infrastructure development effort).	SE, Bfx	3–24 12 months initial version; 24 months production version
1C. Provide validation and characterization methods for tools used for assembly, binning, pathway reconstruction, and other metagenome analyses.	SE	12–24
1D. Develop an environment facilitating easy discovery, assessment, and access to key data sources. This includes:		
• Initial common access mechanisms to data sources and a clearinghouse of data sources.	Bfx, IT, SE	0–6
• Plan for agreement of common descriptive metadata and annotation format and data formats.	Bfx, IT	0–6

Metacommunities 1: *Model Metabolic Processes within Microbial Communities*

Table 4.2 Staffing Resources for Metacommunities 1:

Milestones Task 1

(SE = Software engineering; Bfx = Bioinformatics; IT = Information technology; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
• Develop commonly agreed descriptive metadata and annotation format and data formats for all resources.	B, Bfx, IT	6–60
• Production-level clearinghouse of all relevant data sources and their content.	Bfx, IT, SE	6–24
• Provide common access mechanisms to data sources.	Bfx, IT, SE	6–60 (declining effort profile)
(Major IT components done as part of the Infrastructure development effort.)	B, Bfx, IT	24–60
1E • Develop a workflow environment (repository, shared development, execution). • Develop a common tool platform for *ad hoc* experimentation and workflow development. (Development of a workflow system is a major part of the Infrastructure development effort.)	Bfx, IT, SE IT, SE	3–24 3–36 (prototype available at 12 months)
1F. Provide computational and intermediate storage resources (Infrastructure).	B, IT, SE	0–60
1G. Develop and maintain curated data repositories.	B, Bfx, IT, SE	3–60 12 months initial repository; after that, ongoing development and maintenance effort

Task 2. Metagenomic sequence data processing and assembly.

2A. *Identify sources of metagenomic sequence data and provide integrated discovery of and access to them.*

As sequencing technologies proliferate and become increasingly affordable, the number of sources for metagenomic sequence data is growing worldwide. Providing access to high-density coverage and comparable quality sequencing data for the studied communities is essential for subsequent phylogenetic analysis and metabolic modeling of them.

2B. *Determine additional or future needs for assembly tools for metagenomic data.*

- Implement or provide access to assembly tools.
- Develop or implement new assembly tools as sequencing technology evolves.

Because sequencing technologies continue to evolve, assembly methods must as well. Current tools have been adapted or developed for 454, Illumina, and SOLiD data, but as technologies from Pacific Biosciences, Helicos, and others come online, these tools will likely require further development and modification. Current assembly methods tend to be computationally memory-intensive, and we will need to determine whether assembly remains memory-intensive or other computational challenges arise with new sequencing technologies.

Metacommunities 1: *Model Metabolic Processes within Microbial Communities*

Table 4.3 Staffing Resources for Metacommunities 1:
Milestones Task 2

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
2. Metagenomic sequence data processing and assembly		
2A. Identify sources of metagenomic sequence data and provide integrated discovery of and access to them.		
• Identify sources of metagenomic sequence data.	Bfx	0–3
• Provide integrated discovery and access to the identified data sources (Infrastructure).	CS	3–60 6 months initial access to key resources 18 months semantic access to wider selection of data resources 36 months tools for self-registration of data sources from institutes and research groups 36–60 months Ongoing support in integrating new data sources semantically into Kbase
2B. Determine additional or future needs for assembly tools for metagenomic data.		
• Implement or provide access to assembly tools.	SE	0–6
• Develop or implement new assembly tools as sequencing technology evolves.	SE	6–60

Task 3. Phylogenetic analysis.

3A. *Make microbial ecology tools designed to analyze community diversity available to scientists and provide through Kbase the means for continually developing or integrating them with metabolic modeling efforts.*

These tools use phylogenetic methods, usually based on 16S/18S rRNA sequence data. They are essential to rapidly sample communities and correlate community diversity with enviromental parameters, thereby associating metabolic phenotypes with species ecotypes or guilds.

3B. *Develop, validate, and combine phylogenetic binning methods into an integrated workflow and quantify uncertainty and address its propagation.*

A number of phylogenetic binning methods have already been developed, but more effort is needed in validating their performance and characterizing under which circumstances they perform best. Assembling multiple binning methods into a single binning workflow may allow us to combine the best features of each, since different binning approaches may be optimal for different contigs in a metagenome sequence, depending on contig length, presence of phylogenetic markers, or availability of a close reference genome. Metagenomic data are inherently far more noisy and incomplete than single genome sequences. As such, the uncertainty associated with factors like bin assignment, number of strains in a bin, and incomplete coverage of the genomes needs to be quantified and taken into account in downstream analyses as much as possible.

3C. *Implement example workflows for phylogenetic analysis, covering some minimal set of analysis steps to be applied to a typical microbial community.*

This task would involve, for example, combining pyrotag sequencing, clustering, and identification of operational taxonomic units (OTUs) with UniFrac ordination of community composition and then correlating this with salient environmental parameters (e.g., the Mothur wiki provides written workflow descriptions). Another standard workflow for metagenomic sequence data would include sequence assembly, quality control, phylogenetic binning (e.g., based on standard operating procedures used by IMG/M), and analysis of functional categories in each bin.

Metacommunities 1: *Model Metabolic Processes within Microbial Communities*

Table 4.4. Staffing Resources for Metacommunities 1:

Milestones Task 3

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
3. Phylogenetic analysis		
3A. Make microbial ecology tools designed to analyze community diversity available to scientists and provide through Kbase the means for continually developing or integrating them with metabolic modeling efforts.	SE	0–12
3B. • Develop, validate, and combine phylogenetic binning methods into an integrated binning workflow (Infrastructure will provide workflow service).	Bfx	12–36 (could start as soon as binning methods are available on the tool platform or earlier in a more *ad hoc* fashion)
• Quantification and propagation of uncertainty.	S	36–60 (ongoing effort)
3C. Implement example workflows for phylogenetic analysis, covering some minimal set of analysis steps to be applied to a typical microbial community.	Bfx	3–18

Task 4. Metabolic modeling of community members.

4A. *Identify and provide required resources (e.g., KEGG, MetaCyc, BioCyc, SEED) for integrated data discovery, query, and access to enable the assembly and update of reference datasets as well as other uses of data from these resources.*

It will be necessary to provide integrated discovery, query, and access capabilities across the different data resources for both scientists and automated tools.

4B. *Adapt or develop novel pathway inference methods that can handle noisy and incomplete data and implement example workflows.*

These example workflows will demonstrate metagenome sequence assembly, annotation, phylogenetic characterization, and prediction of metabolic pathways of community members. This effort will leverage and integrate with existing resources, including:

- Pathway data repositories (e.g., BioCyc and KEGG).
- Pathway inference engines and resources (e.g., Pathway Tools and BioPAX).
- Ontological data descriptions [e.g., GO, sequence ontology (SO), Chemical Entities of Biological Interest (ChEBI), OBI, and NCBO's BioPortal).
- Semantic search, query, and access (e.g., Bio2RDF).
- Metagenomic analysis tools (e.g., IMG/M, MG-RAST, and CAMERA).
- Community diversity analysis tools (e.g., DOTUR, Mothur, UniFrac, and Primer).
- Workflow development (e.g., Kepler and Taverna), sharing, and execution (e.g., MyExperiment, Galaxy, and CAMERA).

4C. *Assemble a reference dataset of microbial phenotypes and metadata.*

This task involves developing, adopting, and promoting a standardized vocabulary or ontology for microbial phenotypes and other metadata. The assembled comprehensive set of phenotypes should include a large and varied set of reference genomes, at the very least one species per phylum but preferably one per genus. Kbase also should incorporate metabolic, physiological, and morphological phenotypes used to identify species (e.g., from Bergey's "differential characteristics" tables).

4D. *Assemble and maintain a reference dataset of metabolic reconstructions*.

Develop standardized formats for pathway representation and unique identifiers and cross-references for all metabolites, reactions, and enzymes.

- Genome content (e.g., enzymes and transporters).
- Pathway content (e.g. from KEGG, SEED, or BioCyc).
- Available experimental data, including omics, but also biomass composition, detected metabolites, and enzymatic activities.
- Flux balance analysis (FBA) models of available reference organisms.

Metacommunities 1: *Model Metabolic Processes within Microbial Communities*

Table 4.5 Staffing Resources for Metacommunities 1:
Milestones Task 4

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
4. Metabolic modeling of community members		
4A. Identify and provide required resources for integrated data discovery, query, and access to enable the assembly and update of reference datasets as well as other uses of data from these resources.		
• Identify required data resources.	Bfx	0–3
• Provide integrated discovery and access to the identified resources (Infrastructure).	CS	3–60
4B.		
• Adapt or develop novel pathway inference methods that can handle noisy and incomplete data (with Section 2.1, Task 2A).	Bfx S	6–48 (prototype ready at 24, 36 months)
• Implement example workflows (with Infrastructure).	Bfx	12–36
4C. Assemble a reference dataset of microbial phenotypes and metadata (with Section 2.1, Task 1C).	Bfx	0–18 key phenotypes and metadata selected at 6 months
4D		
• Assemble a reference dataset of metabolic reconstructions.	Bfx	0–24
• Develop standardized formats for pathway representation and unique identifiers (with Section 2.1, Tasks 1A and 4A).	Bfx	0–36 (initial standards at 6, 12 months)
• Maintenance of reference datasets.	Bfx, SE	Ongoing, 6–60

Task 5. Metabolic modeling of the community.

5A. *Identify known physiological data pertaining to members of a community as a first step in modeling its metabolic processes.*

Representing metabolites that are potentially exchanged among community members and the environment will be important. In addition, access to databases containing information on the known physiology of microbes, including substrate uptake kinetics, will be critical for individual community members. Finally, methods will be needed for representing the relevant biological objective and constraints suitable for modeling growth as well as intracellular and intercellular flux distributions.

5B. *Develop methods to model the metabolic interactions of species in a community and the response of the community to perturbations and changes over time and space.*

Simulation frameworks should be capable of incorporating models of individual organisms into a community model able to integrate customized workflows for simulation purposes. Essential for modeling interactions among community members are (1) access to tools for functional annotation of transporters (e.g., TransportDB's Transporter Automatic Annotation Pipeline) and (2) incorporation of experimental data on extracellular metabolites and three-dimensional spatial organization of the community.

5C. *Provide HPC resources for simulating large multispecies models, conducting Monte Carlo sampling of alternative metabolic reconstructions from noisy and incomplete metagenomic data, and for performing dynamic simulations in which the concentration levels of extracellular metabolites or the abundance of individual community members may change over time.*

These types of simulations may require several orders of magnitude more CPU cycles than solving a typical single-genome FBA model. However, these simulations can be carried out in parallel and hence could benefit from GPU computing modules.

5D. *Develop hierarchical or multiscale visualization tools for multispecies metabolic models.*

Existing visualization tools typically represent a metabolic network at one of two levels: the whole-genome network or individual pathways. Visualizing the metabolic network even for a single organism quickly becomes overwhelming, let alone for a community with a dozen organisms. New methods are needed to abstract key metabolic processes in each community member so that a useful whole-community overview can be achieved. This is not merely a visualization task, but rather needs to be closely integrated with metabolic network analysis

to identify the key pathways and fluxes in each organism that are relevant to the functioning of the overall community. We will also need to be able to map any available omics data or computational predictions onto this whole-community visualization.

Metacommunities 1: *Model Metabolic Processes within Microbial Communities*

Table 4.6. Staffing Resources for Metacommunities 1:
Milestones Task 5

(SE = Software engineering; Bfx = Bioinformatics; IT = Information technology; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
5. Metabolic modeling of the community		
5A. Identify known physiological data pertaining to members of a community as a first step in modeling its metabolic processes.	Bfx	0–6
5B. Develop methods to model the metabolic interactions of species in a community and the response of the community to perturbations and changes over time and space.	Bfx	0–60 prototype tools available at 24, 36 months
5C. Provide HPC resources and access (Infrastructure).	IT, SE	6–12
5D. Develop hierarchical or multiscale visualization tools for multispecies metabolic models.	Bfx, SE	0–36 prototype available at 24 months

System Releases

Supporting this scientific objective effectively will require access to 100+ terabytes of intermediate and long-term storage, with specific requirements being driven by the uptake rate of the community. Furthermore, a number of data-intensive applications have been identified by the community that will require specific computational resources. The overall computer requirements are difficult to estimate and will vary widely among different computational tasks. However, we estimate that resources on the order of 2,000 cores would be needed.

Release 1 (1–6 Months).

- Identify essential data resources and analysis tools (Task 1).
- Develop initial common access mechanisms to data sources (Task 1).
- Develop common descriptive metadata and annotation formats and data formats (Task 1).
- Develop an initial clearinghouse of data sources and their content (Task 1).
- Set up initial access to computational and intermediate storage resources. This integration activity will be ongoing throughout the project (Task 1).
- Identify sources of metagenome sequence data (Task 2).
- Implement or provide access to assembly tools (Task 2).
- Identify required resources for metabolic pathway data (Task 4).
- Select key phenotypes and metadata for reference dataset (Task 4).
- Develop initial standardized formats for pathway representation and unique identifiers (Task 4).

Release 2 (1 year).

- Develop a repository of essential tools. Implement the initial version during year 1 and the production version during Release 3 (Task 1).
- Develop and maintain curated data repositories. Initial repository would be released at 12 months, with ongoing development and maintenance thereafter (Task 1).
- Provide common access mechanisms to data sources. Initial release at 12 months, with continued development and maintenance thereafter (Task 1).
- Develop a prototype of a common tool platform for *ad hoc* experimentation and workflow development (Task 1).
- Implement or provide access to community diversity tools (Task 3).
- Quantify uncertainty in metagenomic data and address its propogation. Initial release will be at 12 months, thereafter this will be an ongoing activity (Task 3).
- Update standards for pathway representation and unique identifiers (Task 4).

- Provide HPC resources and access for community-level metabolic simulations (Task 5).

Release 3 (2 years).

- Continue developing and improving commonly agreed descriptive metadata and annotation formats and data formats for key initial resources (Task 1).
- Develop a workflow environment (Task 1).
- Release the production-level clearinghouse of all relevant data sources and their content (Task 1).
- Establish a reputation or scoring system for analysis tools and tool developers, datasets, and computational results (Task 1).
- Develop an infrastructure to simplify cross-validation and characterization of tools and methods for assembly, binning, pathway reconstruction, and other metagenomic analyses (Task 1).
- Implement prototype (example) workflows for phylogenetic analysis (Task 3).
- Adapt or develop a prototype pathway inference method that can handle noisy and incomplete data (Task 4).
- Assemble a reference dataset of metabolic reconstructions (Task 4).
- Assemble a reference dataset of microbial phenotypes and metadata (Task 4).
- Add the ability to run a basic descriptive multispecies metabolic model of natural communities based on the previously described integrated pipelines (Task 5).
- Prototype visualization tools for multispecies metabolic models (Task 5).

Release 4 (3 years).

- Improve the common tool platform for *ad hoc* experimentation and workflow development (Task 1).
- Provide integrated discovery of and access to sources of metagenome sequence data (in release 4 and each subsequent release) (Task 2).
- Combine phylogenetic binning methods into an integrated binning workflow (Task 3).
- Develop an intermediate pathway inference method that can handle noisy and incomplete data (Task 4).
- Implement example workflows demonstrating metagenome sequence assembly, annotation, phylogenetic characterization, and prediction of metabolic pathways of community members (Task 4).
- Continue developing commonly agreed standards for pathway representation and unique identifiers (Task 4).

- Perform ongoing maintenance and automatic updating of reference datasets (Task 4).
- Implement production-level visualization tools for multispecies metabolic models (Task 5).

Release 5 (5 years).

- Continue developing commonly agreed descriptive metadata and annotation formats and data formats (Task 1).
- Continue developing pathway inference methods that can handle noisy and incomplete data (Task 4).
- Provide integrated discovery and access to existing pathway data sources (Task 4).
- Perform ongoing maintenance and automatic updating of reference datasets (Task 4).
- Improve capabilities for descriptive multispecies metabolic modeling of natural communities based on the previously described integrated pipelines (Task 5).

Release 6 (10 years).

- Incorporate other networks, including regulatory, signaling, and intercellular interaction (Task 1).
- Integrate predictive metabolic models with models that incorporate spatial and temporal distribution of metabolic activity (Task 1).
- Provide capabilities for multispecies interacting metabolic modeling that predicts response to perturbation (for the purpose of environmental remediation or other desirable functional behavior) (Task 1).

Metacommunities Scientific Objective 2

4.2 Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function

Summary of Objective and its Requirements

Relevance

One reason to study microbial communities is to determine novel functions and genes within these communities. Reliable functional annotations are critical prerequisites of a successful research program in systems biology. This objective will potentially accelerate efforts aimed at characterizing the function of currently understudied genes. Additionally, the tools developed as part of this project will be a valuable asset to scientists generating new datasets by allowing them to leverage Kbase-associated datasets in the analysis process and to generate actionable hypotheses.

Objective

Data generated in large-scale metagenomics projects can provide the information necessary to better understand the function of poorly characterized genes. As metagenomic data are rapidly coming online, a critical scientific objective is to:

- Mine the data to identify previously unknown genes and ensure that they can be tracked across datasets and databases.
- Leverage the wealth of metadata associated with metagenomic datasets, as well as gene-organism co-occurrence information to identify testable hypotheses about the function of newly identified or poorly characterized genes.

In the longer term, more complex analyses could be applied, such as using various differential equation models to analyze longitudinal data in order to understand the mechanistic interactions among genes, genes and organisms, and genes and environmental parameters.

Roughly a third of all the genes in the *E. coli* genome have no known function (Hu et al. 2009), despite the fact that this bacterium is among the best studied organisms. Although scientists are slowly elucidating the function of some of these genes (Weber et al. 2010), their efforts cannot keep up with the wealth of data being generated in both traditional genomic projects and through large-scale metagenomic efforts. The magnitude of the problem is perhaps best exemplified by the number of novel protein sequences identified by the Global Ocean Sampling expedition (Yooseph et al. 2007). The authors of this study identified more than 1700 genes with no similarity to any known protein families. Efforts to understand the function of these genes cannot be effectively conducted without first prioritizing the genes on the basis of their importance to pressing biological questions. But how can we know which genes are important if we do not even know what they do?

The key to this problem lies in the metagenomic datasets themselves. Specifically, metagenomic data are not simply comprised of DNA sequences; they also contain a rich set of metadata, information linking the sequences to location (e.g., latitude and longitude, height, or depth), to physical characteristics of the environment (e.g., temperature, pH, and salinity), and

to time. Also, information is available that links together multiple metagenomic datasets (e.g., data generated from the same location at different points in time). Prioritization of experimental and annotation efforts, as well as possible hypotheses about the function of a gene or group of genes, can be derived from available metadata. For example, a particular gene might only be found in samples taken from communities known to perform a particular biological process (e.g., a gene or group of genes only found in oil-contaminated water, implying their possible role in hydrocarbon metabolism). In addition, some genes might only be found in conjunction with genes whose function is known, thereby implying their involvement in similar biological processes.

The data required to meet this scientific objective are standard, but there is a critical need for data exchange standards and ways to describe and link the data and to return data search results. To provide such an integrated system, a core requirement is to better define and incorporate metadata related to the data and to its processing. Specifically, a coordinated set of standards needs to be implemented so that the Kbase infrastructure can handle diverse types of metadata, existing standards are extensible, and a governance structure ensures that people comply with the standards.

Potential Benefits

Parts of this high-priority, near- to mid-term objective could be carried out in 1 to 3 years, other portions in 3 to 5 years. The development of methods for extracting information about gene function from metagenomic datasets and associated metadata will have far-reaching impacts on biological research in general and on DOE's mission in particular. Resulting data will provide actionable hypotheses about the function of many genes that have yet to be studied in detail. Additionally, scientific efforts associated with this objective will lead to the discovery of new genes that perform useful biological functions relevant to DOE priority areas such as energy production and environmental remediation. Improvements in identifying unknown genes and their function will help to stem the potential of error propagation in gene-calling databases. These efforts also could lead to the development of sensitive markers of ecosystem health.

Synergies with Other Projects and Funding Agencies

All of the metagenomic sequencing efforts undertaken by DOE and NIH to date could be leveraged for this scientific objective. Moreover, similar efforts are likely in other research fields that are starting to apply metagenomic methods, so potential overlap exists with projects funded by a broad range of agencies, including NIH, NSF, the National Aeronautics and Space Administration, USDA, and the Food and Drug Administration. Maintaining regular communication between DOE and these agencies will be necessary, as will active and broad dissemination of the results of work performed through Kbase.

Illustrative Workflow

One proposed workflow would have the following elements:

(1) Metagenomic sequences are assembled. (2) Genes are found within the assembled contigs and are compared to other datasets registered within Kbase and to public databases. (3) Homologies are detected, and appropriate identifiers are assigned to enable tracking the same gene across datasets. (4) A data matrix is constructed from user-selected or automatically suggested datasets. (5) Statistical computations are performed on the data matrix based on user-defined criteria and column permutations (e.g., "interesting" columns are selected based on a combination of metadata, and genes significantly enriched or depleted in these columns are identified using statistical software). (6) A graph is created of the connections among genes, genes and neighboring gene functions, genes and organisms, and genes and environmental parameters and is annotated with strength of the connection or statistical significance. (7) Resulting data can feed into new hypotheses or predictive models of gene interactions. (see Fig. 4.2, below, for a illustration of this workflow and Section C.3 in Appendix C for additional workflow details.)

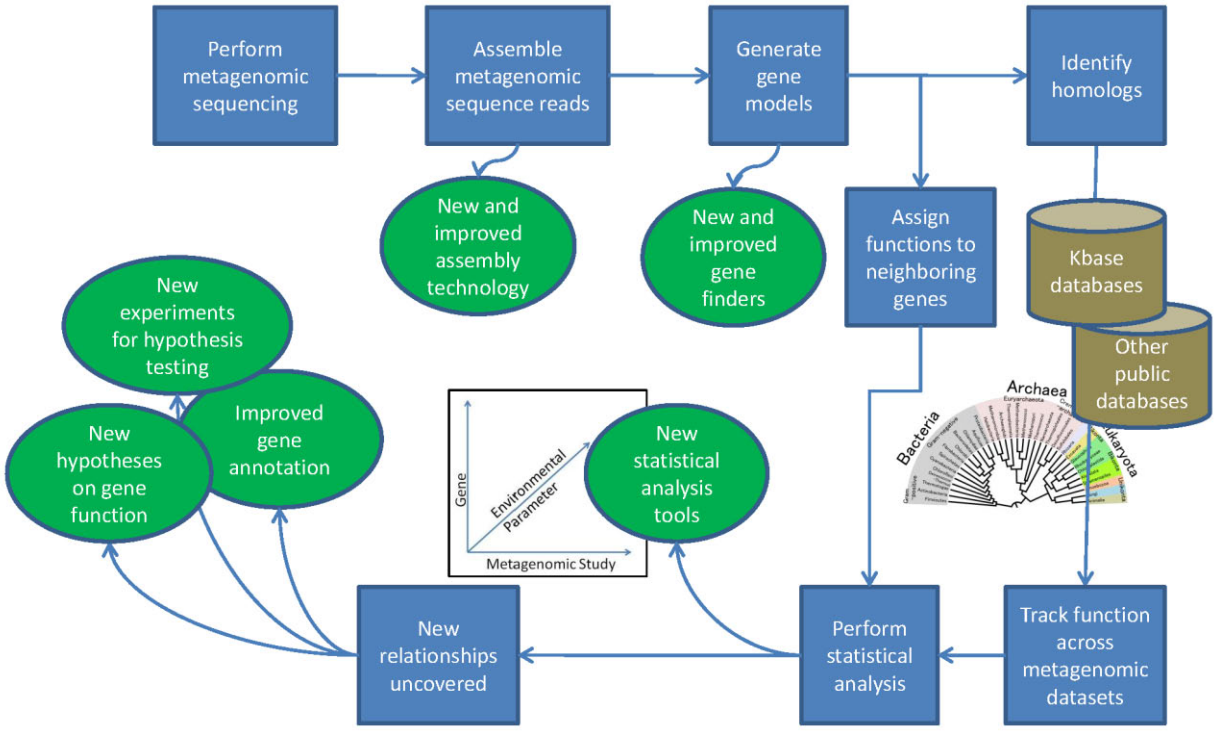


Fig. 4.2. Workflow for Mining Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function. Unknown genes from metagenomic studies may be assigned hypothetical functions based on their occurrence across a range of genomic and metagenomic datasets, the environmental and metabolic parameters associated with these data, and any functional annotations for neighboring or co-occurring genes.

Implementation Plan for Mining Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function

System Capabilities

The system capabilities needed to expand our understanding of poorly studied genes must take into account the projected advances in sequencing technology. Metagenomic projects will produce increasingly more data and become more frequent as costs per project decrease and as sequencing technologies advance. The high-level capabilities required to meet this challenge can be summarized in the development of three overarching capabilities: identifying previously unknown genes in metagenomic datasets, mining metadata to elucidate the potential biological roles of previously uncharacterized genes based on patterns of occurrence with environmental parameters, and supporting the generation of testable hypotheses about the function of newly identified or poorly characterized genes.

A compilation of known and unknown genes and the correlations of these genes with metadata will enable systematic searches for the functions of the unknown genes. These correlations will enable us to advance the scientific objective of understanding poorly characterized genes. A set of evolving consensus protocols for performing the assembly and translation from DNA reads to contigs to genes for metagenomic sequences is an essential feature of this implementation plan for Metacommunities 2. While no single method is best suited for all datasets, a set of standard protocols would improve our ability to repeat results, perform comparisons, and improve quality dramatically.

Tasks

Task 1. Develop resources for assembling metagenomic datasets into consensus sequences.

A prerequisite to gene prediction is a good consensus sequence generated during the assembly process. Because gene prediction methods generally rely on sequence composition, the longer the consensus sequence, the more accurate the gene predictions become. Current approaches to assembling metagenomic sequencing reads involve a binning phase. When assembling a large metagenomic dataset, the GC content varies significantly, enabling binning based on sequence composition. However, not all bacterial species are easily distinguished based on GC content alone. Improvements in binning and assembly methods that deal with closely related species are needed.

1A. *Provide quality control and quality filtering on sequence read datasets.*

Sequence data need to be normalized for low-quality regions and artifacts (technical replicates) that complicate downstream analysis. The establishment of community consensus on a small number of protocols suited to the various data types is a key deliverable of this task. A strong emphasis should be placed on integrating existing open-source methods.

1B. *Improve the binning phase of the assembly process to utilize information about the distribution of closely related strains and species in the metagenomic dataset and integrate the binning and assembly processes more closely.*

Preprocessing metagenomic sequence reads to obtain information about the distribution of reads based on sequence composition characteristics can inform the binning and assembly process. Metagenomic samples will vary in complexity, ranging from a few to thousands of organisms. Understanding the complexity prior to binning and assembly can inform the selection of parameters used during these processes. Several protocols exist for binning sequences prior to assembly. The best of these protocols function by closely integrating both the binning and assembly processes. Establishing community consensus on a small number of protocols suited to the various data types is a key deliverable of this task.

1C. *Improve the assembly phase of the assembly process to produce a pan- or core genome that is thought to be representative of bacterial taxa at various taxonomic levels.*

A longer-term deliverable of improving the process by tuning the binning and assembly based on the desired taxonomic granularity will enable investigators to refine their assembly based on sample complexity and the scientific question being addressed. Improvements to the binning and assembly phases will be required so that concepts like a pan-genome or core genome can drive the binning and assembly processes.

1D. *Develop a model for representing polymorphisms when assembling multiple taxa (strains, species, and genera) into a single consensus sequence.*

Functional specificity can be influenced by single changes in a sequence. When looking for common or unique functions across different environments, understanding the key structural positions of proteins and if these positions are polymorphic will influence the quality of the correlations. Data structures that capture, persist, and make available polymorphism information are the key deliverable of this subtask.

1E. *Extend the assembly resource to include meta-RNA sequence datasets*.

Using RNA sequence data will represent a more accurate picture of which functions are active in the community. These data can be treated as a *de novo* assembly problem or in a manner similar to existing fragment recruiting approaches. A unique feature of RNA data that impacts the assembly process involves character composition. Approaches to binning will need to be evaluated to understand the differences between genome and RNA sequence data.

Task 2. Improve gene-finding algorithms.

The function of gene-finding works best with whole genes, however partial genes will result from metagenomic data. The community at large has yet to adopt a standard approach to gene identification in metagenomic sequence data. Several gene prediction methodologies for metagenomic sequences exist, including Metagene, MetaGeneAnnotator, FragGeneScan, MetaGeneMark, and others for metagenomic sequences. Establishing community consensus on a small number of protocols suited to the various conditions is a key deliverable of this task. A strong emphasis should be placed on improving open-source methods.

2A. *Identify the best set of gene-finding algorithms for identifying gene fragments.*

Test on short-read archives or multiple-sized artificially fragmented sequenced genomes. Comparing gene-finding algorithms requires a method for comparing results and some gold-standard datasets. Such datasets should contain genomes with different GC contents and sets that have accurate N-terminus sequences supported by proteomics data.

2B. *Improve the best gene-finding algorithms for use on datasets having a significant mixture of assembled and unassembled reads.*

Ideally, this would involve active collaboration on an open source gene finding software package.

Task 3. Produce reliable functional annotations based on information derived from correlations between orthologs and environmental parameters across metagenomic datasets.

3A. *Identify orthologs among metagenomic datasets.*

The community has established a well-understood set of algorithms that can identify orthologs when comparing two organisms. The most common group of algorithms is based on some form of all-against-all search coupled with a form of reciprocal best-hit requirement. Several variants of this approach exist in the public domain. The utilization of these approaches and optimizations in run-time performance will be an important part of this subtask. Determining groups of orthologs within a metagenomic assembly might also be accomplished using something like TIGRFAMS or FIGfams that are specific enough to define orthologs. The use of existing models (TIGRFAMS and FIGfams) will also require focusing on run-time performance.

3B. *Track orthologs across metagenomic datasets.*

Using the methods described in 3A, metagenomic datasets should be linkable based on the presence or absence of particular orthologs. Universal gene identifiers that link homologs across datasets will need to be generated. Minimally, this identifier should associate a gene id, an ortholog id, the strength of orthology, and the metagenomic dataset.

3C. *Normalize metadata produced by different investigators.*

Applying standards to metadata for consistent data representation is necessary for correlating findings across metagenomic datasets. This will require some manual curation of metadata, involvement in proactive support of metadata standards, and development and use of synonym tables. Nomenclature for metadata in Kbase should conform to existing community standards. Metadata values will need to be normalized across datasets. Similar to the transform in exact, transform, and load (ETL) architectures, metadata values will need to be transformed into a common representation. As a simple example, environment temperature of one sample is collected in Celsius while another metagenomic study uses Fahrenheit. This example is overly simplistic but is illustrative of the subtask for data normalization.

3D. *Incorporate additional metadata when possible.*

Methods for obtaining additional metadata will need to be developed. For example, a metadata document might contain the date, location, and time of sample collection but nothing about the weather conditions preceding the collection (e.g., average temperature, air quality, precipitation, and humidity). Weather information might be obtained by querying the national weather service information systems with a location and date. Similar external data should be identified and used for marine and ground samples and other environments.

3E. *Develop methods for identifying correlations between genes and environmental conditions.*

The development of correlation-based annotation between a gene model and the environment where it is found will provide general annotation or hints at the function. Terms that capture these hints or implications must be developed and applied at the appropriate level of granularity. The effects of the strength of orthology on the correlation value between gene and environmental conditions should be considered in later years in methods for representing this strength in various investigations, including correlation studies. Methods for providing this information to users will need to be developed.

3F. *Identify genes or proteins that display the same activity but lack sufficient similarity.*

Genes that seem to lack common origin are said to have analogous activity. The implication is that analogous proteins followed evolutionary pathways from different origins to converge upon the same activity. Thus, analogous genes or proteins are considered products of convergent evolution. Analogs have homologous activity but heterologous origins. Methods for linking analogs across metagenomic datasets are more difficult and are seen as long-term elements of this implementation plan.

Task 4. Support experimental-based annotation derived from high-throughput assays.

This will be extremely important for the long-term success of this plan. Unless new high-throughput technology is developed, we will be unable to verify the vast majority of the implications that will be identified by the analysis being developed in this plan. Initially, we should develop technology to speed up assays we already know how to do. This will create the greatest amount of biological information for the least money and will simultaneously allow for vastly improved automated annotation. The structure of the data produced, access to that data, and application of it to Tasks 3 and 5 will help elucidate the function of poorly characterized genes. This task will require Kbase to leverage experimental biology efforts to perform these collaborative verifications.

4A. *Develop appropriate data models.*

Working with high-throughput assay laboratories, identify appropriate experimental data and subsequently the data structures (models) for representing it. Because of the complexity of biological data and the need for representing the relationships between different biological concepts, an approach that uses semantic web technologies rather than focusing entirely on relational database technology will be required. Open efforts to define ontologies should be used and contributed to.

4B. *Develop methods for updating relationships among metagenomic datasets based on new understanding of the functions that exist in a microbial community.*

As new information from functional assays becomes available and as the results of correlation analysis shed light on poorly characterized genes, these insights need to be captured, quantified with a level of certainty, and made available to the community. Correlation analysis using the results from high-throughput characterization assays will be essential as new high-throughput functional assays come on line.

Task 5. Provide the capability to visually and computationally navigate and discover relationships among genes, between genes and organisms (pan- and core genomes), and between orthologs and environmental parameters.

Include in these graphs a measure of confidence of the relationship.

5A. *Develop appropriate data structures to represent concepts of function and environment.*

The complex nature of the relationships among concepts related to biological function and environmental conditions and between function and environment will require that advances in semantic web technology play an important role in data models. Development of these models must leverage existing activities in the biological and environmental sciences related to controlled vocabularies, ontologies, and associated standards.

5B. *Extend existing software to map and visualize the interrelationships of multiple genomes and environments using the latest computer architecture and visualization tools.*

Multiple open-source packages exist that allow for efficient navigation of graphs. Data structures that will be navigated will be graph-like in nature due to the multiple levels of relationships among biological and environmental concepts. This software will be evaluated for its fitness to visually represent relationships among genes, orthologs, and environmental parameters. The selection of a software package for visualizing graph-based relationships should be influenced by the ease with which the software can be extended.

Resources

***Metacommunities 2**: Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function*

Table 4.7 Hardware Resources for Metacommunities 2

Hardware Purpose	Type	Size
Data management	Storage	Petabytes
Data analysis	Processing	Large (more than 100 cores)

***Metacommunities 2**: Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function*

Table 4.8 Staffing Resources for Metacommunities 2

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
1. Develop resources for assembling metagenomic datasets into consensus sequences		
1A. Provide quality control and quality filtering on sequence read datasets.	SE Bfx	1–6
1B. Improve the binning phase of the assembly process to utilize information about the distribution of closely related strains and species in the metagenomic dataset and integrate the binning and assembly processes more closely.	CS Bfx	1–24

Metacommunities 2: *Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function*

Table 4.8 Staffing Resources for Metacommunities 2

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
1C. Improve the assembly phase of the assembly process to produce a pan- or core genome that is thought to be representative of bacterial taxa at various taxonomic levels.	CS Bfx	12–48
1D. Develop a model for representing polymorphisms when assembling multiple taxa (strains, species, and genera) into a single consensus sequence.	SE	12–24
1E. Extend the assembly resource to include meta-RNA sequence datasets.	CS, Bfx	36–60
2. Improve gene-finding algorithms		
2A. Identify the best set of gene-finding algorithms for identifying gene fragments.	Bfx	1–6
2B. Improve the best gene-finding algorithms for use on datasets having a significant mixture of assembled and unassembled reads.	CS	6–30
3. Produce reliable functional annotations based on information derived from correlations between orthologs and environmental parameters across metagenomic datasets		
3A. Identify orthologs among metagenomic datasets.	Bfx, CS	1–36
3B. Track orthologs across metagenomic datasets.	SE	24–36
3C. Normalize metadata produced by different investigators.	B	1–60
3D. Incorporate additional metadata when possible.	SE	36–60
3E. Develop methods for identifying correlations between genes and environmental conditions.	S	36–60
3F. Identify genes or proteins that display the same activity but lack sufficient similarity.	Bfx	48–60
4. Support experimental-based annotation derived from high-throughput assays		
4A. Develop appropriate data models.	CS	24–60

Metacommunities 2: *Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function*

Table 4.8 Staffing Resources for Metacommunities 2

(SE = Software engineering; Bfx = Bioinformatics; CS = Computer science; B = Biology; S = Statistics)

Task or Subtask	Expertise	Duration (Months)
4B. Develop methods for updating relationships among metagenomic datasets based on new understanding of the functions that exist in a microbial community.	S, SE	36–60
5. Provide the capability to visually and computationally navigate and discover relationships among genes, between genes and organisms (pan- and core genomes), and between orthologs and environmental parameters		
5A. Develop appropriate data structures to represent concepts of function and environment.	CS, B	12–48
5B. Extend existing software to map and visualize the interrelationships of multiple genomes and environments using the latest computer architecture and visualization tools.	CS, SE, B	12–48

System Releases

Release 1. System capabilities in Release 1 will provide a set of standardized protocols for quality control and quality filtering on datasets of metagenomic sequence reads, improvements to the binning phase of assembly, and an integrated binning and assembly process. The resulting consensus sequences will be tunable based on taxonomic granularity, and a prototype for representing polymorphisms when assembling multiple taxa (strains, species, and genera) into a single consensus sequence will be available. Work will have started on determining the best set of gene-finding algorithms for identifying gene fragments, and recommended procedures will be available to the community for use and improvement.

Release 2. System capabilities in Release 2 will demonstrate improvements in the assembly process that produces a pan- or core genome thought to be representative of bacterial taxa at various taxonomic levels. Improvements to the best gene-finding algorithms will have been tested on datasets having a significant mixture of assembled and unassembled reads, and a standard methodology for evaluating gene finders against standardized datasets will be available. Additionally, tools to identify orthologs among metagenomic datasets will appear in Release 2 with the feature of being able to track orthologs across these datasets. This release will provide initial visualization tools that begin to represent the interrelationships of multiple genomes and environments.

Release 3. In Release 3, the assembly process will handle meta-RNA sequence datasets. When possible, additional metadata will be brought in and linked to existing metadata. These additional metadata represent data from sources outside the immediate scope of the metagenomic studies, including databases maintained by the U.S. Environmental Protection Agency, the National Oceanic and Atmospheric Administration, and other federally sponsored resources. Methods and easy-to-use tools for identifying correlations between genes and environmental conditions will be available in the third release, as will support for new experimental data.

5. Mid-Term Science and Leveraged Annotation Needs

The workshop identified several other feasible medium- and high-priority needs that are highly important for the DOE Systems Biology Knowledgebase (Kbase). Three are mid-term scientific needs that could be completed in 3 to 5 years. Another two are tied to improved annotation in the near- to mid-term time frame (1 to 5 years). All five were developed into scientific objectives and requirements and could be developed into a component of the Kbase implementation at a later time. Brief summaries of these follow below.

Three Identified Mid-Term Science Goals

5.1 Analyze Understudied Microbial Phyla

The goal of this scientific objective is to understand the role of unclassifiable members of a microbial community in terms of genetic and phenotypic comparison. To achieve this objective, physiologic and metabolic datasets must be linked to metagenome annotations to provide context and evidence. This linkage will create a more informative and flexible product. The specific datasets to be utilized are the genomes and accompanying physiologic and metabolic data of understudied microbial phyla. Questions that this objective would address are: (1) where are members of a new phylum found, (2) how do we facilitate phylogenetic binning to minimize orphan gene assignment, and (3) what are the emerging concepts of their metabolomes? This is a mid-term (3 to 5 years) priority that requires infrastructure and tool development to accomplish the goals. Elements of this objective are included in the scientific objective titled “Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function” (see Section 4.2).

5.2 Interpret Metagenomic Data to Identify Conditions Required for Growth by Key Microbial Communities Relevant to DOE Missions

Using a partial single microbial genome found within microbial communities, can we predict how to cultivate (and isolate) this target species? Put another way, can we predict culture conditions from genomic information? This will require metagenomic sequence, assembly into species genomes, and pathway analysis of these partially assembled genomes. Existing workflows can perform some of these tasks, but they will need to be developed much further and altered to make use of supercomputing facilities to handle gap-finding exercises. It is not clear if relevant tools exist to accomplish this objective, which was given medium priority because it will take 5 to 10 years to develop.

5.3 Construct, Simulate, and Validate Plant Life Models

Enable semiautomated inference, construction, simulation, validation, and query of complex, multilevel (gene, protein, metabolite, small RNA, organelle, cell, and tissue) plant life models, focusing on models useful for integrating and exploring experimental data types collected during studies of biomass recalcitrance, the carbon cycle, and environmental remediation. Four proposed subobjectives are: automation and streamlining of model construction, development of a semiautomated model validation process, development of an advanced semantic querying capability targeted to biological models and representations, and phylogenetic inference of functional networks (itself a model construction exercise). Model construction and validation are very closely aligned with Kbase objectives. Exploratory model construction is completely dependent on a conceptual framework, together with multiple datasets (annotated genome, proteomic, metabolomic, transcriptomic) to populate instances of this framework. Validation depends on well-structured and well-annotated experimental data, yet the dependencies are modular, which facilitates separate software development for specific or more generalized tasks. Semantic query will enable scientists to more rapidly and precisely develop hypotheses and conclusions from the complex metabolic and regulatory models that arise from genome-scale studies. This science objective requires interfacing with existing plant genomic databases, as well as the Kyoto Encyclopedia of Genes and Genomes (KEGG), gene ontology (GO), MetaCyc, and Plant Metabolic Pathway Database (PMN). This high-priority objective could take up to 10 years to achieve in stages.

Two Identified Science Needs Tied to Improved Annotation

Scientific needs in annotation could be leveraged by the Kbase project. Annotation improvements for both microbes and plants are high priorities and could begin in the near term. DOE's Joint Genome Institute (JGI) is the lead organization in primary sequencing and annotation for organisms of DOE and community interest. The DOE JGI is pursuing and developing plans for improving approaches to incorporate technology advancements. Programmatically, the DOE JGI would have the primary mission to develop and carry out implementation of improved annotation pipelines. These two summaries are included to reflect the importance that the community places on these efforts, as well as to provide input into these plans. The DOE JGI's relationship with Kbase is described further in Section 6.1.

5.4 Integrate Descriptions and Annotations of Microbial Genomic Features

This objective will create the ability to represent and update experimental data and inferred knowledge about genes and genomes so experimental and computational results drive progressively richer and more accurate gene models and predictions. This capability would allow users to access existing genomic sequence information, upload new experimental data to define and refine models, and test consistency between the two. Kbase will address a component of this objective by integrating relevant experimental data that support the specific scientific objectives outlined in Chapter 2, Near-Term Microbial Science Needs Supported by Kbase. This objective requires integration with the DOE JGI, Integrated Microbial Genomes (IMG), and National Center for Biotechnology Information (NCBI), as well as data standards

development and access to large-scale computing resources. Achievement will take 1 to 3 years. This objective also will support metagenomic analyses.

5.5 Improve Plant Genome Annotation Datasets and Make Them More Accessible

Plant genomes typically are annotated in isolation and with varying methods. Even more problematic is that the annotation is rarely, if ever, updated. Consequently, annotation across genomes is not comparable, rapidly becomes stale, and frequently is of undocumented quality. Without confidence in gene model annotations, biological interpretations will be greatly hampered, if not erroneous. The research goal is to generate high-quality, documented, uniform, and integrated annotation for plant genomes. Six target genomes have been identified (*Brachypodium*, *Chlamydomonas*, sorghum, *Populus*, switchgrass, and *Miscanthus*). The goal is to develop a platform that results in annotations that are higher quality than those provided to date rather than to annotate more genomes. In the initial phase, only two genomes that are phylogentically diverse will be annotated in years 1 and 2. Subsequently, in years 2 and 3—with platform refinements—another two genomes will be annotated, and the platform will be further refined. In years 3 to 10, all of the genomes will be iteratively annotated to capture newly available empirical data and algorithmic improvements. This scientific objective would need to be coordinated with the omics data integration objective described in Section 3.2 and with the DOE JGI, NCBI, iPlant, and the plant science research communities. This high-priority objective could be accomplished in 1 to 3 years.

6. Kbase Relationships with Existing or New Resources

The DOE Systems Biology Knowledgebase (Kbase) is providing a unique impetus toward support and acceleration of the biological research community's efforts. However, Kbase is not operating in isolation. There are critical partnerships that it will leverage. Four such partnerships and their relationship with Kbase are described below. These include (1) DOE's lead DNA sequencing facility, the Joint Genome Institute (JGI); (2) DOE's Office of Advanced Scientific Computing Research (ASCR), which is DOE's lead for computational and networking tools; (3) the National Center for Biotechnology Information (NCBI), the National Institutes of Health's (NIH) national resource for molecular biology information; and (4) the iPlant Collaborative, a National Science Foundation (NSF) project in computational plant sciences. These programs can work together to leverage new tools and data and also coordinate efforts in standards development and other areas. The following sections were written in collaboration with representatives of these critical partners.

6.1 Kbase Relationship with the DOE JGI

Microbes

The DOE JGI's Microbial Genomics Program (JGIMGP) will remain a leader in microbial sequence, assembly, and annotation. This program is continuing to automate and accelerate its processes and procedures. The science objective described in Section 5.4, Integrate Descriptions and Annotations of Microbial Genomic Features, is a natural fit and extension to ongoing DOE JGI efforts. These efforts include tools like Integrated Microbial Genomes (IMG). They also include exploring phylogenetic diversity such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA) and confirmation of hypothetical and putative protein gene predictions in collaboration with DOE's Environmental Molecular Sciences Laboratory. JGIMGP sees a clear role in partnering with Kbase to provide improved annotations and integrating with experimental datasets from the research community. DOE JGI will work to integrate and lead in this area and to incorporate the objectives and requirements from the Kbase workshops into its planning efforts.

Plants

The DOE JGI's Plant Genomics Program (JGIPGP) is involved in several areas of sequence-based science that are synergistic with some of the Kbase objectives outlined in this implementation plan. JGIPGP is currently responsible for the assembly, annotation, distribution, and visualization of several reference plant and algal genomes (e.g., *Populus trichocarpa*, Glycine max, Sorghum bicolor, and *Chlamydomonas reinhardtii*) and biomass candidates (switchgrass and *Miscanthus*). Some of these (e.g, *Populus* and *Chlamydomonas*) have already been through two or more new assemblies and annotations, while updates to others are in the planning stages. As noted earlier, "JGI would have the primary mission to develop and carry out implementation of improved annotation pipelines." It would be appropriate going forward for JGIPGP to effectively communicate both with the Kbase steering committee and the larger

plant biology community its roadmap (both in terms of capabilities and schedule) for pipeline development as well as specific genome updates. Such communication would enable more effective project planning within the overall user community dependent on JGIPGP output and would provide a basis for coordination of certain JGIPGP activities with Kbase deliverables.

On the data and analysis systems side, JGIPGP currently supports the Phytozome (www.phytozome.net) platform for plant genomic data visualization, comparison, and distribution. This platform is built around open standards [mainly the GBrowse and BioMart components based on the Generic Model Organisms Database (GMOD)] that the plant genomics science community has already widely adopted. The VISTA comparative platform, which generates pairwise and multiple alignment of plant genomes for comparative genome visualization, is integrated into all plant GBrowse viewers within Phytozome. A distributable stand-alone VISTA package for alignment integrated with visualization of comparative data is currently in the final stages of development at the DOE JGI. The open-source Galaxy framework for analysis workflows (main.g2.bx.psu.edu) is scheduled for incorporation into Phytozome in the next 6 months. It is essential that systems developed within Kbase remain compatible (at some level to be determined) with both JGIPGP data systems and dominant open-source components currently in use. This level of compatibility should, at the very least, include the ability to output and input data in standard formats but could extend to deeper interoperability (e.g., via the Galaxy platform through the use of GBrowse plugins). Discussions should be initiated between JGIPGP and Kbase concerning which existing JGIPGP and broader community components should be adopted or extended (and what the corresponding resource requirements are) versus which should be developed *de novo* as Kbase implementations.

Metagenomics

Kbase has the opportunity to develop software, ontologies, and infrastructure in collaboration or in coordination with the DOE JGI's metagenomics program (JGIMP). JGIMP and Kbase have several similar objectives to support metacommunity analysis. The similarities span both of Kbase's metacommunities-related objectives (see Chapter 4, Near-Term Metacommunity Science Needs Supported by Kbase). Common goals include the exploration and integration of methods for sequencing, assembly, binning, and downstream analysis of metacommunities and integration with isolate genomes; integration of other "omics" data (e.g., expression data, proteomics); development of metadata ontologies and databases; and development of tools that allow efficient visualization, exploration, and analysis of large datasets produced using different sequencing technologies.

Both JGIMP and Kbase objectives include the development of resources for top-to-bottom and bottom-to-top processing of metagenomic datasets (assembly, gene prediction, phylogenetic analysis, and binning, as well as metabolic reconstruction of organisms and communities). JGIMP has pioneered evaluations of data analysis tools, focusing on comparison of available tools for gene calling, assembly, and phylogenetic binning using simulated datasets and the development and update of pipelines that facilitate the accurate analysis of metacommunities. JGIMP is participating in the DOE-funded Metagenomics, Metadata and MetaAnalysis, Models and MetaInfrastructure (M5) initiative, which aims to design metacommunity processing

pipelines using state-of-the-art tools, develop exchange standards, and distribute data from a central data repository center.

Moreover, Kbase and JGIMP have the objective to generate reliable functional annotation of genes and integrate DNA sequencing data with other omics data. JGIMP has developed a set of tools that enable function prediction using existing protein families [e.g., pfam, TIGRFAMS, clusters of orthologous groups (COG), KEGG Orthology (KO) database], pathway collections [e.g., Kyoto Encyclopedia of Genes and Genomes (KEGG), MetaCyc, SEED], and additional nonhomology-based information (gene context, pathway completion), as well as methods for the validation of such predictions. Furthermore, JGIMP has worked toward integrating expression data (transcriptomics, proteomics) with existing genomic data from isolate genomes and metacommunities and developing systems that allow horizontal (genome and community annotation and metabolic reconstruction) as well as vertical annotation propagation (using protein families).

Both Kbase and JGIMP focus on developing ontologies and metadata collections for isolate genomes and metagenomes. JGIMP is developing the Genomes OnLine Database (GOLD), which has been internationally accepted as one of the main metadata catalogs for organisms and metacommunities and is being used by the NIH Human Microbiome Project.

To support the analysis of metacommunities, JGIMP has developed IMG and IMG with Microbiome sample (IMG/M) systems that allow data from isolate genomes and metacommunities to be integrated in a user-friendly environment at the levels of genes, functions, organisms, and communities and published as primary and curated data. Furthermore, to facilitate dataset analysis in the era of tens of thousands of genomes, JGIMP has developed “data compression” strategies such as pangenomes for more efficient representation and analysis of large groups of organisms and communities.

With the commitment of these JGIMP capabilities and data as part of the DOE JGI’s primary responsibilities in genomics and annotation, the DOE JGI would work with the Kbase effort to serve the needs of the metagenomics research community and meet the scientific objectives recommended by the Kbase workshops.

6.2 Kbase and Extreme-Scale Computing Efforts in ASCR

In this section, two important Kbase programmatic issues are addressed: does Kbase have computing needs that require access to exascale capability? If so, does Kbase need to participate in co-design?

Co-design is the iterative process whereby applications, software, and hardware are designed together in such a way that cost-and-performance benefits are freely traded among the three aspects of the complete system’s design rather than the usual process of adapting software and applications to the hardware after the hardware is specified. The successful co-design centers have one or, at most, a few overarching problems that they are targeting. These problems provide a focus for their engagement with the design process for exascale hardware and software. While the overall scope of Kbase is too broad to effectively optimize exascale for all

aspects (plus it is not needed), it is likely that some Kbase elements could play an important role in exascale co-design.

There are several ways that the Kbase effort could leverage the exascale computing capability being developed in ASCR. Most of the Kbase scientific targets are data driven and at a scale that, for the foreseeable future, would allow their computing requirements to probably be met by modest-scale commodity clusters. However, several subproblems will require substantially more computing than likely will be available outside the leadership computing facilities. Examples of these computations are outlined below.

An important issue to resolve in the near future is whether the requirements for computational biology applications for Kbase and similar efforts are fundamentally unique in any way such that proposed exascale hardware and software systems should be considering these requirements from the beginning (i.e., whether there is programmatic justification for a Kbase or "computational genomics" co-design center).

Examples that seem particularly unique and relevant to the exascale and co-design effort include the following areas (all of which are completely different from anything currently being considered in existing co-design centers).

Combinatorial Analysis and Optimization of Biological Networks

During the process of cell network reconstruction, it is often convenient to formulate the desired solution as a discrete or mixed/integer optimization problem and then to search through large numbers of possible configurations to find good solutions based on the optimization criteria. The resulting network reconstructions can be incrementally improved. This method is now being widely used in the reconstruction of metabolic networks but is also expected to be used in the reconstruction of transcription regulatory networks, as well as in the integration of metabolic and transcription networks. Additionally, functional annotation consistency can be formulated as an optimization problem that would enable consistency, accuracy, and comparative analysis to be computing in parallel for all microbial genomes simultaneously. The result could be dramatically improved annotation quality. From an exascale co-design standpoint, the need is twofold. First, the system should be very good at core matrix operations for integer and mixed integer linear programming problems. Second, these problems can be formulated in sets of between 10^6 and 10^{12} subproblems that then need to be solved together. Consequently, low-level support for many-task parallelism at the hardware and operating system is needed, which is something that other applications are not yet driving in the exascale co-design requirements. This supports Microbial Scientific Objective 1: Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function.

Metagenome Indexing, Assembly, and Analysis

As metagenome sequencing gets deeper (i.e., we are able to apply more reads per sample), it will become increasingly possible to extract whole genomes from metagenomics samples through enhanced genome assembly methods. These methods, due to the large size of the datasets, will need considerable computing capabilities (clearly in the petascale to exascale range over the next decade) and, more important, will need large aggregate memory and

tightly coupled processors. Existing prototype implementations are demonstrating that one can effectively use millions of cores and tens of terabytes of random-access computer memory (RAM), and the methods are highly scalable. What makes these different from other applications being considered for exascale co-design is that they are data- and communication-intensive and can benefit from low-level hardware support for fast string comparisons and associative memory type operations. In the future, the bulk of new microbial genomes likely will be sequenced directly from environmental samples, and this capability will directly support those approaches. This ties in with Metacommunities Scientific Objective 2: Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function.

Computing Sequence Similarities and Indexing Kbase Reference Databases

Kbase will be assembling and curating many databases over time. These databases will need to be continuously integrated and periodically updated on a regular cycle. Typical update cycles in existing integrated bioinformatics systems such as SEED, IMG, MetaCyc, and KEGG are on a biweekly to monthly basis. During these update cycles, the sequence similarities between all genes and proteins (and perhaps in the future all metagenomic reads) need to be (virtually) updated or recomputed. This problem is formally an $O(n^2)$ problem that is known not to scale relative to the computing capabilities available in the future. A variety of new methods are being developed to enable the computational integration of datasets in addition to the similarity; however, similarity will continue to be important. Hardware acceleration for local alignments and K-mer indexing and associative arrays would be ideal to support these integration efforts, as would computational (hardware) support for graph indexing and comparisons, and clustering methods. These are relatively unique requirements that are not yet represented in the co-design centers. Addressing them supports the data-intensive computing aspects of the Kbase infrastructure as described further in Chapter 7, System Architecture, and Chapter 8, Kbase Infrastructure Tasks and Timeline.

Computational Screening of Structures, Functions, and Networks

Although predicting all of the ways Kbase will be used in the future is difficult, one clear use will be support of the computational screening of various biological entities. Computational screening is widely used in the pharmaceutical industries as a way to focus limited wetlab resources on the targets and candidates most likely to yield results. It is becoming a key strategy in materials research and in many other areas. Kbase likely will enable screening of proteins for applications in energy, biotechnology, and the environment. Some of these computational screens will be structural, and others will be based on database or computational properties. Many different search and screening computations are possible and often can scale to all of the available resources (i.e., running on millions of cores is not a problem, as they tend to be highly parallel). Screening applications can use serial components or parallel components. They typically need a software coordination layer and scoring functions that may be computationally intensive. These applications, like the optimization applications above, can benefit from operating system and hardware support for many-task parallelism, an aspect of the core services for the Kbase infrastructure (see Chapter 8 and Section 8.7).

6.3 Kbase Relationship with NCBI

NCBI is the major repository of primary sequence data that includes raw sequence reads (both traditional traces and new-generation sequencing), genome assemblies, transcript data, and protein sequence translation products from the coding regions annotated on genomes. More recently, NCBI started collecting more comprehensive information for various types of projects (genome, transcriptome, proteome), as well as descriptive information for samples and phenotypes. This effort led to the development of several new databases: Database of Genotypes and Phenotypes (dbGaP), BioProject, and BioSample. NCBI is also the primary archive for bibliographic biomedical data (PubMed) that allows researchers to connect sequence data with experimental data described in the literature. While sequence data is accumulating in public databases very rapidly, analysis and understanding of organism biology seems to be falling behind.

The Kbase project can fill the gap by creating an infrastructure that will provide users with a single portal to a variety of tools, resources, and multiple data types. For example, these would include metabolic pathways, gene regulatory models, and protein interactions. This will create a good platform for the NCBI-DOE collaboration to be mutually beneficial without duplicating efforts.

There are several areas of potential collaboration: data sharing, cross-referencing of resources, developing community-supported standards for new data types, and developing and reusing data analysis and data visualization tools. NCBI and DOE can work together to make sure that both agencies benefit from complementary approaches and better serve the needs of the scientific community. Certain items have already been identified for further interactions between Kbase and NCBI. A working group will be formed for the Kbase-NCBI relationship that will meet on a regular basis to establish areas of potential collaboration. This working group will identify groups and individual researchers that will work together on specific tasks. These tasks include: (1) identify subprojects that are within the scope of Kbase that overlap with projects NCBI already has or plans to develop; (2) register all Kbase-relevant projects in the NCBI BioProject Database and use BioProject ID for future cross linking; and (3) work together on community-supported standards for genome and metagenome assembly and annotation. This will include standards for metadata (e.g., environmental, ecological, and geochemical), quality of genomic sequence data, and quality of protein functional annotation (e.g., experimental support, metabolic pathways, and cell location).

6.4 Kbase Relationship with iPlant

Kbase has the opportunity to develop software and cyberinfrastructure in collaboration or in coordination with various groups, including the NSF-funded iPlant Collaborative, a 5-year, $50 million project driven by needs of the plant science research community. The Kbase and iPlant projects have several similar objectives to support plant biology research. Although the ultimate goals of the DOE and NSF projects are unique, the solutions have several potential synergies. These include integration of datasets relevant to the understanding of plant and microbial biology, development of standards and semantic technologies, development of tools to support social networking among researchers, and creation of high-performance computational approaches to empower biologists to efficiently use next-generation, ultra-high throughput data generation.

Both the Kbase and iPlant objectives include technologies that will empower biologists to use ultra-high throughput DNA sequence data (including RNA-Seq, polymorphism identification, and transcript quantification). In addition, statistical inference tools to allow efficient association between genotypes and phenotypes are objectives of both Kbase and iPlant. These inference tools include more efficient general linear models and the use of general-purpose graphics processing units (GPUs) to accelerate statistical association studies. iPlant is also supporting the development of an image analysis cyberinfrastructure platform to facilitate integration of image analysis software and provide storage for plant images useful for phenotyping, an outlined Kbase goal. A similar iPlant cyberinfrastructure platform is being discussed and designed to support both statistical and predictive modeling.

To support plant breeders, iPlant is collaborating with a project funded by the Bill and Melinda Gates Foundation called the Integrated Breeding Platform (IBP). IBP is a $20 million project designed and led by highly experienced breeding experts with the Consultative Group on International Agricultural Research (CGIAR). IBP objectives include support for seed storage, phenotyping databases, pedigree support, portable software and hardware tools useful for field biologists, and software to facilitate the use of modern genomics technology and data for crop improvement in developing countries. iPlant software to enhance phylogenetic studies includes the capability to accelerate the determination of phylogenetic relationships through maximum likelihood (RAxML) and neighbor-joining (NINJA) algorithms. These software accelerations include the addition of checkpointing and parallelization to popular phylogenetics approaches. Downstream analysis to assist in the study of trait evolution via comparative genomics is also an iPlant objective. To facilitate collaborations, iPlant has developed a social networking tool for phylogentics researchers called MyPlant.

Finally, it is essential for all publicly funded efforts to work together to support the development of standards such as the Minimum Information About a Plant Phenotyping Experiment (MIAPPHE) and semantic technologies to empower data integration and software interoperability. The Kbase and iPlant initiatives have the opportunity to consider appropriate collaborations or coordinated activities because both are at an early stage of development.

7. System Architecture

7.1 Kbase Architecture Principles

The DOE Systems Biology Knowledgebase (Kbase) will be a large-scale system that:

- Provides access to massive amounts of biological data through hosted services and as links to external resources.
- Provides high-performance and scalable computational resources.
- Supports a large user community with tools and services that enable Kbase utilization.

To meet these requirements, Kbase must be designed with a highly *elastic* architecture that enables continual expansion and scaling to accommodate new data, computational platforms, and software innovations. This necessitates that the architecture be designed and implemented according to a core set of architectural principles described below.

Open

Kbase will provide a published set of open-source application programming interfaces (APIs) to enable the community to access Kbase resources programmatically. APIs will make it possible to create new tools that can exploit data through Kbase and to extend existing tools so that they can exploit Kbase-accessible datasets.

Extensible

Kbase APIs will enable community-driven extensions to the core Kbase resources. For example, new analytical tools that exploit Kbase APIs can be installed as a resource in Kbase. The APIs will enable the tool to be registered in Kbase and be included in tool directories so that Kbase users can utilize the technology in their own analyses.

Federated

Kbase will be a federation of physically distributed heterogeneous compute and data resources. Kbase data will be physically distributed across the federation, utilizing resources that already exist at DOE laboratories and other institutions, as well as newly acquired Kbase-specific systems. A replicated data and resource directory will enable Kbase users to transparently locate and access data as well as execute analysis on Kbase compute platforms.

Integrated

Kbase will create mechanisms to integrate existing community resources that are essential for the DOE systems biology community. By integrating external databases and tools, Kbase leverages community efforts and becomes a hub through which community resources can be discovered and accessed.

Exploit Data Locality

To maximize its performance, Kbase will exploit data locality in its processing. To this end, the Kbase infrastructure will provide transparent dataset replication to provide greater performance and availability. In addition, the Kbase infrastructure will transparently implement mechanisms able to move requested analyses to execution sites that can best exploit data locality and provide maximum performance. These mechanisms will exploit Kbase historical performance logs, metrics, and heuristics associated with Kbase tools to dynamically determine an execution site that provides the best performance.

Modular

Kbase APIs will promote modular, component-based design for codes that execute in Kbase. The Kbase component model will ensure that codes are encapsulated by interfaces that clearly define the services and operations that a code can provide, along with the data types it requires. A component definition will also specify any external dependencies (both data and tools) that a code has, as well as the data types that it outputs. These interfaces will make it possible to easily compose codes represented as components into pipelines that chain together codes to execute complex, multistep analyses.

Scalable

Kbase system architecture will scale simply through the addition of more computational and storage resources. The Kbase software infrastructure will be designed to transparently incorporate new resources so that users and tool builders do not have to be aware of the underlying system architecture.

The overall architecture goal is to provide a set of services and underlying scalable and high-performance mechanisms to support the creation of a broad-based, scalable Kbase. The Kbase architecture is the key enabler in achieving this goal and is essential for Kbase efficiency and low-cost sustainability.

7.2 Architecture Recommendations

Layered Architecture Blueprint

The foremost task for the Kbase platform is to provide the user access to the underlying Kbase associated data, while shielding the user from how that access is achieved (e.g., federated versus centralized, cloud-based versus central server). It should also provide the user with elementary analysis and visualization tools to apply to that data, a way to store intermediate results, data standards to allow data to be exchanged between tools, and ways to chain analysis tools together to create *ad hoc* workflows. In addition, the platform should provide a low-threshold infrastructure for tool development, reuse, and dissemination.

To satisfy these requirements, we recommend that the Kbase system architecture be organized in a series of layers, as depicted in Fig. 7.1 (next page).

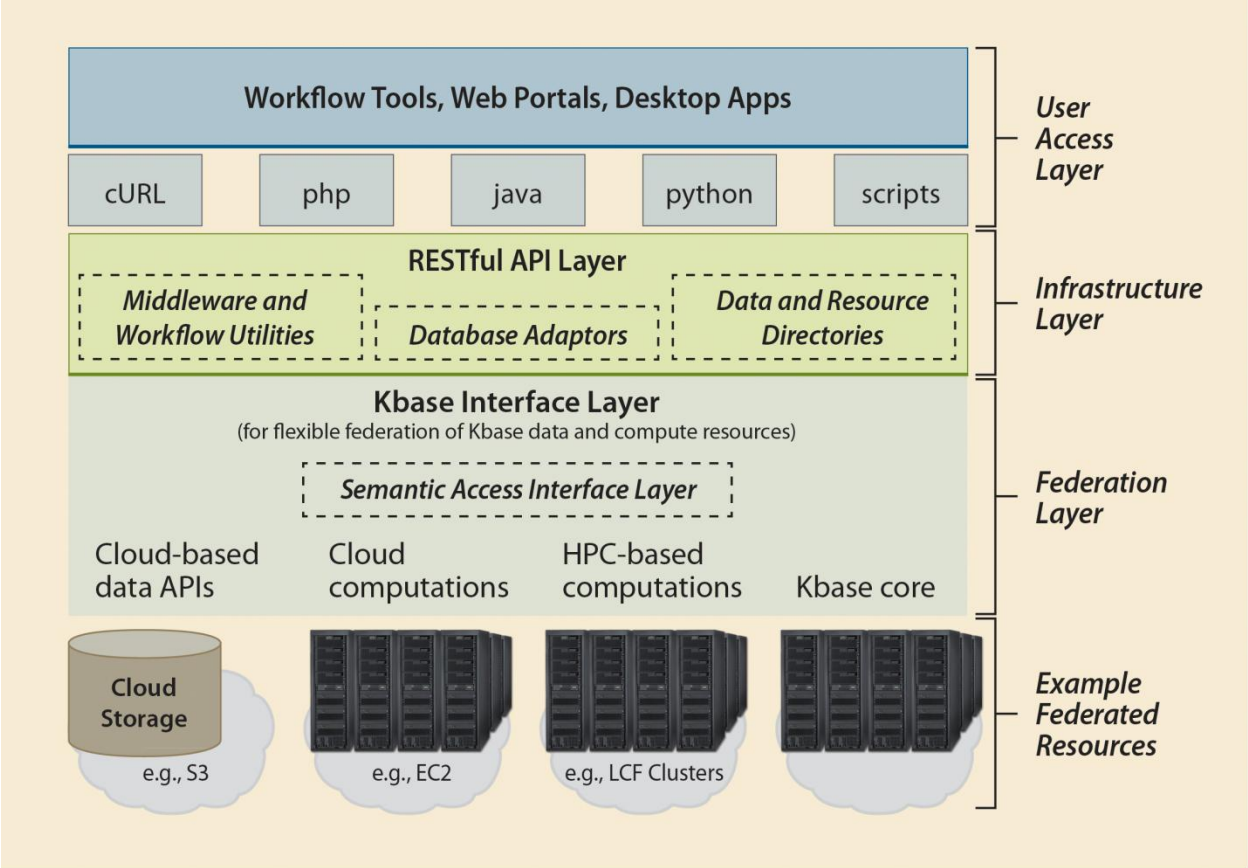


Fig. 7.1 Kbase Architecture Overview. The architecture comprises four layers: user access layer, infrastructure layer, federation layer, and federated resources. The purpose of each is described below.

User Access Layer

The user access layer (UAL) is responsible for facilitating external access with Kbase. It comprises a set of tools that enable biologists to browse, search, download, and upload data from and to Kbase. We envisage a Kbase user environment similar to social networking sites such as Facebook. Users would be able to contribute their own data and tools, form collaborations with scientists from other institutions, specify the visibility of their data, and interact with other Kbase users in *ad hoc* ways (e.g., chat spaces and electronic whiteboards). Tools will also be provided for users to define, execute, and share workflows that leverage Kbase resources to perform complex analyses.

The UAL also comprises a set of libraries that support a published, open-source API. This allows software developers to create new tools and analysis methods to manipulate Kbase datasets. These APIs and development tools will be packaged as a Kbase software development kit (SDK). The SDK must also support a variety of programming languages to allow Kbase developers to leverage development technologies they are most familiar with and to port existing tools to Kbase with minimal modification.

Infrastructure Layer

The infrastructure layer provides the functionality and services needed to support the UAL and employs various mechanisms to associate user requests with the data and compute resources managed by Kbase. This infrastructure forms the core of the Kbase system architecture and includes:

Data and Resource Directories

These directories are the Kbase address book. They advertise the datasets available to Kbase users and the tools and workflows that users can invoke to analyze data in various useful ways. Each entry in the directories is associated with rich metadata that provide a collection of attributes about the resource. For datasets, these may include the data originator, experimental conditions, and additional semantic definitions to unambiguously define the data or software used to produce the data. For tools and workflows, these metadata might include purpose, creator, input formats handled and formats produced, and a summary of execution times from previous runs. These directories will be searchable with user browsing tools and through the Kbase API.

Middleware and Workflow Utilities

Kbase users need to be able to connect various datasets and tools into analytical pipelines that perform complex and often long-running tasks. The Kbase infrastructure will facilitate the definition and execution of pipelines by Kbase compute resources. Based on the tool definitions and metadata in the tool directory, analytical codes can be *componentized* and offer standard interfaces that define the data they require and produce. These components can then be composed into workflows by users and executed by the Kbase infrastructure. In addition, Kbase will provide *data location–aware* mechanisms that can select optimal execution sites for tasks based on dataset availability.

Database Adaptors

Kbase will provide a framework for accessing data resources that must be accessed through a specific API. These resources may exist either external to Kbase or be part of the Kbase federation. The database adaptor framework will make it simple to programmatically integrate various biological databases and obtain results that can be stored in Kbase and made available through the data directory. Kbase will provide adaptors for the most commonly required databases, as well as a set of libraries in the Kbase SDK for developers to create their own database adaptor that can be integrated with Kbase.

Federation Layer

The federation layer will provide the necessary mappings from logical identifiers to physical addresses for Kbase data and resources. Users and applications refer to Kbase resources using logical names represented in the data, tool, and workflow directories. The federation layer is responsible for binding these logical names to actual data and tools. For example, a requested dataset may be stored in a block device in a cloud-based storage system or as a file in an online data archive. As another example, users may wish to invoke an analysis tool, specifying the

input data from their personal storage area in Kbase. The federation layer loads the virtual machine image associated with this tool and launches the software on the specified datasets. Essentially, the federation layer provides a unified view of the underlying physical resources that comprise Kbase.

The federation layer also supports the *semantic access interface*, which supports advanced semantics-based searches from users and tools that operate across the federated Kbase resources. Underlying this interface is a semantic data store that captures relationships between datasets in Kbase by leveraging both metadata and the controlled vocabularies and ontologies supplied by the science community.

Kbase Federated Resources

The problems biologists face require a variety of computing and data platforms and applications that do not all fit onto one single hardware and software platform. The physical compute and data resources that comprise Kbase must be a rich and diverse collection of hardware and systems services. This collection of hardware and services include data repositories and semantics-based metadata (such as ancillary experimental data, ontologies, controlled vocabularies, and data models). These hardware and services would be located at multiple locations and would support virtualization, commodity data parallel computing (e.g., Hadoop based), cluster computing, and high-performance computing (HPC). With the inclusion of the Energy Sciences Network (ESnet) as the underlying network backbone, the Kbase cloud will be a unique and valuable resource for biologists.

Kbase hardware will be a heterogeneous collection. Some applications, such as those related to molecular modeling, require standard HPC platforms. Such platforms are exemplified by DOE's Office of Advanced Scientific Computing Research's (ASCR) National Energy Research Scientific Computing Center based at Lawrence Berkeley National Laboratory in California and the Leadership Computing Facilities based at Oak Ridge National Laboratory's Center for Computational Sciences in Tennessee and at Argonne National Laboratory in Illinois. Smaller compute clusters, not the large ASCR-signature HPC machines, are more generally the target for deploying software developed by bioinformaticists. Smaller compute clusters (generally ranging from 100 to 1000 cores) that support virtualization are needed for a wide range of bioinformatics applications.

Other biology applications are not well suited to HPC platforms. New architectures that focus more on the data and its location are needed. **New computing paradigms where the location of data becomes the primary driver of the location of the computations are leading to the emergence of new technologies and different hardware configurations.** This is already evident in Google's use of the MapReduce architecture and the Apache-supported open-source implementation of MapReduce called Hadoop.

Scientific data centers located at strategic sites on the 100-gigabyte (GB) ESnet will be the hosts for data being generated, analyzed, and shared. These data centers are the likely sites for computations when the computations should be performed near the data. Additionally, these data centers become key elements of a reliable infrastructure where data replication is automated and transparent.

It is recommended that Kbase initially consist of one to seven ESnet data centers upgraded to interconnect at 100 GB. Each scientific data center would be associated with one of the six scientific objectives, and one data center would be associated with the Kbase core infrastructure. Although having these centers co-located would offer some benefits in management and operation efficiency, there are technical reasons for the centers to be dispersed to improve bandwidth to the research community and provide redundancy. This approach would provide an opportunity to evaluate the benefits of multiple data centers. As the number of data centers increases, the apparent bandwidth increases for data delivery to the research clients.

7.3 Kbase Data Representation

A key Kbase aspect will be its ability to provide users with all the data required for a particular analysis through a uniform interface and in a common format. A tremendous challenge in computational biology today is the vast array of formats and schema used for storing data. In addition to providing appropriate storage and access mechanisms, Kbase will provide the integration mechanisms necessary to support comprehensive analysis workflows. Kbase will devise a common vocabulary and a common set of formats to store biological data. Defining the common vocabulary will be a community activity and will leverage extensively the existing and emerging standards efforts throughout the biological community. Initially, a core set of terms will be defined based on community-accepted standard metadata and ontology definitions. The vocabulary will be augmented by a type registry and an associated set of data file formats, which will allow the extension of Kbase to support new data types as they emerge. The initial set of data formats will be limited to only those needed by the use cases.

In addition to the data type registry, Kbase will also implement a data-source registry and semantic search capability to dynamically track which data are available within Kbase. The data resources to be integrated include both the extensive existing databases and file-based data collections (experimental results) currently available to the community, as well as new resources established within Kbase. Community-wide efforts will be supporting the development of agreed data formats, metadata standards, ontologies, and ontology mappings. This work will further require the implementation of metadata resources to aid identification of relevant repositories and federated querying and reasoning. The data-source registry will be utilized by the semantic search service to identify relevant data resources for specific queries or offer those as choices to the user. The repository will contain multiple semantic attributes about data resources that can be used to direct the search. The semantic search will be available both through web- and desktop-based user interfaces (UIs), as well as through the Kbase API for programmatic utilization. As with all systems that provide data through a federation of resources, critical capabilities will be to trace the origin of a particular resource and to make results reproducible. Kbase data services will incorporate a comprehensive system of provenance. Whenever a data request is made, the Kbase data-management system will pass to the calling service an associated set of metadata that will provide the origin, date, and version of requested data. More complicated workflows will carry metadata that provide the provenance of all derived results, including the original provenance of all data included.

The Kbase API will also support role-based access to data. While some Kbase services will be publicly available to any user who connects, many will require authentication. The Kbase data services will work in conjunction with Kbase identity services to allow restricted dataset access to particular users or groups. The API will support research teams depositing data in Kbase to be used for those teams' exclusive pre-publication analysis before being made available to the broader community at a later date.

Kbase data services will exist atop the Kbase storage services, which will provide support for a robust, replicated, and scalable data-storage federation. Kbase will support a multipetabyte online data-storage infrastructure for Kbase datasets and databases. This infrastructure will be expandable to accommodate the expected doubling of data requirements every 2 years, as well as support different storage requirement—from short-term scratch to long-term curation and from simple addressable storage to shared name spaces with high quality services, including database systems. Due to the varied requirements generated by the different use cases, Kbase will support a variety of underlying file storage systems (e.g., parallel file systems, cloud file systems, and tape archives) and will support a variety of replication and retention policies. The federated data directory service will hide from the user the complexity of accessing these many and varied storage systems.

7.4 User Environment

The user environment will provide the interfaces that biologists will use to interact with Kbase collaboratively to exploit data and computational services. The user environment will be open and extensible, enabling incorporation of new applications into the Kbase environment. We anticipate the primary user environment will be web-based and support loosely coupled integration through a data-exchange framework with specific desktop tools used by key Kbase communities for specific scientific needs.

A key attribute of the user environment is to enable biological tool development and integration by providing an open-developer platform analogous to the Facebook platform or Google applications API (see sidebar, The Facebook Platform, next page). This attribute will allow outside developers to produce novel analysis and visualization tools that can query Kbase directly and display and exchange results through the common Kbase UI. There will always be disagreement among research communities on which analysis is best for any particular data type. However, Kbase should not be in the position of enshrining one type of analysis over another. It should provide the platform, allow individual researchers to develop the tools, and let the community reach a consensus.

The Facebook Platform

Facebook released its "Facebook Platform" in May 2007, enabling users to "build the next generation of applications with deep integration into Facebook, mass distribution through the social graph, and a new business opportunity." The Facebook experience has shown that this is an excellent way to involve the community in platform development. Users immediately took advantage of the opportunity and started generating tools and widgets—sometimes in direct competition with tools Facebook had already implemented. The platform provides multiple integration points for applications to integrate seamlessly into the existing Facebook user interface. Many Facebook applications turn out to be useless or poorly designed and disappear into obscurity, but some are absolute hits and propagate rapidly throughout the community, resulting in far more high-quality tools than the Facebook developers could ever have implemented themselves. As of June 2009, 2 years after the introduction of Facebook Platform, Facebook reported 350,000 active applications from over 950,000 developers. A significant part of the platform infrastructure itself was open sourced in 2008, and it is possible that some pieces of this could be leveraged, although Kbase platform needs are likely to be very different from those for a social networking site like Facebook. Note, however, that the underlying Facebook database is much larger than existing genomic databases and has orders of magnitude more users and hits.

This open-development platform is crucial because, regardless of the size and quality of the Kbase development team, there will inevitably be more developers, talent, and ideas (not to mention time to implement) "outside" than "inside" Kbase. Hence, Kbase should be a vehicle to leverage the talent within the scientific community to develop and choose the best tools. Many novel bioinformatics tools suffer from a "failure to launch," never reaching beyond the initial journal publication. By enabling biological analysis tool developers to integrate their methods with the Kbase platform and tie directly into the UI, we can connect a wider variety of analysis tools to a wider range of users and enable more users to become involved in the development process. A consequence of opening Kbase tool development to the community is that some mechanisms are needed to enable the community to disseminate, vote, and prioritize the highest quality tools. Tool reputation may be based on a number of factors, including direct user votes and usage statistics (how many other tools incorporate this tool and how frequently it is actually called). A credible mechanism for attribution and credit potentially could be used to drive tool developers to participate in the Kbase effort. This mechanism could include a tool impact factor that would be comparable to journal impact factors.

7.5 Risk Analysis and Mitigation

The following major risks must be addressed to ensure that Kbase is designed and built to meet current and future community needs.

Requirements

It is essential that the science communities agree on clear requirements and that high-priority use cases are available to drive the design. To mitigate this risk, we propose engaging leaders in each key scientific community at the start of the project to further refine the requirements and test the demonstrations. We will create working groups that include science community members to continuously validate requirements, designs, and implementations. The project will follow a highly iterative design-development life cycle to ensure that demonstrations are available on 1- to 2-month time scales, ensuring continuous validation.

Complexity

Kbase is a complex project in several technical dimensions, including wide-ranging requirements, large-scale heterogeneous data needs, and complex computations. It is essential that the Kbase architecture create solutions that are as simple and uniform as possible to address these complexities. This requires the design and implementation to eschew additional complexity, carefully manage scope, and focus on creating core, extensible capabilities.

Organizational complexity is also an area of risk. In creating the Kbase development teams, care should be taken not only to find highly skilled individuals and groups, but also to organize around teams that have a coherent focus on key tasks, as defined in the infrastructure implementation plan in Chapter 8. A small, core architecture team should be created to drive the overall project design and development, address cross-cutting concerns that permeate all architecture layers, and provide oversight on project progress.

8. Kbase Infrastructure Tasks and Timeline

8.1 Overview

The DOE Systems Biology Knowledgebase (Kbase) will provide users with advanced services to support and enhance their science. In summary, these services are:

Kbase Data Services. Kbase users will be able to create, access, share, and analyze datasets managed by Kbase data services or datasets held in repositories linked to Kbase. Data services will include advanced, semantics-based data searching, access, and integration capabilities and will support storage of large datasets.

Kbase Computational Services. Kbase users will be able to execute both simple analyses and complex workflows using the Kbase computational services. Access to different computational resources will be provided to meet the Kbase community's wide range of needs. These needs will include, for example, petascale high-performance computing (HPC) platforms, clusters supporting virtualization for existing applications, clusters supporting advanced data-parallel applications, and cloud computing.

Kbase Platform Services. The Kbase platform will enable users and developers to easily exploit Kbase data and computational services. The platform environment will support user collaboration and sharing for data and computations, capabilities for creating workflows that execute on Kbase, a software development kit (SDK) for Kbase developers, and the necessary security and integration services to facilitate seamless scientific collaboration within the Kbase community.

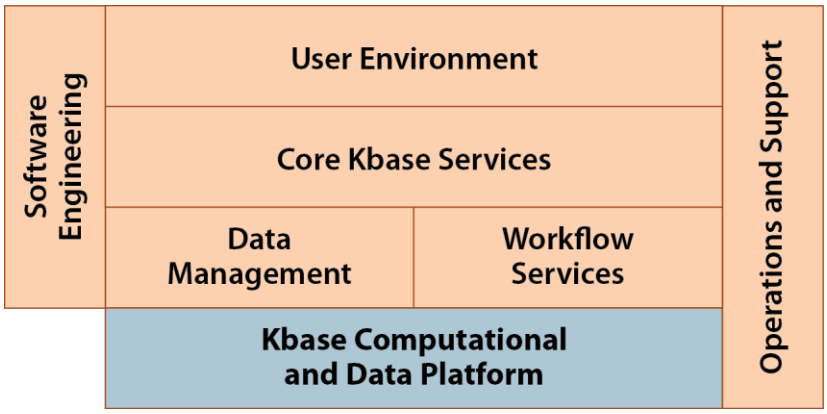


Fig. 8.1. Task Breakdown for Kbase Infrastructure Implementation.

Figure 8.1 and Table 8.1, respectively, provide overviews of the task breakdowns detailed in this chapter for Kbase infrastructure implementation and the associated resources required to achieve the first two Kbase releases in the project's first 3 years, as well as an estimate of the mid-term (5 years total) resources required for the project. Starting with the Kbase computational and data platform (bottom layer of Fig. 8.1), the scope, subtasks, resource

estimates, and timelines for all tasks and hardware infrastructure in Figure 8.1 are described in the remainder of this chapter. These tasks are designed to be as orthogonal as possible and to decompose the overall development of the Kbase software and hardware infrastructure into major software subsystems with clear interfaces.

Table 8.1 Resource Summary for Infrastructure

Deliverable	Duration
Kbase version 1.0	18 months after project start
Kbase version 2.0	36 months after project start
Kbase version N	60 months after project start
Total	**5 years**

8.2 Kbase Computational and Data Platform

Overview

The Kbase infrastructure must be a rich collection of services and hardware. The problems scientists face require a variety of computing and data platforms and applications that do not all fit onto one single hardware and heterogeneous software platform. This collection of hardware and services includes data repositories; data storage or data warehouses; semantics-based metadata clearinghouses; data centers at multiple locations; virtualization; commodity, data-parallel computing (e.g., Hadoop based); cluster computing; and HPC. With the inclusion of the Energy Sciences Network (ESnet) as the underlying network backbone, the Kbase infrastructure is a cloud-based system with a unique and valuable resource for biologists, offering:

Platform as a Service. Kbase will provide a software platform for users to store, access, and share heterogeneous data and to deploy existing and new bioinformatics applications aimed at Kbase-supported science. The platform will support users in exploiting the computational and data resources available in the Kbase cloud.

Infrastructure as a Service. This will allow users to leverage Kbase hardware, thereby reducing local operational costs associated with purchasing, installing, and maintaining hardware, as well as reducing the burden on the facility to house the hardware. Advancements in hardware virtualization now make it possible for users to create images of their local system that can be shared through Kbase with other users, enabling sharing of analysis environments and replication of scientific results.

Data as a Service. This will allow users to store and curate data in Kbase, reducing the need to buy additional storage and to scale their existing infrastructure and data curation services.

Providing data services to the biological research community at a time when data accumulation rates are increasing exponentially will enable research scientists to focus more resources on biological problems.

Hardware Requirements

The hardware behind the Kbase cloud will be a heterogeneous collection. These hardware requirements are discussed in detail under Kbase Federated Resources in Section 7.2, Architecture Recommendations.

Smaller compute clusters—not the large, DOE Office of Advanced Scientific Computing (ASCR)-signature HPC machines—are the target for deploying software developed by bioinformaticists. Smaller compute clusters (generally ranging from 100 to 1000 cores) that support virtualization are needed for a wide range of bioinformatics applications.

Data Services Requirements

Kbase must support a multipetabyte online data-storage infrastructure for Kbase datasets and databases. This infrastructure must be expandable to accommodate the expected doubling of data requirements every 2 years. It also must support different storage requirements, from short-term scratch to long-term curation and from simple addressable storage to shared name spaces with high-quality services, including database systems and managed data repositories that effectively serve data-intensive computing on demand. These requirements can be satisfied by a cluster that runs as a cloud-based, data-as-a-service system.

Kbase also must provide a multipetabyte backup facility of multisite mirroring. In addition, Kbase must provide resources to operate its data services, such as searching metadata clearinghouses, inference or data warehouses, and curated data repositories.

Kbase Cluster Compute Resources

Kbase needs “front end” compute resources to run the Kbase user-access services and data-management (DM) systems and to allow users to create virtual machine images that they can configure for specific, diverse, and typically smaller computational needs. These requirements can be satisfied by a cluster that runs as a cloud-based, infrastructure-as-a-service system.

HPC Requirements

The systems biology computations that Kbase must support typically can be accommodated by a 1000-node compute cluster. These jobs have runtime durations of several hours to several days. In addition, many jobs can run on a smaller amount of nodes with shorter duration runtimes. Larger jobs of petascale and beyond should exploit existing DOE leadership class machines, which Kbase will require access to as computational needs demand. We anticipate that such large-scale jobs are the exception rather than the rule in the foreseeable future. Consequently, we would first seek to partner with existing DOE HPC centers.

Systems biology codes are highly diverse in their programming language runtime requirements and database needs. This means that the compute environment must run a standard version of Linux so that this large variety of codes can run without change.

The compute cluster will require a significant amount of scratch storage.

8.3 Recommendations for Kbase Core Computational and Data Platform

Figure 8.2 summarizes the recommendations for the Kbase core computational and data platform. This platform will seamlessly federate to external compute and data resources as dictated by Kbase science requirements.

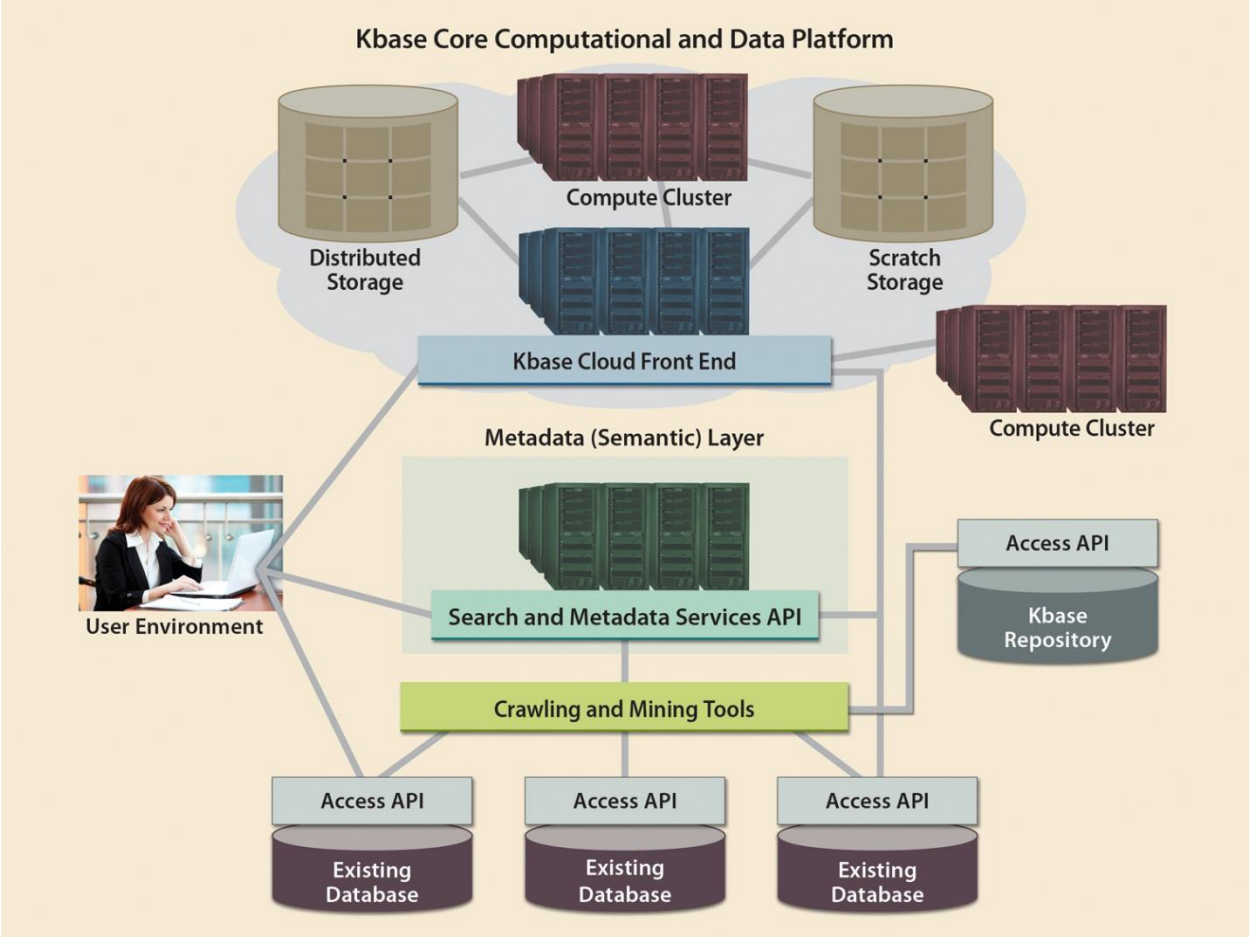


Fig. 8.2. Kbase Core Computational and Data Platform.

Resources

The Kbase cloud system infrastructure includes:

- Scientific data centers (1 to 6) and a Kbase core data center located on ESnet with petabyte (PB) storage capacity.
- HPC resources provided by existing ASCR facilities.
- Cluster compute resources to support commodity, data-parallel applications based on Hadoop, and a virtualization compute cluster located at the data centers. Facilities must have expandable space and infrastructure.

The specific requirements are as follows:

Compute Cluster to Support Data-Parallel Applications

- 256 to 512 nodes (assuming 8 cores per node) for initial configuration; 2-terabyte (TB) minimum local storage (depending on the expected size of the Kbase user community).
- Hadoop running on nodes.
- Nodes running standard Linux and cloud-based resource managers such as Ubuntu.
- Gigabyte (GB) Ethernet interconnect (high-speed interconnect optional).

Online Data Services

- 4 PB "spinning disk" online storage per data center (expected to double in size every 2 years).
- Database servers for scientific databases; metadata databases; semantics-based metadata clearinghouses; and repositories for data, applications, and workflows.

Compute Cluster to Support Virtualization

- Minimum 1000 nodes (assuming 8 cores per node); maximum 3000 nodes depending on the expected size of the Kbase user community.
- Nodes running standard Linux.
- High-speed interconnect (GB Ethernet possible).
- 1 to 2 PB scratch storage, depending on node count.

8.4 Operations and Support

Scope

This task is responsible for providing Kbase systems operations and support. The tasks range from installing and operating Kbase hardware resources to providing support for ongoing Kbase software and hardware.

Subtasks

Establish Kbase Hardware Infrastructure. This subtask is responsible for the acquisition, installation, and initial configuration and support for Kbase hardware resources (see Section 8.2, Kbase Computational and Data Platform for details). The deliverable will be a hardware and software environment that can be used for testing and subsequent deployment of the Kbase version 1.0 system.

Create and Support Federated Kbase Platform. This subtask will perform the necessary system configuration, hardware expansion, and ongoing support to integrate the Kbase compute, utility computing, online storage, and backups. This deliverable will be a fully operational federated Kbase platform that supports the Kbase version 2.0 system.

Ongoing Kbase Platform Operations and Support. This subtask provides the resources needed to operate, maintain, update, and support the Kbase computational and storage platforms. The deliverable is a reliable Kbase platform operating with high availability.

Table 8.2 Milestones for Operations and Support Tasks

(IT = Information technology)

Task/Deliverable	Expertise	Duration (Months)
Establish Kbase Hardware Infrastructure: Kbase hardware platform running and available. This includes establishing data centers and acquiring and standing up clusters for virtualization and data-parallel computations. Access to DOE computational resources also established.	IT	0–12
Create and support federated Kbase platform: Kbase version 1.0 automated build-and-test suites.	IT	12–36
Ongoing Kbase platform operations and support: Highly available Kbase platform.	IT	37–60

Resources

Experienced compute and data-systems administrators and network operations and database administrators will be required for operational support.

8.5 Data Management

Scope

The DM task is responsible for designing appropriate data storage, query, access, and integration mechanisms for Kbase, as well as supporting higher-level tools for collaborative working and data sharing in a secure environment. This task will involve working closely with the science teams to understand their data needs, controlled vocabularies, ontologies, and provenance, and then implementing appropriate data services including:

- **A Kbase data-publishing service based on a data-source registry.** Elements in the data registry represent the fact that a dataset exists. This registry would be used by both users and the automated pipelines that generate analysis results. Pre-computed analytical results would be an example of a dataset that is automatically registered.
- **A Kbase data-discovery service that users use to search the data registry.** The data-discovery service enables the development of both a graphical user interface (UI) and application-programming interface (API) methods to query the data registry for the existence of a dataset or pre-computed analysis result.
- **A Kbase data-retrieval service that enables users to retrieve data once a reference to it has been found in the registry.** This service will enforce access policies.
- **A Kbase data-transformation service (DTS) that automates retrieval, transformation, and load operations to or from, for example, a storage location, analytical packages, remote sites, and alternative storage containers.** DTS allows data to be transformed and used from heterogeneous sources using relational databases, or text-only files, into any supported application format. DTS would allow data transformation to be automated on a scheduled basis and would be able to perform additional functions such as FTPing files and executing external programs. Additionally, DTS interfaces with version control and backup components when used in conjunction with a version control system and ultimately provenance tracking.

Subtasks

Design Core DM Vocabularies and Standard Data Formats. Defining the common vocabulary is a community activity that could take a considerable amount of time and therefore should be started immediately. A step-wise approach should be taken to define a core set of terms early on, leveraging existing community-accepted standard metadata and ontology definitions, to facilitate the core DM system development. In addition to defining a common vocabulary, common, file-based data formats need to be defined for each data type provided by Kbase (e.g., pathways). Again, data formats should be limited to those required by the use cases.

Design Core DM System. This subtask will work with the science and core Kbase services team to design a suitable DM approach for Kbase. Because of the heterogeneous and distributed nature of the datasets required, as well as the federated Kbase nature, this is a complex task. The deliverable will be a design document for the core DM system.

It will be necessary to identify the data sources required to satisfy the needs of the use cases and determine the services necessary to integrate these needs effectively. Similarly, the required basic DM services and their integration will need to be defined. It is expected that the design will include a semantic interface layer on top of existing data sources, a cloud-based data-storage system for file-based data and databases, and a semantic-based metadata resource.

Implement Core DM System. This subtask will create the core DM services required for Kbase. These services will include the necessary underlying schemas and data formatters, mechanisms for managing dataset catalogs and metadata, ontologies, controlled vocabularies, and search utilities across heterogeneous data resources. The deliverables will be part of the Kbase version 1.0 release.

Design Additional Core Semantic Access Integration and Inference Tools. This subtask will build upon the initial DM subsystem to incorporate more advanced semantic technologies that are able to provide sophisticated search and inference capabilities for Kbase users. This subtask will work solely with the user environment and core services teams on an optimal design for incorporating semantics into Kbase. The deliverable will be a design document and proof-of-concept prototype.

Implement Additional Core Semantic Access Integration and Inference Tools. This task will implement the user tools and backend services and mechanisms to provide semantic annotation, search, and inference capabilities to Kbase users. Ontologies developed by the scientific communities will be leveraged wherever possible and built into the Kbase infrastructure. The initial set of tools and services will be delivered as part of the Kbase version 2.0 release.

Design and Implement Provenance Services. This subtask will build upon the core DM versioning capabilities and design and implement configurable approaches to capturing provenance for analyses performed by Kbase. Provenance will be captured for both simple, individual analyses and complex workflows that invoke a sequence of tasks. Users will be able to control the level of provenance they wish to capture, and tools and services will be provided for Kbase users to browse provenance data and produce reports that detail the heritage of a particular set of results.

Evolve DM System. This subtask will extend and improve the DM system to incorporate changes in scientific requirements and evolutions in DM technology to ensure that Kbase remains state of the art. Deliverables will be part of the annual Kbase system releases.

Table 8.3 Milestones for Data Management Tasks

(SE = Software engineering; Bfx = Bioinformatics)

Task/Deliverable	Expertise	Duration (Months)
Design Core DM Vocabularies and Standard Data Formats: Ontology and format specifications.	Bfx	0–12
Design Core DM System: DM system document.	SE	0–6
Implement Core DM System: Kbase system version 1.0.	SE	7–18
Design Additional Core Semantic Access Integration and Inference Tools: Semantic tools design document.	SE	15–24
Implement Additional Core Semantic Access Integration and Tools: Kbase version 2.0.	SE	24–36
Design and Implement Provenance Services: Provenance services as part of a Kbase version 4.0 release.	SE	37–60
Evolve DM System: Annual releases of Kbase system.	SE	37–60

Resources

The staff required for these tasks must have the following range of skills:

- Database design and implementation.
- Semantic technologies.
- Large-scale DM.

8.6 Workflow Services

Scope

The workflow services task will create the necessary user-driven design and execution tools enabling Kbase users to create workflows by defining the automated execution of tool sequences available in Kbase. The resulting workflows will be made available through a registry for other Kbase users to leverage and modify.

Subtasks

Design Workflow Services. This subtask will involve working with the science teams to understand the requirements for user-defined Kbase workflows. It will design a component-based approach that meets these needs, leverages virtualization, and allows workflows to be published and shared by Kbase users through a registry. Where possible, existing workflow infrastructure and description tools will be leveraged and extended to meet the more demanding Kbase needs. The deliverable will be a workflow document and proof-of-concept prototypes.

Implement Initial Workflow Services. This subtask will create the first version of the Kbase workflow services. It will allow users to create, store, and execute linear workflows, or pipelines, on Kbase. It will comprise user tools for creating workflows and backend services and mechanisms to execute workflows. The deliverable will be part of the Kbase version 1.0 release.

Implement Advanced Workflow Services. This subtask will extend the initial Kbase workflow services by adding advanced features for users to exploit and improve performance. The task will improve the user toolset by abstracting advanced features and making the toolset simpler for the user to build and share workflows. Workflow execution also will be made "data aware" so that data movement can be minimized during workflow execution. This subtask also will work with the DM provenance subtask to design and implement suitable provenance capture hooks into the workflow infrastructure. Deliverables will be part of the Kbase 2.0 release for user tools and execution improvements, and part of subsequent annual releases for the provenance.

Evolve Workflow Services. This subtask will maintain and evolve workflow services to meet new Kbase user requirements and ensure that the technology remains state of the art. The improvements will be delivered as part of the annual Kbase systems releases.

Table 8.4 Milestones for Workflow Services Tasks

(SE = Software engineering)

Task/Deliverable	Expertise	Duration (Months)
Design Workflow Services: Workflow services document.	SE	0–9
Implement Initial Workflow Services: Kbase system version 1.0.	SE	9–18
Implement Advanced Workflow Services: Kbase version 2.0 release (36 months). Kbase version 4.0 release (60 months).	SE	19–60
Evolve Workflow Services: Annual releases of Kbase system.	SE	37–60

Resources

This task will require deep skills in middleware, scalable systems design, distributed and HPC, and workflow systems.

8.7 Core Kbase Services

Scope

The core Kbase services subtask is responsible for designing and building a flexible, scalable software infrastructure for Kbase, and providing a Kbase SDK for Kbase developers to exploit these core services. These services and infrastructure provide the mechanisms needed to handle external Kbase requests and serve as the glue that ties together the user environment, data, computation, and workflow services in order to satisfy requests. Tools also will be provided as part of the SDK to exploit virtualization and create machine and application images that can be executed in Kbase. The ability to save, update, and retrieve virtualized computing environments directly supports the ability to reproduce analytical results and to share complex scientific workflows without the need for every biologist to have his or her own cluster.

Subtasks

Design Core API. This subtask will design the core API for handling requests from the user environment for Kbase resources, as well as APIs for associated partners to offer their data, application, or computational resources to the Kbase user community. These APIs provide the backend implementation for the facilities offered in the user environment. This task will be carried out in conjunction with the design tasks for all other Kbase subsystems and be based on Kbase science driver requirements. The deliverable will be a design document and prototype that supports the demonstration of the prototype Kbase user environment.

Design Federated System Infrastructure. This subtask will investigate and design suitable mechanisms to transparently federate distributed Kbase data and computational resources. The deliverable will be a proof-of-concept prototype that validates key mechanisms to reduce risk in the implementation phase.

Implement Core API. This subtask will implement and test the Kbase core API. The API will be implemented in conjunction with the Kbase user environment. It will be built so that it can be trivially scaled through stateless services replication to support a growing user base. The deliverables will be API implementation as part of the Kbase version 1.0 release, the Kbase SDK version 1.0, and associated documentation.

Implement Federated System Infrastructure. This subtask will implement and test the necessary mechanisms to provide a seamless federation across federated Kbase resources. These will be initial but functional implementations designed to be extended as new Kbase federation requirements emerge. The solutions will include security, data access and replication, and launching computations across the Kbase federation. The deliverables will be the software implementation as part of the Kbase version 1.0 release and associated documentation.

Design Extensible Tool API. This subtask will work with the science tasks (microbial, plant, and metacommunities) to understand the requirements and design a suitable approach for creating an API and tools to enable users to add Kbase applications. The deliverables will be a design document and proof-of-concept prototypes.

Implement Extensible Tool API. This subtask will implement the APIs and user tools that enable user-provided Kbase extensions. These will be tools that can be made available for use by other Kbase users and that either execute on the Kbase computational infrastructure or are downloadable for local use. The deliverables will be the Kbase SDK version 2.0 and API implementation as part of the Kbase version 2.0 release.

Evolve Core Kbase Services. This subtask will design, implement, and deliver annual releases of the core Kbase services. These releases will modify, improve, and extend existing features to meet emerging use requirements and introduce new capabilities to ensure that Kbase remains state of the art.

Table 8.5 Milestones for Core Kbase Services Subtasks

(SE = Software engineering)

Task/Deliverable	Expertise	Duration (Months)
Design Core API: API design document and prototype implementation.	SE	0–6
Design Federated System Infrastructure: Proof-of-concept prototypes.	SE	0–9
Implement Core API: Kbase version 1.0. Kbase SDK version 1.0.	SE	7–18
Implement Federated System Infrastructure: Kbase user environment version 1.0.	SE	9–18
Design Extensible Tool API: Demonstrable prototype extensible user environment.	SE	19–24
Implement Extensible Tool API: Kbase version 2.0. Kbase SDK version 2.0.	SE	24–36
Evolve Core Kbase Services: Annual releases of Kbase core services.	SE	37–60

Expertise Required

The expertise required for the Kbase core services tasks includes middleware design and implementation; web services; scalable, server-side design and implementation; and systems-level programming.

8.8 Software Engineering

Scope

The software engineering task is responsible for creating and managing the work environment required to support the complete life cycle of building, testing, deploying, and maintaining Kbase-supported software, web applications, and services. This work environment should include facilities that support software design, development, testing, and deployment, as well as application services such as team collaboration.

Subtasks

Establish Open-Source Development Repository. This subtask will create the infrastructure for Kbase development teams to share and manage their code, manage and resolve error reports, and generate metrics. The deliverable will be a software repository ready for Kbase development teams to utilize.

Create Automated Build-and-Test Suites. This subtask will create a software engineering environment that is able to perform automated "build-and-test" cycles on a regular basis (e.g., daily or weekly). This environment will streamline Kbase development and ensure a higher-quality product capable of finding errors more quickly. Scripts and test suites will be built and delivered along with the Kbase version 1.0 release.

Manage Ongoing Software Development Efforts. This subtask will continue to evolve the build-and-test infrastructure for subsequent versions of Kbase software systems. Each release will be associated with extensive regression testing suites.

Table 8.6 Milestones for Software Engineering Tasks

(SE = Software engineering)

Task/Deliverable	Expertise	Duration (Months)
Establish Open-Source Development Repository: Software development repository.	SE	0–6
Create Automated Build-and-Test Suites: Kbase version 1.0 automated build-and-test suites.	SE	7–18
Manage Ongoing Software Development Efforts: Automated build-and-test suites for each Kbase release.	SE	19–60

Resources

Experienced software build-and-test engineers will be required.

8.9 User Environment

Scope

The user environment task is related to creating the software interfaces that biologists will use to interact with Kbase and exploit the data and computational services both on their own and collectively as collaborating groups. We anticipate that these will be primarily web-based for basic services, along with a number of desktop tools for more advanced users that supplement the web environment for more complex tasks identified by specific scientific needs. The user environment is targeted at scientific Kbase users and is envisaged to be open and easily extensible, enabling users to add their own applications. If possible, the user environment should be based on a suitable existing framework with an extensive user base to leverage prior developments for core functionalities that would allow researchers to establish their own secure individual environment within Kbase, as well as share it with a broader user-defined group. This environment would be user configurable and enable information sharing through the use of tools such as MediaWiki, WordPress blogs, and other web-development environments that allow biologists to create personalized web content based on their individual interests and extend their environment to the research community. The user environment would support a system for web-based seminars, tutorials, and demos for user training and dissemination of Kbase capabilities and science.

As part of this development, the integration of social networking tools will facilitate the Kbase objective of stimulating scientific interaction and adoption of the community collaboration needed to address the substantial upcoming grand challenges in systems biology. Kbase will need to provide not only computational, workflow, and data- and science-related tools, but also a supporting infrastructure that encourages expertise sharing and work collaborations within Kbase. Part of this offering should be measures to stimulate:

- Scientific interactions and platform adoption.
- A fair, yet expressive reputation-scoring system for analysis tools and tool developers, datasets, and computational results.
- Incentives for community buy-in and participation in an open platform aligned (or at least not in conflict) with more traditional scientific incentives such as publication records, intellectual property rights, funding, and tenure.

Subtasks

Initial Design and Prototype. This subtask involves working with the science leads and core Kbase services group to design and rapidly prototype a Kbase user environment. The prototype will be limited in scope, covering the basic actions that a user needs to perform when using Kbase. These will include, for example, loading datasets, searching available data, launching tasks, and viewing and downloading datasets of interest. The deliverable will be a demonstrable user environment that can be used to exhibit key Kbase features and gain user-community feedback that guides the design and implementation of the initial user environment software.

Implement Core User Environment. This subtask will design and implement the first version of the Kbase user environment and be based on the prototype design. This design will be web-browser based and enable users to interact with Kbase to load, search, and access data. It will also implement the initial set of policies for data governance, security, and sharing among user-defined subgroups. This task will be undertaken in close collaboration with the efforts in the core Kbase services task. The deliverable will be the first release of the Kbase user environment for utilization by the community.

Design Extensible User Environment. This subtask will design and prototype the features required for users to incorporate extensions into the Kbase user environment. Such extensions will enable users to add tools and services that utilize the Kbase API to augment the Kbase user environment and be made available to the whole community through the Kbase infrastructure. In the Kbase context, this will specifically address the extension of Kbase UIs for the microbial, plant, and metagenomics communities. The deliverable will be a demonstrable prototype of an extensible user environment exhibiting key features so that community feedback can be obtained and incorporated into the final design.

Implement Extensible User Environment. This subtask will design and implement the first version of the extensible Kbase user environment. Extensive testing will be required to ensure that the user extensions are safe and that applications developed by the community cannot destabilize Kbase. This task will be undertaken in close collaboration with the efforts in the core Kbase services task. The deliverable will be the first release of the extensible Kbase user environment for the community to utilize.

Integrate Existing Tools. This task will integrate existing community desktop tools into the Kbase user environment. It will require working with the scientific groups to identify and prioritize tools, and then design suitable integration mechanisms for each tool type. Once integrated, tools will be made available for the community to download and install from a repository into an individual's Kbase user environment. The deliverable for this task will be periodic releases of the Kbase user environment, with a progressively more comprehensive toolset repository available.

Evolve the User Environment. This task will design and implement extensions and improvements to the Kbase user environment based on requests from the scientific community. The deliverables will be periodic releases of the Kbase user environment, each with new and improved features for the Kbase user base to exploit.

Table 8.7 Milestones for User Environment Tasks

(SE = Software engineering)

Task/Deliverable	Expertise	Duration (Months)
Initial Design and Prototype: Demonstrable prototype Kbase user environment.	SE	0–6
Implement Core User Environment: Kbase user environment version 1.0.	SE	7–18
Implement Extensible User Environment: Demonstrable prototype extensible user environment.	SE	15–24
Implement Extensible User Environment: Kbase user environment version 2.0.	SE	24–36
Integrate Existing Tools: Community tools integrated into Kbase user environment versions 1.0 and 2.0.	SE	12–36
Evolve the User Environment: Annual releases of the Kbase user environment.	SE	37–60

Expertise Required

The primary skill sets required are UI design and implementation both for web and desktop environments.

9. Governance

Governance in the enterprise software domain of the DOE Systems Biology Knowledgebase (Kbase) can be thought of as the organizational approach toward enabling development of and adherence to policies and procedures. Policies are the design decisions combined with the incentives to adhere to the design. Since a primary goal of a good architecture is to define a modular system and well-defined abstractions, related choices made along the way need a level of enforcement. Governance starts with a vision of what the governance process will accomplish and how it will be achieved. The following is not a complete governance handbook, but it represents the guiding principles and initial process for establishing an ongoing governance system (especially targeting the near-term 1 to 3 years). This vision should be the collective effort of those who will use, design, build, and finance Kbase. Consensus should be the social norm in the Kbase governance model. This will also affirm initial guiding principles for architecture and operations, recognizing that *DOE has the primary responsibility to ensure that goals are met and that Kbase project management has the primary responsibility for implementation.*

9.1 Vision

Kbase-recommended policies will be developed under a consensus governance model in which the scientific community is actively engaged in governance and in developing and driving Kbase goals and objectives. In addition to the scientific leadership, expertise in computational infrastructure, bioinformatics, and project management is needed in the overall governance body. Scientific leadership is required to facilitate the project launch and create community support. Computational infrastructure and bioinformatics expertise is necessary to ensure that decisions on adopting specific technical applications are appropriately informed. Project management is needed to effectively run the project on a day-to-day basis consistent with DOE needs and under DOE's oversight. The governance body represents Kbase community stakeholders and plays the role of a policy board associated with recommending Kbase design and operations.

Certain broad principles underlie Kbase governance. These include:

- Open access to data and open-source software development to the greatest extent possible, while simultaneously respecting a reasonable level of protection and temporary embargoes to allow publication and career development.
- A federated model with centralized, facilitated coordination.
- Community engagement with stakeholder representation.
- High-level policies, such as open-source development and standard establishment, recommended by the governance body and executed by project management working closely with DOE management and the stakeholder community.

Defining and formulating these principles into policies will be a primary initial task of the governance body in collaboration with project management, DOE management, and the broader Kbase community.

9.2 Governance Body

The Kbase governance body should be composed of representatives of various disciplinary experts (e.g., experimental research scientists, computational infrastructure experts, and bioinformatics scientists) who assume leadership roles appropriate to their expertise within the governance body. Care should be taken that experts recommended to the governance body represent both the disciplines and the range of stakeholders. Project management will execute the Kbase project development strategy with feedback from the governance body, DOE, and appropriate external stakeholders. Considerable effort should be made in the early stages of Kbase formation to ensure a comprehensive stakeholder group is engaged in the project planning and agile development process. The governance body can recommend subcommittees to increase expertise and share the workload. If desired, the governance body can function both as a technical advisory group and as a policy recommendation group to DOE and the Kbase project management.

Responsibilities

The governance body is responsible for recommending the development and maintenance of Kbase policies. The governance body will recommend the establishment of necessary policy and standards committees to define needed policies and draft their implementation, management, and resource requirements. If requested, the governance body can advise on policy implementation and management and their resource requirements. Project management will provide the governance body with the required technical resources to support these activities. Project management also will develop operating procedures, performance metrics, and other structures required to implement and measure the effectiveness of the policies based on standard DOE metrics. Project management will regularly report to the governance body and DOE on the execution of these procedures and on measurements of the associated metrics.

Relationship of Governance to Project Management and Stakeholders

Stakeholders and DOE Genomic Science community members will participate in the governance body and may provide input directly to it. Stakeholder groups represented on the governance body should include: (1) end users of Kbase facilities (researchers); (2) developers of tools supported by or incorporated into Kbase; and (3) producers of data, models, and knowledge incorporated in, or otherwise used by, Kbase and its associated tools. The governance body, as requested, will periodically facilitate evaluation of the project's implementation of policies and achievements against performance metrics (see Section 9.5) and recommend corrective actions to project management as necessary.

Governance Process

The governance body will lead in producing prioritized policy recommendations, xml schemas, wsdl documents, ontologies, and other artifacts that must be distributed to the Kbase community of users and developers. The governance body will recommend policy committees early in the Kbase formation to produce draft policies. Examples of such committees might include an "ontology committee" or a "security policy committee." The committees will be responsible for drafting appropriate policies. Committee policy recommendations will be reviewed by the governance body and sent to project management and DOE program officers for final approval. Adherence to the policies will be a project management responsibility.

9.3 Engaging Community Stakeholders

The governance body will recommend a strategy for engaging the DOE community and creating a "stakeholder" community from interested DOE community collaborators. The governance body will also promote appropriate collaborations to grow that stakeholder community by setting policies and guidance that provide needed scientific and networking tools to empower cross-disciplinary discovery. The governance body will be composed of experimental scientific leaders and computational infrastructure and bioinformatics experts, each of whom will be responsible for engaging their respective communities. This strategy is required to transition from largely independent efforts to a community-driven effort.

The governance body will promote the establishment of "agile" software development practices that engage the user-stakeholder community in the software requirement and development process. Development or pilot projects that demonstrate success are one way to keep the Kbase user community engaged and invested in the software development life cycle so that they feel some ownership of the success. The governance body will also support extensive use of electronic networking tools and strategies.

Enthusiastic communities for Kbase interaction and deployment will be determined by each development project upon initiation. Examples of such communities may include the systems biology community, DOE Bioenergy Research Centers, and microbial and plant science research communities.

Kbase stakeholders will constitute a consortium of community members from large and small projects and an array of institutions. It is assumed that the various participants in Kbase planning are members of this consortium, as well as representatives of centers of excellence, and could serve as part of the Kbase governance body. Such centers also have prior experience in interoperability standards and their development.

There are at least two communities that must be both served and enabled by Kbase. One focus needs to be the biologists using computational analyses to understand their experimental results. Another focus is to enable tool builders. Kbase will not remain structurally static, receiving increasing amounts of similar data types that are analyzed by only one set of tools. Instead, Kbase will comprise a combination of new experimental data and tools that access the growing reference data. By having common access to quality data, tool builders will also have the data product transformations in one place. This should accelerate the evolution of

transformations and provide a better process for designing new data products. Some innovative ideas in this arena were suggested by workshop participants. These include potential tools registries (see the *DOE Systems Biology Knowledgebase Workshop Report from the 5th Annual JGI User Meeting* in Appendix D), challenges, and challenge grants to answer "tool needs," and Facebook-style entries of "my experiment" to advertise. A vision is to improve data analysis sufficiently enough so that experimental sample generation becomes a bottleneck, despite the massive amounts of data generated per experimental sample.

Governance policies must also respect the career development needs and paths of these communities as we move into shared big projects (see the *DOE Systems Biology Knowledgebase Workshop Report from the 5th Annual JGI User Meeting* in Appendix D). Some consideration must be given to institutional technology transfer philosophy and the career impact of open-source software development and use in Kbase. Using the framework of bioinformatics as an integrating force within Kbase, there are perhaps two classes of bioinformatics tasks: (1) publishable research, which develops new algorithms or methods for key problems, and (2) infrastructure support and development, which is likely to be much less publishable and where methods are more likely to be mature. The infrastructural element is analogous to core facilities. There needs to be a strong research activity to generate solutions for next-generation problems in bioinformatics. The open-source policy for algorithms could be similar to the current DOE Office of Biological and Environmental Research (BER) data release policy where there is some allowance for limited access before data is made public. We need to identify proper incentives to enable sustained careers for top people in innovative tools, core support, and experiments. Ultimately, all are required, working together, to attain the ambitious scientific objectives of the future.

9.4 Interoperability Framework and Necessary Standards

The governance body will develop an interoperability framework that includes the necessary interoperability standards and their details. This is seen as the primary standards need during the first year. An interoperability framework should list the standards that Kbase will use, point to reference information, and indicate the status of the choice (e.g., approved, *de facto*, trial, active, deprecated, obsolete). Standards do not generally exist for all time. New standards will be identified by the governance body and managed through an active life cycle process. Typically standards pass through the following stages:

- **Trial Standard.** A trial standard has been identified as a potential Kbase standard but has not been tried and tested to a level where its value is fully understood. Projects wishing to adopt trial standards may do so, but under specific pilot conditions so that the viability of the standard can be examined in more detail.
- **Active Standard.** An active standard defines a mainstream solution that should generally be used as the approach of choice.
- **Deprecated Standard.** A deprecated standard is approaching the end of its useful life cycle. Projects that are reusing existing components can generally continue to make use

of deprecated standards. Deployment of new instances of the deprecated standard is generally discouraged.

- **Obsolete Standard.** An obsolete standard is no longer accepted as valid within the Kbase landscape. In most cases, remedial action should be taken to remove the obsolete standard from the landscape. Change activity on an obsolete standard should only be accepted as part of an overall decommissioning plan.

The governance body and project management will periodically review all standards within the Kbase architecture to ensure that they sit within the right stage of the standards life cycle. As a part of standards life cycle management, the impact of changing the life cycle status should be addressed to understand the landscape impact of a standards change and plan for appropriate action to address it.

Standards to expedite data and file sharing are important. Gene sequence data is relatively established as a standard. An mRNA expression (MIAME; Minimum Information About A Microarray Experiment) standard and other standards are being developed. However, workshop participants had a range of opinions on the priority of standards (i.e., when do we focus on the standards?). Historically, standards development by community consensus has taken a very long time, and there is a need for this effort to move faster. Part of this long duration is driven by the desire to make the standards do all things for all people and uses. For example, required metadata lists quickly become wish lists of all possible information. There have also been "dictatorial" attempts at setting standards. These can lead to frustration as they are outgrown, such as in the file formats used for annotation over the last decade. There is a minimalist view that standards are actually formalized file formats, but the discussions of required metadata move beyond that interpretation. Nevertheless, at a minimum, there was agreement on the need to have some standards for file-sharing formats to expedite transfer (interoperability protocols or interoperability standards). Another consensus was that if we do the needed work, the standards will sort themselves out. If the data exist and there is a need to share, "someone" will create a protocol for sharing, which in effect is a small *de facto* standard. The challenge here is that this leads to duplication and balkanized tools. *Standards are important, but comprehensive standards setting is not the top priority of building a Kbase community. The Kbase effort needs to focus on the science needs and what the initial Kbase version will do.* Beyond the need for interoperability standards, it was not clear that a major effort is required in standards setting in the first year or two. Broadly, the first 2 years of Kbase should focus on implementation, data, and tools to enable specific science.

9.5 Recommended Areas for Initial Governance Board Policy Development

To the fullest extent possible, Kbase will follow an open-source development model using a federated implementation. The governance body will need to recommend policies to promote this strategy.

Definitions

The first policies will require a "definition" stage to define:

- "Open source" and "open contribution." What does this mean? How does it affect Kbase users? How does it affect contributors of both code and data?
- Editorial process policies and organization.
- Methods to engage and retain contributions for data and analysis methods.
- Development of standards policies for the relevant data and analysis methods.
- Embargoes. What will the policy be? How will it affect data and code? This will draw on the existing DOE BER data policy.
- Federation. What does this mean? How does this apply to distributed data (both in its generation and management, and similarly for analysis tools) and distributed computing architecture?
- Development of licensing policies based on open-source and open-development principles and in compliance with DOE and other laws and regulations.

Performance Metrics

The governance body can contribute to the development of key performance indicators and metrics for the principal Kbase services and for Kbase as a whole. Possible initial metrics include:

- Number of users per unit time, number of new users, and number of repeat users.
- Number of publications attributed to Kbase data and tools.
- Results of user surveys:
    - How important is Kbase to your research?
    - What would be the impact if Kbase went away?
    - How responsive has Kbase staff been to your requests for assistance and new tools?
    - How many new tools have been developed based on Kbase infrastructure? What was their impact?
- Individual service key performance indicators (KPIs; e.g., availability, response time).

9.6 Compliance

The Kbase project management structure will be responsible for implementing policies recommended by the governance body through effective operating procedures, architectures, services, and implementation projects. Typically, project management will propose appropriate service-level measurements (metrics) for operational services and projects. The governance body will review and advise on these KPIs, metrics, and service levels. Project management will report regularly to the governance body and DOE on its policy implementation efforts and the achieved (measured) performance levels on the agreed metrics and KPIs. Kbase project management will ensure that any subprojects comply with any procedures required by approved policies. Project management will report to the governance body on policy compliance.

The governance body functions as an advisory board and is charged with making policy recommendations and providing advice on direction. As requested by project management, the governance body will facilitate regular reviews of Kbase performance and recommend, as necessary, modifications to execution plans and procedures.

Kbase project management will ensure that any subprojects develop and comply with any procedures required by policies developed by the governance body. Project management will report to the governance body on subproject progress and policy compliance. The governance body will review subproject performance and recommend, as necessary, modifications to execution plans and procedures.

Project management will report project performance and progress to DOE.

The governance body will make recommendations on granting exceptions to adopted policies. These exceptions should occur and be granted only with adequate justification based on strategic considerations and value to Kbase stakeholders.

Revision, Feedback, Update and Outreach. The governance body and project management will continuously solicit input from Kbase stakeholders with regard to project priorities, policies, and performance. This information will be used in revising policies, priorities, defined service levels and targets, KPIs, and metrics.

9.7 Tasks and Milestones

Governance Tasks and Timelines

- Y1-Q1: First face-to-face governance body meeting.
- Governance body meets, sets initial tasks, and assembles the required subcommittees.
    - Two subcommittees are identified above: Interoperability standards and Definitions (e.g., open source, embargoes).
- Y1-Q3: A draft definitions policy will be provided to the governance body for comment. The governance body revises and approves the initial policy at a second face-to-face meeting. This meeting will also include project management to provide plan feedback.

- Y1-Q4: The interoperability subcommittee must work closely with the Kbase infrastructure and demonstration projects. An initial interoperability standard should be provided by the end of Y1.
- Y2-Q1: The governance body meets and sets goals for Y2 policies and revisions. This would include the establishment of a licensing subcommittee.
- Y2-Q3: The governance board meets to review implementation and Kbase metrics provided by project management. This meeting will be held annually to provide feedback to project management.
- Y2-Q4: Licensing subcommittee provides draft for review by the governance body.
- Y3 continues a similar schedule of establishing priorities, working committees, draft reviews, and feedback provision to project management and DOE.

The governance body will have two face-to-face meetings and two teleconferences per year.

Table 9.1. Potential Risks and Mitigation Strategies

Risk	Mitigation Strategy
Governance body members are insufficiently engaged to accomplish governance tasks.	Select only members who are stakeholders (e.g., have a professional, vested interest in Kbase success). Provide a mechanism for the governance body to easily replace members who are unable to engage at the level required.
Governance body members have insufficient time to accomplish governance tasks.	Compensate governance body members for a portion of the time required for Kbase issues. Select only members who are stakeholders (e.g., have a professional, vested interest in Kbase success). Provide the governance body with administrative and technical support (direct staffing of the governance body or backfilling administrative support at the governance body chairperson's home institution provided by Kbase or DOE BER). Provide a mechanism for the governance body to easily replace members who are unable to devote the required time to governance body activities.
Governance body enforcement lacks adequate authority to enforce policies and priorities.	Project management controls Kbase; DOE provides appropriate incentives through funding and other mechanisms.

10. Project Management

Project management for the DOE Systems Biology Knowledgebase (Kbase) must enable multi-institutional and open research community contributions to a project that provides software, data, and infrastructure needed to meet high-priority scientific objectives for systems biology. The project will involve research, development, and infrastructure to produce a distributed computational system that will be a major advance for the biological research community. Therefore, the overall project management plan should include project management software elements. Both aspects are described in this chapter. The first section covers high-level project management requirements, and the second section focuses on requirements specific to software system construction.

10.1 Essential Project Management Responsibilities

Provide Proper Project Coordination

Kbase scientific and engineering activities will be multi-institutional. Consequently, project management will need to ensure that efficient and productive coordination of activities occurs. Multi-institutional partners must participate in planning and managing change. The management structure should include individuals with experience in managing science and engineering activities across multiple institutions and coordinating changes across an entire project.

Ensure Work Performed is Consistent with Scientific Objectives

The community has defined scientific objectives, and these objectives provide the scope for the work being performed. A process for generating new objectives and reviewing current objectives helps to manage change in scope over time. Allowing the community to define the scientific objectives keeps the project's activities tightly coupled with the goals of the systems biology research community. High-level software requirements derived from the scientific objectives provide further definition of project scope. Periodic reviews that result in new scientific objectives or changes to current ones propagate to software requirement modifications.

Provide Timely Completion of Project Activities within Approved Budgets

Project management will rely on and use the implementation plans for each scientific objective as inputs to the work breakdown structures, scheduling, and resource allocation using management tools such as the Gantt chart (see end of chapter) for these implementation plans. The generation and management of these implementation plans will be an essential project management function.

Ensure Project Outcomes Satisfy Scientific Objective Requirements

Management must establish mechanisms for evaluating overall project performance on a regular basis to guarantee that the project is providing value and utility to the scientific community in conjunction with the governance board and in consultation with DOE. Being

responsible for project quality means that management develops and takes necessary steps to ensure the project will satisfy the needs for which it was undertaken. Periodic project reviews by the Kbase governance board and review teams assessing scientific advancement, engineering soundness, and operational efficiency offer assurance that the project is providing value to stakeholders and meeting expectations.

Ensure Human Capital Productivity

A key management function is selecting staff who can ensure a successful interdisciplinary team and manage staffing changes. Staff development, mentoring, and team building are important project management aspects. Team development through face-to-face time, reward and recognition, and training will be important in the envisioned multi-institutional project. Performance measurement conducted by project management staff will have varied metrics that can include publications and software functionality and usability.

Provide Timely and Appropriate Information Sharing

Management needs to enable and require information distribution and sharing in accordance with the open Kbase philosophy, while respecting individual rights to publication and intellectual property. In short, management must determine who needs what information when and then use the appropriate mechanism for information dissemination. Scientific and technical information as well as project- and task-related information need to be shared with appropriate distribution groups. Communication with the Kbase community will require multiple forms, including user support, training, and outreach. Using social media tools to foster discussion can be a mechanism for informing users of new developments and providing tutorials. Management should ensure that outreach includes symposia and exhibits with live demonstrations at conferences associated with the Kbase project's scientific and technical domains.

Identify, Analyze, and Respond to Project Risks

Management must define and execute a process that identifies project risks, evaluates potential outcomes, defines steps to take in response to outcomes, and, most importantly, takes steps to mitigate risks before they require a response. Risk management must be continuous throughout the project's life cycle as new risks are identified and existing risks become obsolete. Periodic review of the effectiveness of risk mitigation strategies will ensure that the mitigation strategy is being executed and will allow for strategy adjustments.

Provide Process and Oversight for Obtaining the Necessary Computing and Data-Storage Infrastructure

Project management must ensure that the required operational infrastructure is identified and implemented. This includes the computer hardware and facilities to operate that hardware. Management, with DOE input on larger items, will determine what to procure and when and will execute solicitation planning and source selection in adherence to DOE and local instructional regulations. Once in place, project management will administer any subcontracts for maintenance, operation, or other vendor-related services.

10.2 Essential Software Management Responsibilities

In many research environments, investigators develop software to perform their analyses of interest as efficiently as possible. Software code developed and used by a large community and often distributed at different geographic sites requires a different approach. What distinguishes the two approaches is the ability to scale the development process to include multiple developers and to produce code that is usable by many people rather than just the person who writes the code. Kbase project principles include providing an open-development and open-contribution environment. Project management must ensure that these principles are adhered to in an open and productive manner while also accomplishing the scientific objectives.

Ensure Software Requirements are Derived from the Scientific Objectives

Software requirements are captured and managed to establish a baseline for the software development activities. If these are not captured and kept current, development activities can get off track and drift out of the scope defined by the scientific objectives. Project management is responsible for reviewing and ensuring that software requirements are necessary and sufficient.

Establish Software Design Approaches Consistent with the Complexity and Importance of the Software Being Produced

The process of designing the system's architecture, components, and interfaces requires a more formal approach than the typical small software project. The products of this design process will be of critical importance to the system's interoperability, usefulness, and extensibility. Fundamentals such as architectural and design patterns, as well as addressing key design issues such as concurrency, distribution of data and computation, and error handling need to be considered during the design phase and not be delayed until the construction or testing phases.

Enable Software Construction Consistent with Open-Contribution Philosophy

Management must ensure that the software construction process is consistent with the design and embraces an open-contribution philosophy, as well as fundamental concepts of minimizing complexity and anticipating change. Software construction of the scale envisioned for the Kbase project will require good software testing and configuration practices.

Ensure Sufficient Software Quality to Meet Project Goals

Project management is responsible for ensuring software quality by using good software test engineering practices. These practices rely on forms of dynamic or runtime verification that the code does what is expected. This includes testing at multiple levels—from the testing of individual software system parts to the testing of all the parts interacting together as one system. The former is often referred to as unit testing and the latter as system testing, although the conditions under which system tests are performed can lead to user acceptance testing.

Ensure Software Configuration is Available at Distinct Times and Applied within the Context of Software Change Control

Software configuration management is at the very core of software management and is the first area of management that most software-intensive projects embrace. What is at stake is the ability of software developers to share and enhance code within a community of developers. Management must ensure an organizational environment that has the necessary processes and tools in place to allow software versions to be re-created and bug fixes to be applied to existing releases, even as new releases continue in development, changes are approved, and change histories are maintained.

Provide Processes that Support Continuing Software Evolution

A management strategy for maintaining Kbase-produced software is important. Since a large amount of Kbase software will be produced, software maintenance after the initial development period requires forethought and planning. Tools exist that allow users to submit reports identifying bugs, and good configuration management practices can enable changes to current or previous releases in a manner that does not introduce new bugs. Without planning for software maintenance, a simple bug fix might never get implemented and impede a scientific objective or, worse, continue to provide incorrect results.

Establish a Software Engineering Process

Management will define a software engineering process and management functions for monitoring and measuring the processes involved in building software. Key activities would involve process definition, process implementation and change, process assessment, process and product measurement, and improvement of the software engineering process.

Provide Software that Adds Value to the Biological Scientific Community

In addition to testing software, management must ensure that the resulting software actually addresses its intended purpose and is suitable for use. In short, the software must add value to the community. Software quality management relies on the establishment of a healthy software engineering culture, ethics, value, and costs of quality and quality improvement. Quality assurance depends on software verification and validation, reviews, and audits to ensure that the software meets stated requirements. Practical considerations such as understanding quality requirements, defect characterization, and software quality measurement are not well understood in today's bioinformatics community, and management will need to take the necessary steps to introduce these considerations into the practitioner's daily routine.

10.3 Illustrative Management Structure

The management structure shown in Fig. 10.1, below, illustrates some of the key aspects of managing a distributed project heavily involved in software development based on scientific objectives. Managing a project in a scientific setting that requires solid software engineering practices is not new, and DOE has sponsored such projects in the past.

The lead institution would be accountable to DOE for the project milestones and deliverables. A project director located at the lead institution would assemble a management team that ensures success in the management areas outlined in the two preceding sections (10.1 Essential Project Management Responsibilities and 10.2 Essential Software Management Responsibilities). Resource control, in strict accordance with DOE procedures, will be the responsibility of the project director and management team, who will be responsible for achieving project milestones and deliverables. The project director and management team will establish a change control process with various thresholds for tasks, partners, and budgets. Higher thresholds of change will require DOE concurrence or notification.

A governance body consisting of broad expert disciplinary representatives (e.g., experimental research scientists, computational infrastructure experts, and bioinformatics scientists) would advise the project director and management team on stakeholder objectives and policy recommendations for Kbase design and operations. Partner institutions will have their local management teams. Project management, in consultation with the governance body, would appoint technical committees for areas such as verification and validation, support, software engineering standards, and biological data representation standards.

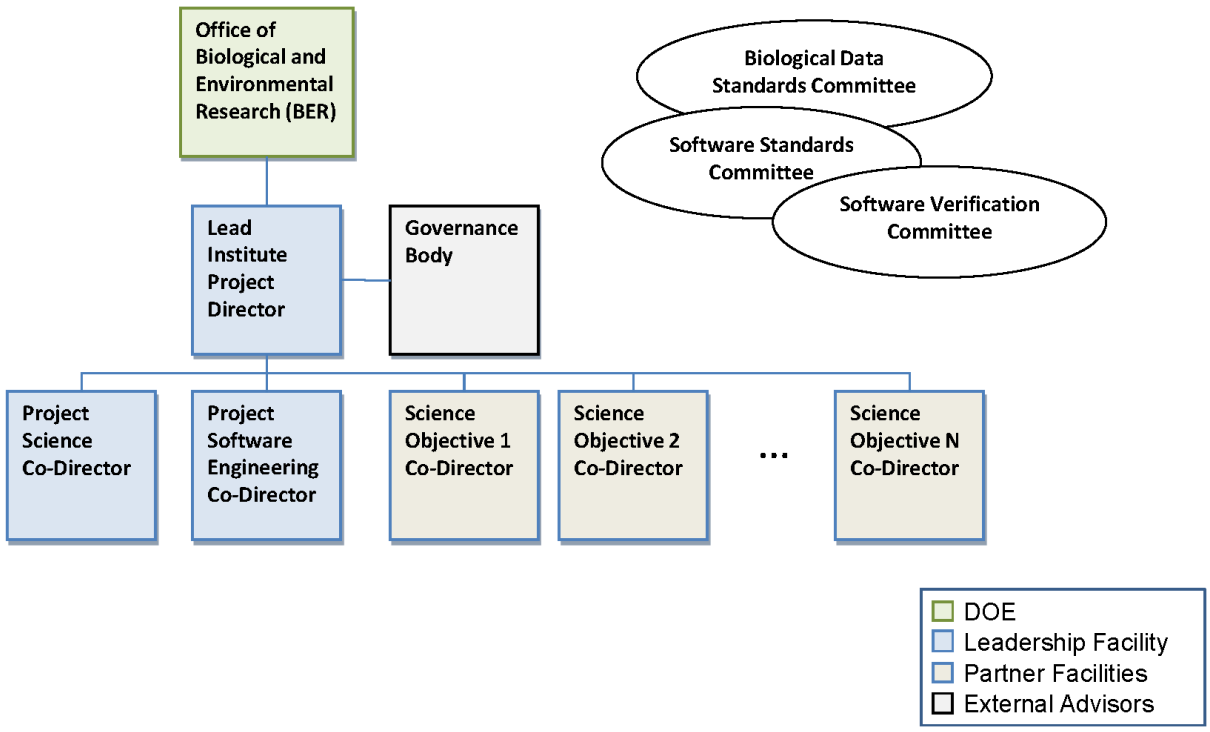


Fig. 10.1. Illustrative Management Structure.

10.4 Overall Project Risk Analysis and Management

A significant PM responsibility is project risk management. Essential risk management components include:

- Identifying the risks by stating known risks and developing a process for uncovering future risks.
- Developing a process for risk analysis.
- Defining a risk response process and proposing responses for known risks.
- Describing how responses will be executed and controlled.

Kbase project risk management is already under way. The implementation plans include an element for identifying risks. These risks are summarized in the software requirements derived from each of the six near-term scientific objectives. Two of these objectives are presented with their associated risks in Table 10.1. While included here for illustrative purposes, the development of a scientific objective and accompanying software requirements and implementation plan is a natural process that includes identifying and documenting risks and mitigation strategies.

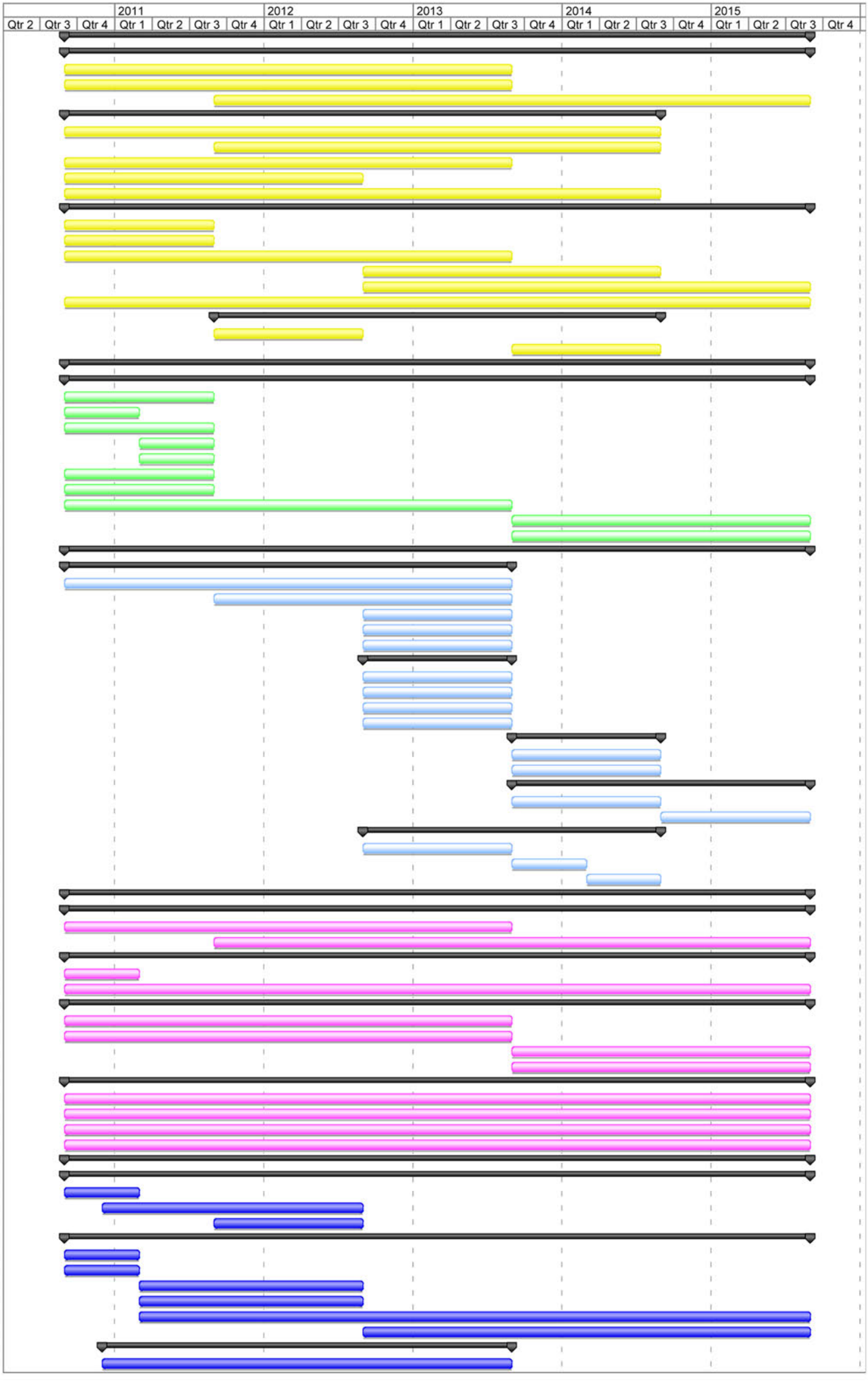
Table 10.1. Potential Risks and Mitigation Strategies.

Microbial Science	
Scientific Objective 2.2: Define Microbial Gene Expression Regulatory Networks	
Risk	**Mitigation Strategy**
1. Stakeholder disagreement over objectives and approaches could undermine the project's ability to produce tools that will find widespread use. This risk is high because various stakeholders have been involved in different stages of Kbase development, and not all were present at a single forum that could have allowed a consensus to be reached. This risk is a frequent Achilles heel in large-scale bioinformatics projects, and there is evidence of this risk in the Kbase project, especially among the microbial contingent.	Continue efforts to achieve consensus and carefully select goals that will achieve the widest buy-in among stakeholders.
2. Unanticipated technological changes (e.g., sequencing, microarray) that would significantly change the requirements or implementation plan.	Anticipate changes and adjust the requirements and implementation plan as soon as possible.
3. Inadequate data or poor data quality that precludes a productive workflow as currently designed.	Test typical datasets for adequacy and quality. Modify experimental protocol to correct and change minimum standards.

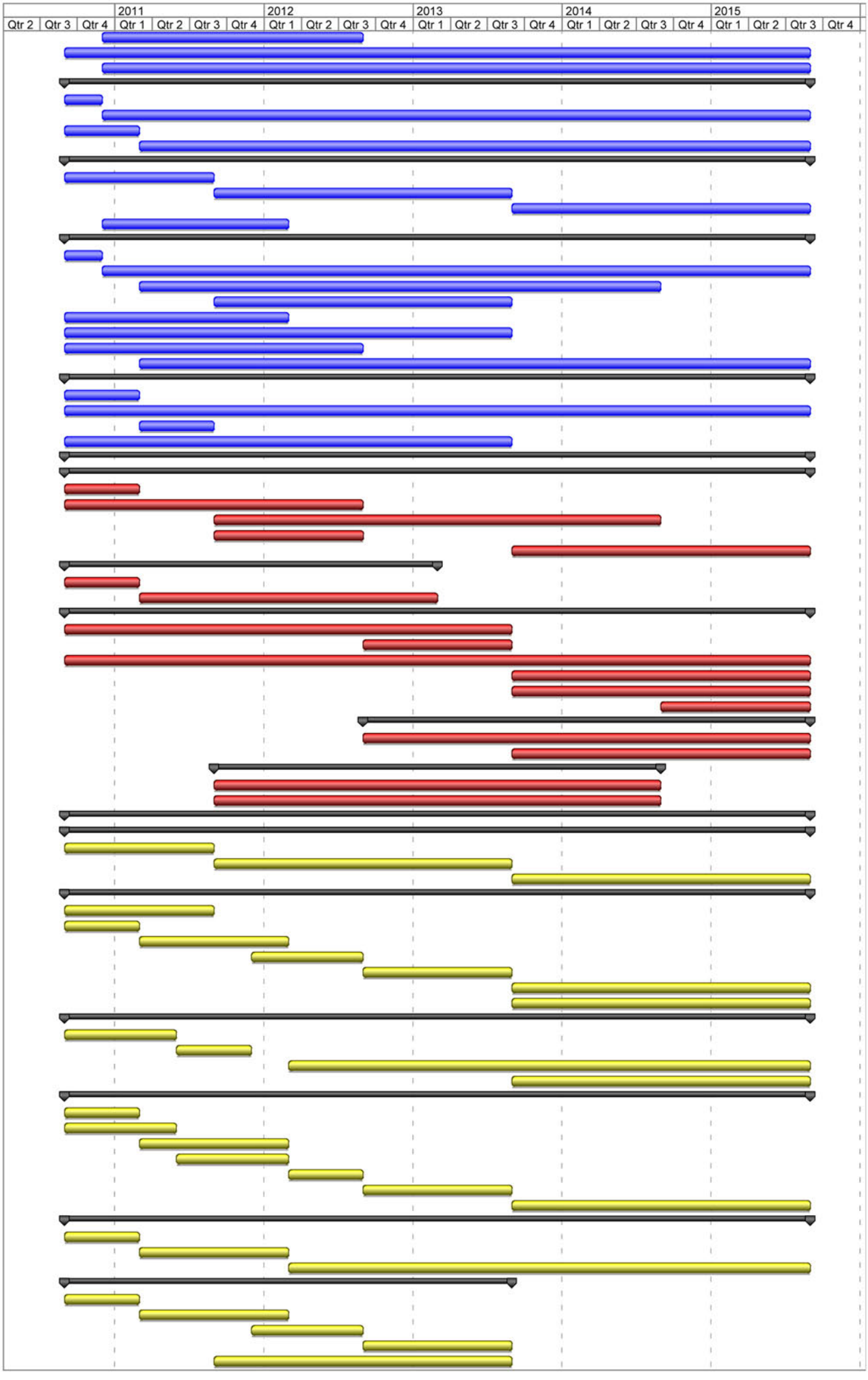
Table 10.1. Potential Risks and Mitigation Strategies.	
4. Cluster analysis on these datasets requires more resources than currently anticipated.	Modify algorithms by accepting some additional error in return for higher performance speed. Allow clustering on subsets to manually find the optimum with reduced error.
Plant Science	
Scientific Objective 3.2: Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Analysis, Comparisons, and Modeling	
Risk	**Mitigation Strategy**
1. Unanticipated slow adoption of one or more target species by the plant biology community, and limitations or delays in the availability of genome-scale datasets for one or more target species.	Prioritize the target species for funding on genome-scale resource development and communicate and collaborate with other funding agencies to ensure adoption and support of genome-scale research in those species.
2. Unanticipated changes in omics technology (namely, high-throughput sequencing and proteomics) that would significantly change the requirements or implementation plan.	Anticipate changes and adjust the requirements and implementation plan as soon as possible.
3. Inadequate omics data or poor data quality that prevents productive workflows as currently designed.	Assess available datasets for adequacy and quality. Modify the platform by adjusting workflows to conform to available and projected datasets.
4. Anticipated algorithm and software improvements for several project aspects (reference-guided transcript assembly, *de novo* transcript assembly, analysis of RNA-Seq data for gene expression profiling, cross-platform expression clustering, and analysis of high-throughput screening derived epigenetic and RNA degradome data) require more resources (software engineering) than currently anticipated.	Anticipate improvements in open-source algorithms used as workflow components and adjust the requirements and implementation plan as soon as possible.
5. Bioinformatic analysis on these datasets requires more computational resources (random-access computer memory, cores) than currently anticipated.	Modify algorithms or workflows to improve performance in terms of speed or hardware requirements, while possibly accepting increased error or other negative performance characteristics.

Gantt chart starts on next page

ID	WBS	Task Name
1	**1**	**Microbial 1: Reconstructing and Predicting Metabolic Networks to Manipulate Microbial Function**
2	**1.1**	**1. Databases**
3	1.1.1	1A. Create a repository of growth data for organisms of importance to DOE in validating growth-prediction algorithms.
4	1.1.2	1B. Create a repository of metabolic flux data.
5	1.1.3	1C. Develop gold standard, manually curated metabolic reconstructions
6	**1.2**	**2. Software**
7	1.2.1	2A. Improve fully automated metabolic reconstruction systems
8	1.2.2	2B. Methods for integration of metabolic and regulatory models
9	1.2.3	2C. Evaluate existing tools and methods for automated design of pathways for metabolic engineering
10	1.2.4	2D. Create tools for comparing of metabolic models with simulation results and with experimentally determined fluxes.
11	1.2.5	2E. Create tools for predicting rate limiting steps within metabolic networks.
12	**1.3**	**3. Applications**
13	1.3.1	3A. Convert into SBML all flux balance models currently unavailable in this format
14	1.3.2	3B. Convert stoichiometric maps into SBML format
15	1.3.3	3C. Decompose the hundreds of existing microbial SBML and CellML kinetic models into individual reaction steps and rate laws
16	1.3.4	3D. Provide better access to an online metabolic regulatory map
17	1.3.5	3E. Integrate gene functional annotations and genome-scale metabolic reconstruction and simulation capabilities
18	1.3.6	3F. Validate metabolic models at five successively harder levels
19	**1.4**	**4. Interoperation and Standards**
20	1.4.1	4A. Exchange and align metabolic models
21	1.4.2	4B. Establish round-trip testing of metabolic models between different platforms and software tools
22	**2**	**Microbial 2: Defining Microbial Gene Expression Regulatory Networks**
23	**2.1**	**1. Enable Automated Inference of Gene Regulatory Networks (short term)**
24	2.1.1	1A. Finalize the definition of regulatory network reconstruction workflow (6 months)
25	2.1.2	1B. Identify specific network inference algorithms
26	2.3.3	1C. Collate existing expression data for microbes of interest or those available.
27	2.3.4	1D. Make available for general use a capability for inference of regulatory networks from expression data
28	2.3.5	1E. Create and make available inferred regulatory network from existing expression datasets
29	2.3.6	1F. Create a controlled vocabulary for meta-information to capture experimental design parameters
30	2.3.7	1G. Provide a user interface for importing and displaying existing datasets, inferred TRN, predicted binding sites
31	2.3.8	1H. Standardize interfaces and APIs for interoperation across selected data repositories, algorithms, and visualization software
32	2.3.9	1I. Generate standards for regulatory network representations
33	2.3.10	1J. Incorporate other data types into regulatory network models [TSS, ChIP-Seq, proteomic, regulator-binding site specificity]
34	**3**	**Plant 1: Integrating Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype**
35	**3.1**	**1. Develop a semantic infrastructure for concepts related to plant phenotype, chemotype, genotype and growing environment**
36	3.1.1	1A. Use, extend and develop controlled vocabularies that apply to plant phenotype, chemotype, genotype and growing environment
37	3.1.2	1B. Translate semantic structures to a consistent schema for database design
38	3.1.3	1C. Provide necessary data services to register, store, query and retrieve data from the data model
39	3.1.4	1D. Apply the meta-model developed in sub task 1A to relevant existing phenotypic and physiological data
40	3.1.5	1E. Apply the meta-model developed in sub task 1A to relevant existing image and multidimensional datasets
41	**3.2**	**2. Develop software for data collection that utilizes the semantic infrastructure**
42	3.2.1	2A. Develop software clients for collecting data in the field
43	3.2.2	2B. Develop server software that will accept, validate and add data from a variety of clients
44	3.2.3	2C. Enable users to save and store routines or configurations used by client software for experimental data collection
45	3.2.4	2D. Enable rapid deployment of barcoding systems within a field setting
46	**3.3**	**3. Implement interactive methods for manipulating, describing, and assessing the quality of data and metadata**
47	3.3.1	3A. Develop server software features that enable interactions (e.g. additions or modifications) with data and metadata
48	3.3.2	3B. Aggregate related datasets, identify outliers, duplicates, and irrational values, and summarize experimental metadata
49	**3.4**	**4. Provide an infrastructure for data mining and analysis based on statistical procedures**
50	3.4.1	4A. Evaluate, extend and develop data models for genetic diversity and phenotype to align with the semantic infrastructure
51	3.4.2	4B. Implement a basic set of analyses for a genome wide association study, QTL study, or for applying genome-wide selection
52	**3.5**	**5. Provide feature recognition software for extracting and quantifying features in raw data (e.g. images and spectra)**
53	3.5.1	5A. Adopt and integrate existing software for detecting features in photographic images for bioenergy applications
54	3.5.2	5B. Incorporate spectroscopic data and provide quality metrics
55	3.5.3	5C. Implement methods to analyze data sets of correlated features to provide predictive ability. (NIR, mass spectrometry, images)
56	**4**	**Plant 2: Assemble Regulatory Omics Data for Target Plant Species in Common Platforms To Enable Analysis, Comparisons and Modeling**
57	**4.1**	**1. Establish a reference plant genome platform with capabilities for visualizing, comparing and automating curation of genomes.**
58	4.1.1	1A. Develop platform and methods for better comparing plant genomes
59	4.1.2	1B. Establish a curatorial process and third party curation tools for continual improvement
60	**4.2**	**2. Develop a platform for access to consolidated omics data**
61	4.2.1	2A. Develop standards and methods for locating, transporting, storing and retrieving plant omics data.
62	4.2.2	2B. Develop appropriate semantic meta models to apply to omics data
63	**4.3**	**3. Extend the platform to support the generation of pre-computed and on the fly analyses of plant omics datasets.**
64	4.3.1	3A. Develop a configurable pipeline(s) to analyze RNA sequencing reads.
65	4.3.2	3B. Develop appropriate semantic meta-models to apply to pre-computed analysis results and to the more stable on-the-fly analyses
66	4.3.3	3C. Extend analysis pipelines to include proteomic, RNA degradome, and epigenetic data sets.
67	4.3.4	3D. Extend semantic meta-models to incorporate proteomic, RNA degradome and epigenetic data
68	**4.4**	**4. Provide an easy to use user interface that supports both plant biologists and plant bioinformaticists**
69	4.4.1	4A. Develop a graphical user interface access to the data
70	4.4.2	4B. Develop an application programming interface to the data
71	4.4.3	4C. Provide a graphical user interface for constructing and executing on-the-fly analyses.
72	4.4.4	4D. Provide an application programming interface for constructing and executing on-the-fly analyses.
73	**5**	**Metacommunities 1: Modeling Metabolic Processes within Microbial Communities**
74	**5.1**	**1. Providing a common platform**
75	5.1.1	1A. Identify essential resources and analysis tools.
76	5.1.2	1B. Develop a repository of essential tools and workflows. (Repository implemented as part of the Infrastructure development effort.)
77	5.1.3	1C. Develop infrastructure for cross validation and characterization of methods for assembly, binning, and pathway reconstruction tools.
78	**5.1.4**	**1D. Develop an environment for facilitating easy discovery, assessment and access to key data sets.**
79	5.1.4.1	Initial common access mechanisms to data sources and clearinghouse of data sources.
80	5.1.4.2	Plan for agreement of common descriptive metadata and annotation format and data formats.
81	5.1.4.3	Develop commonly agreed descriptive metadata and annotation format and data formats for key resources.
82	5.1.4.4	Production level clearinghouse of all relevant data sources and their content.
83	5.1.4.5	Provide common access mechanisms to data sources.
84	5.1.4.6	Develop commonly agreed descriptive metadata and annotation format and data formats for all resources.
85	**5.1.5**	**1E. Develop a workflow environment (repository, shared development, execution)**
86	5.1.5.1	Develop a common tool platform, for ad hoc experimentation and workflow development.


2011
2012
2013
2014
2015
Qtr 2
Qtr 3
Qtr 4
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Qtr 1
Qtr 2
Qtr 3
Qtr 4

ID	WBS	Task Name
87	5.1.5.2	Develop a workflow environment (repository, shared development, execution).
88	5.1.6	1F. Provide computational and intermediate storage resources. (Infrastructure)
89	5.1.7	1G. Develop and maintain curated data repositories
90	**5.2**	**2. Metagenomic sequence data processing and assembly**
91	5.2.1	2A. Identify sources of metagenomic sequence data and …
92	5.2.2	2A. Provide integrated discovery and access to identified data sources (Infrastructure)
93	5.2.3	2B. Implement or provide access to assembly tools
94	5.2.4	2B. Develop or implement new assembly tools as sequencing technology evolves
95	**5.3**	**3. Phylogenetic Analysis**
96	5.3.1	3A. Implement or provide access to community diversity tools
97	5.3.2	3B. Develop, validate and combine phylogenetic binning methods into an integrated binning workflow (Infrastructure will provide workflow service.)
98	5.3.3	3B. Quantification and propagation of uncertainty
99	5.3.4	3C. Implement example workflows
100	**5.4**	**4. Metabolic modeling of community members**
101	5.4.1	4A. Identify required data resources
102	5.4.2	4A. Provide integrated discovery and access to identified resources (Infrastructure)
103	5.4.3	4B. Adapt or develop novel pathway inference methods that can handle noisy and incomplete data (with Section 2.1 Task 2A.)
104	5.4.4	4B. Implement example workflows (with Infrastructure)
105	5.4.5	4C. Assemble a reference dataset of microbial phenotypes and metadata (with Section 2.1 Task 1C.)
106	5.4.6	4D. Develop standardized format for pathway representation, unique identifiers (with Section 2.1 Tasks 1A and 4A.)
107	5.4.7	4D. Assemble a reference dataset of metabolic reconstructions
108	5.4.8	4D. Maintenance of reference datasets
109	**5.5**	**5. Metabolic modeling of the community**
110	5.5.1	5A. Identify necessary physiological data and methods to represent the community in modeling its metabolic processes
111	5.5.2	5B. Develop methods to model metabolic interactions of species in a community and the community response to perturbations and changes
112	5.5.3	5C. Provide HPC resources and access (Infrastructure)
113	5.5.4	5D. Develop heirarchical and multiscale visualization tools for multispecies metabolic models
114	**6**	**Metacommunities 2: Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function**
115	**6.1**	**1. Develop resources for assembling metagenomic data sets into consensus sequences**
116	6.1.1	1A. Provide quality control and quality filtering on sequence read data sets
117	6.1.2	1B. Improve the assembly process binning phase to utilize information about the distribution of closely related strains and species in the data
118	6.1.3	1C. Improve the assembly phase of the assembly process to produce a pan or core genome at various taxonomic levels
119	6.1.4	1D. Develop a model for representing polymorphisms across taxa (strains, species, and genera) in a single consensus sequence
120	6.1.5	1E. Extend the assembly resource to include meta RNA sequence datasets
121	**6.2**	**2. Improve gene finding algorithms**
122	6.2.1	2A. Identify the best set of gene finding algorithms for identifying gene fragments
123	6.2.2	2B. Improve the best gene finding algorithms for use on datasets having a significant mixture of assembled and unassembled reads.
124	**6.3**	**3. Produce functional annotation derived from correlating orthologs and environmental parameters across metagenomic datasets**
125	6.3.1	3A. Identify orthologs among metagenomic data sets
126	6.3.2	3B. Track orthologs across metagenomic data sets.
127	6.3.3	3C. Normalize meta-data produced by different investigators
128	6.3.4	3D. Incorporate additional meta-data when possible.
129	6.3.5	3E. Develop methods for identifying correlations between genes and environmental conditions
130	6.3.6	3F. Identify genes or proteins that display the same activity but lack sufficient similarity
131	**6.4**	**4. Support experimental-based annotation derived from high-throughput assays**
132	6.4.1	4A. Develop appropriate data models
133	6.4.2	4B. Develop methods for updating relationships among metagenomic data sets based on newly discovered functions in a microbial community
134	**6.5**	**5. Provide visual and computational navigation of the relationships among genes, organisms, and environmental parameters**
135	6.5.1	5A. Develop appropriate data structures to represent concepts of function and environment.
136	6.5.2	5B. Extend existing software to map and visualize the interrelationships of multiple genomes and environments
137	**7**	**Kbase Infrastructure**
138	**7.1**	**Operations and Support**
139	7.1.1	Establish Kbase hardware infrastructure including data centers and clusters for virtualization and data-parallel computations.
140	7.1.2	Create and support Federated Kbase Platform: - Kbase version 1.0 automated build and test suites
141	7.1.3	On-going Kbase Platform Operations and Support: - Highly available Kbase platform
142	**7.2**	**Data Management**
143	7.2.1	Design Core DM Vocabularies and Data Formats: Ontology and format specifications.
144	7.2.2	Design Core DM system: - DM system document
145	7.2.3	Implement Core DM System - Kbase system version 1.0
146	7.2.4	Design Semantic Access Tools: - Semantic Tools Design Document
147	7.2.5	Implement Semantic Access Tools: - Kbase version 2.0
148	7.2.6	Design and Implement Provenance Services: - Provenance Services as part of a Kbase version 4.0 release
149	7.2.7	Evolve DM System: - Annual releases of Kbase system
150	**7.3**	**Workflow Services**
151	7.3.1	Design Workflow Services: - Workflow Services document
152	7.3.2	Implement Initial Workflow Services: - Kbase system version 1.0
153	7.3.3	Implement Advanced Workflow Services: - Kbase version 2.0 release (36 months)
154	7.3.4	Evolve Workflow Services: Annual releases of Kbase system
155	**7.4**	**Core Kbase Services**
156	7.4.1	Design Core API - API design document and prototype implementation
157	7.4.2	Design Federated System Infrastructure: - proof-of-concept prototypes
158	7.4.3	Implement Core API: - Kbase version 1.0
159	7.4.4	Implement Federated System Infrastructure: - Kbase user environment version 1.0
160	7.4.5	Design Extensible Tool API - Demonstrable prototype extensible user environment
161	7.4.6	Implement Extensible Tool API: - Kbase version 2.0
162	7.4.7	Evolve Core Kbase Services - Annual releases of Kbase Core Services
163	**7.5**	**Software Engineering**
164	7.5.1	Establish open source development repository - Software development repository
165	7.5.2	Create automated build and test suites: - Kbase version 1.0 automated build and test suites
166	7.5.3	Manage ongoing software development efforts: - Automated build and test suites for each Kbase release
167	**7.6**	**User Environment**
168	7.6.1	Initial Design and Prototype - Demonstrable prototype Kbase user environment
169	7.6.2	Implement Core User Environment - Kbase user environment version 1.0
170	7.6.3	Implement Extensible User Environment - Demonstrable prototype extensible user environment
171	7.6.4	Implement Extensible User Environment - Kbase user environment version 2.0
172	7.6.5	Integrate Existing Tools: - Community tools integrated into versions 1.0 and 2.0 of Kbase user environment


2011
2012
2013
2014
2015
Qtr 2
Qtr 3
Qtr 4
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Qtr 1
Qtr 2
Qtr 3
Qtr 4

APPENDIX A

Supporting Scientific Objective and Software Requirement Documents for Near-Term Microbial Science Needs

This appendix provides the working documents for the two selected microbial scientific objectives and their requirements. These documents were the core output from the final DOE Systems Biology Knowledgebase (Kbase) workshop held June 1–3, 2010 (see Appendix D for the report). The scientific objectives must answer the question, "What is the scientific or research goal that needs to be solved?" The related "requirements" establish workflows and provide details on the needs to accomplish these objectives. The process of identifying the scientific objective, determining its software requirements, and then developing an implementation plan from the two was described in Chapter 1. These working documents are provided as backup to their implementation plans, which are described in Chapter 2 and contain the final revised judgment by the community concerning the tasks needed for each objective.

In research, a scientific objective is satisfied by creating hypotheses and doing one or more experiments. For every experiment, there are rationales, protocols to be executed, a number of data inputs (data sources) and outputs (results), and analysis tools. Workflows are sequential procedures that describe the envisioned steps to answer questions. They are the bioinformatic equivalent of an experimental protocol. Detailed workflows are bridges between the experimental research and computing communities and thus are key to translating research objectives into computing requirements that will most effectively advance the science. Workflows were developed for these objectives. From these workflows and the underlying objectives, the requirements could be defined that lead to the articulation of an implementation plan with tasks and scope to achieve the scientific and technical goals.

In the microbial science area, the first objective is to improve the accuracy of metabolic network models, especially for microbes important in biofuel production and environmental remediation, so that metabolic engineering produces more predictable results. The second objective is to enable automated inference of gene regulatory networks based on data from gene expression profiling. Predicted networks then would be validated to determine their accuracy and refined to improve prediction of cellular behavior and fitness. Both objectives have tasks in developing data repositories and workflows that link into the infrastructure.

A.1 Microbial Scientific Objective 1: *Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function*

The aim of the DOE Systems Biology Knowledgebase (Kbase) is to provide a set of interoperating databases and software tools to achieve the scientific objective of evaluating the metabolic potential of an organism; predicting the phenotypic outcome of specific metabolic or environmental interventions; and developing quantitative, validated metabolic models. This knowledge will lead to the informed modification of one or more specific enzymes or the introduction of entirely new enzymes and pathways. This would allow the community to better

determine strategies for manipulating biomass flow in microorganisms. More specifically, this objective is concerned with reconstructing metabolic networks, predicting the growth phenotypes of organisms from their metabolic networks, understanding the metabolic potential of an organism, and providing scientists with software tools to interrogate and interactively visualize metabolic networks and enable engineers to quickly determine the strategies necessary to remodel metabolism for specific purposes. The goal is to move beyond the current state of the art to increase the speed and automation with which metabolic networks can be reconstructed and to increase the accuracy of metabolic network predictions. Kbase should include access to a variety of data to help achieve the overall objective. Such data would include genomic-level metabolic maps, both stoichiometric and regulatory; enzyme concentration and activity levels; qualitative data on enzyme regulation and known substrate, product, and cofactor dependencies; enzyme kinetic data if available; suggested kinetic rate laws or reasonable approximations; metabolic flux maps (predicted or measured); sensitivity data (rate limitingness and control coefficients) if available; time-course data on metabolite changes and enzyme concentration changes; and finally relevant thermodynamic data (computed or measured) on individual metabolic reactions.

This objective does not currently include protein-protein and protein-DNA interactions; phosphorylation, glycosylation, or acylation states; or signaling pathways. However, part of this information (e.g., covalent modification) is currently held in resources such as MetaCyc. These processes were not within the original scope of the initial objective but could easily be included in a future revision, particularly in relation to dynamics. Gene regulatory networks also are not included in this objective because they are considered in a separate objective description. Ultimately, all three network repositories must be linked, given the close relationships among them.

Probably the three main primary sources of online metabolic data in use today are MetaCyc (www.metacyc.org), Kyoto Encyclopedia of Genes and Genomes (KEGG; www.genome.jp/kegg/), and Braunschweig Enzyme Database (BRENDA; www.brenda-enzymes.org/). MetaCyc and KEGG provide metabolic stoichiometric map data, although access for noncomputer specialists is not always easy. BRENDA has a huge range of diverse enzyme data (83,000 enzymes listed), but the data are incomplete on individual enzymes and at times tedious to access. The System for the Analysis of Biochemical Pathways–Reaction Kinetics (SABIO-RK, sabio.villa-bosch.de/) is one of the few enzyme kinetics databases. The current range of sources is therefore scattered, not always easy to use, and lacks important information. Repositories such as MetaCyc could be modified or new third-party tools developed to enhance and access the data more seamlessly.

Another source of useful data for the metabolic community is genome-based curated metabolic reconstructions. Although these data are of much higher quality, they too are scattered and stored in a number of conflicting, different, and sometimes undocumented formats. One of the primary reasons is that there is no agreed standard for storing flux balance information.

Kacser et al. (1995) provides a good review of the modern interpretation of rate limitingness in pathways.

Prioritization

Given DOE's interest in metabolic engineering for biofuel production and environmental remediation, which require detailed knowledge of metabolic dynamics, the priority for the proposed data source is high.

PRIORITY: _X_ HIGH ___ MEDIUM ___ LOW

Potential Benefits

Metabolism is a critical component for many applications of interest to DOE. Naturally, DOE researchers should have access to reliable and comprehensive data sources pertaining to metabolism.

Feasibility of Success: Near, Mid, and Long Term

A timeline is presented for development of subobjectives comprising the overall goal of Microbial 1: Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function.

A number of short-, mid-, and long-term goals could be achieved. Examples include the following.

Near to intermediate term (6 months)

- Subobjective 1: Generate first-generation (uncurated) metabolic reconstructions from genome annotations for all genomes within Kbase and make them available to the community via browsable and programmatic access.

Near term (1 to 3 years)

- Subobjective 2: Compare metabolic models developed within different software environments, particularly flux balance models, to facilitate combining the strengths of different environments and different models.

Near to mid term (1 to 5 years)

- Subobjective 3: A number of researchers are now measuring fluxes *in vivo* using isotopic methods. One important scientific objective is to compare flux balance model predictions to flux data generated by isotopic experiments. Databases of maps for different organisms under different conditions could be provided with flux data, both predicted and measured.

Mid term (3 to 5 years)

- Subobjective 4: One important scientific objective is to begin to map out the degree of rate limitingness in metabolic pathways (flux control coefficient, Kacser et al. 1995) and to determine how the patterns change under different conditions. For these efforts, enzyme levels and protein turnover rates should be made accessible via a publicly available database. These data also can be used to assist in building kinetic models and assessing global protein demand.

- Subobjective 5: Increase the accuracy of programs that predict growth and intracellular fluxes from metabolic models.
- Subobjective 6: Develop and make accessible, via web and programmatic interfaces, computational tools that can be used for metabolic engineering. These tools would utilize metabolic reconstructions and models developed within or outside Kbase.
- Subobjective 7: A critical scientific objective is to have validated metabolic models for predicting the metabolic potential of a cell. Five suggested levels for validating a model are proposed; each level is more demanding than the previous and requires additional model-data comparisons. The first level of validation and possibly the easiest to achieve is to compare growth and no-growth phenotypes for wildtype and mutant strains. Related to this is the comparison of flux balance analysis (FBA) predictions with isotopic flux measurements as a way to further validate the flux balance models. The next level is to compare predicted steady-state flux and metabolite levels against experimentally measured fluxes and metabolites. Level 4 validation will test the ability of the model to predict the effect of "small" perturbations in enzyme activity levels and environmental conditions. Finally, the most demanding validation test in this sequence is to compare time-course changes due to major environmental changes, such as shifts in nutrients or O_2.

  Validation Levels

  - Level 1: Compare growth and no-growth predictions for wildtype and mutant strains against experimental data (within 2 years).
  - Level 2: Compare flux balance predictions against isotopic flux measurements (within 3 years).
  - Level 3: Steady-State Validation. Compare predicted steady-state metabolite concentrations and fluxes to experimentally measured values (within 4 years).
  - Level 4: Perturbation Validation. Perturb enzyme levels by specified amounts (e.g., 50% change or less) and recompute fluxes and metabolite changes that result. Compare changes with equivalent experimental perturbations (within 4 years).
  - Level 5: Time-Course Validation. Apply environmental changes and significant enzyme perturbations to track time-course changes in metabolite and flux values. Validate against equivalent changes in the kinetic model (long term, 5 to 10 years).

Long term (5 to 10 years)

- Subobjective 8: To understand the role and distribution of rate-limiting reaction steps (measured using flux and concentration control coefficients) in metabolic pathways under different conditions and in relation to the activity of negative and feedback loops.

This would complement the flux, enzyme activity, and kinetic data and would be used to engineer strains with increased flux through desired metabolic pathways.

- Subobjective 9: Integrate the spatial distribution of metabolites and enzymes with metabolic flux analysis. Data on localization of enzymes and metabolites in cells will inform flux analysis and metabolic models, and such analyses and models should be developed to be able to incorporate localization results as they become available.

Relevance to the Kbase Project

The ability to build predictive metabolic models is an important step in proposing design strategies for affecting flux through metabolic pathways. This is clearly of DOE interest, given the need to re-engineer biofuel production by microbial metabolic pathways and to stimulate microbial communities involved in environmental remediation.

Synergies and Leverages: Potential Overlap with Other Projects or Funding Agencies

The Palsson group and others have created genome-scale reconstructions of metabolism for tens of microbes (reviewed in Oberhardt, Palsson, and Papin 2009) that are available in Systems Biology Markup Language (SBML) and Excel formats. The Karp group has created two highly curated metabolic databases (MetaCyc and EcoCyc) and has generated genome-based and metabolic databases for 670 organisms (expanding on a regular basis). SEED (www.theseed.org) has recently incorporated the development and simulation of metabolic models into its database. Other DOE-relevant U.S. metabolic databases include ShewCyc (from the *Shewanella* Federation), BeoCyc (a database of 33 bioenergy-related organisms from the DOE BioEnergy Science Center), PlantCyc (metabolic databases for *Arabidopsis* and poplar from the Carnegie Institution), FungiCyc (from the Broad Institute), and YeastCyc (from Stanford). Fostering exchange of metabolic models between platforms (e.g., BioCyc, KEGG, SEED, and SBML or Excel) would be desirable to facilitate comparison of models and application of models developed under different platforms.

No concerted effort has been made to collect and curate quantitative data, enzyme levels, and time-course data. There is very little or no overlap with existing projects such as iPlant (www.iplantcollaborative.org), GenBank (www.ncbi.nlm.nih.gov/genbank/), or other efforts at the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov/). Eventually, iPlant may commit more resources to metabolic systems and modeling, but at this point, the level of interest is unknown.

Details

Scientific Discovery Process (Workflows)

No workflows are available for metabolic studies other than the recently published constraint-based model workflow (Thiele and Palsson 2010). There are, however, new experimental high-throughput procedures to measure both protein and metabolite absolute values (Bennett et al. 2008), which should prove extremely useful in helping to achieve this objective.

Inputs

- Sequenced and annotated genomes will be critical for reconstructing genome-scale metabolic networks.

- Absolute values for metabolite (Bennett et al. 2008), gene expression (RNA-Seq), and enzyme levels and activities.

A number of standards can be used in metabolic studies. Largely missing are well-established standards for storing experimental data. Standards such as SBML and BioPAX (Biological Pathway Exchange) can be used to store stoichiometric information, modifiers (e.g., inhibitors), parameter values, and rate laws, but there are very few actual data standards to store flux data, enzyme levels, time-course data and such. One possibility is to use Systems Biology Results Markup Language (SBRML) (Dada et al. 2010) for data storage, together with supporting formats and ontologies that have arisen in recent years.

- Existing standards (defined as community-agreed formats implemented in more than one software tool):
    - SBML (sbml.org): Representation of stoichiometric and regulatory models, including extension packages (e.g., FBA).
    - BioPAX (www.biopax.org): Annotation standard for biological pathway data; could complement SBML.
    - SED-ML (Simulation Experiment Description Markup Language; www.biomodels.net/sed-ml): Standard for describing simulation experiments.
    - TEDDY (Terminology for the Description of Dynamics; www.ebi.ac.uk/compneur-srv/teddy): Ontology for the dynamics of biomodels.
    - KiSAO: Kinetic Simulation Algorithm Ontology (sourceforge.net/projects/kisao/).
    - SBGN: Systems Biology Graphical Notation for unambiguous visualization of pathways (www.sbgn.org).
    - SBRML: (www.comp-sys-bio.org/tiki-index.php?page=SBRML): Can be used to store data and simulation results.
    - CellML (www.cellml.org): Math-based representation of models.

- SBML extension packages conceivably could be developed to accommodate these kinds of data, using the core SBML format to anchor the metabolic network itself.

- To propose new data standards, regular focused workshops would be needed with the stakeholder community, preferably twice a year. Specification documents would need to be drawn up, test data developed, and supporting software libraries written to read and write the proposed data formats. In addition, some kind of governance structure would have to be devised to manage specifications and open-source licensing, assess

proposals, and manage revisions. The entire process should be as open as possible and involve the community at every stage.

Outputs

- A quantitative and dynamic picture of metabolism and metabolic fluxes.
- A series of new proposed standards for data storage, perhaps as package extensions to SBML and using BioPAX as additional annotation.

Tools

Existing software tools include the following.

Noncommercial

- CellDesigner (celldesigner.org). CellDesigner is a network visualization tool that can also integrate with Systems Biology Workbench (SBW) and Complex Pathway Simulator (COPASI) for carrying out simulations.
- Constraint-based reconstruction and analysis toolbox (COBRA). This is a Matlab toolbox for carrying out flux balance analysis.
- COPASI (www.copasi.org). COPASI is a general purpose and SBML-compliant simulator written in C++ (free for academic use).
- FiatFlux (www.imsb.ethz.ch/researchgroup/nzamboni/Software/fiatflux) is a Matlab toolbox for carrying out metabolic flux analysis (isotopic analysis).
- JDesigner (jdesigner.sf.net). JDesigner is an open-source network visualization and simulation tool.
- Jarnac (jdesigner.sf.net). Jarnac is a script-based tool for general-purpose simulation of reaction networks.
- KEGG Tools (www.genome.jp/kegg/)
- Metatool (von Kamp and Schuster 2006). Matatool is a C-coded open-source tool for computing elementary modes.
- Model SEED (http://seed-viewer.theseed.org/seedviewer.cgi?page=ModelView). A resource for generating, optimizing, curating, and analyzing genome-scale metabolic models.
- OptFlux (www.optflux.org). OptFlux is an open-source tool for carrying out flux balance analysis and related optimization procedures.
- Pathway Tools (bioinformatics.ai.sri.com/ptools/)
    - Computational generation of genome-scale qualitative metabolic models.
    - Editing of metabolic models.

  - Querying and visualization of metabolic models.
  - Analysis of metabolic models (reachability analysis, dead-end metabolite analysis, comparative analysis).
  - PySCeS (Olivier, Rohwer, and Hofmeyr 2005). PySCeS is a Python-based SBML-compliant tool for network analysis and simulation.
- XPP (www.math.pitt.edu/~bard/xpp/xpp.html). X-Windows Phase Plane (XPP) is a numerical analysis tool (C/C++/FORTRAN) used for general-purpose simulation and birfurcation analysis.
- SBW (www.sys-bio.org). SBW is an open source integrated SBML-compliant suite of tools for the numerical analysis of reaction networks (written in a variety of languages).

Commercial

- SimPheny (www.genomatica.com/). SimPheny is a commercial tool for flux balance analysis with good visualization functionality.
- SimBiology (Matlab). SimBiology is an SBML-compliant Matlab tool distributed by Mathworks for simulation and analysis.

References

Bennett, B. D., et al. 2008. "Absolute Quantitation of Intracellular Metabolite Concentrations by an Isotope Ratio-Based Approach," *Nature Protocols* **3**, 1299–1311.

Dada, J. O., et al. 2010. "SBRML: A Markup Language for Associating Systems Biology Data with Models," *Bioinformatics* **26** (7), 932–938.

Kacser, H., J. A. Burns, and D. A. Fell. 1995. "The Control of Flux," *Biochemical Society Transactions* **23**(2), 341–366.

Olivier, B. G., J. M. Rohwer, and J. Hofmeyr. 2005. "Modelling Cellular Systems with PySCeS," *Bioinformatics* **21**(4), 560–561.

Thiele, I., and B.A. Palsson. 2010. "A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction," *Nature Protocols* **5**(1), 93–121.

von Kamp, A., and S. Schuster. 2006. "Metatool 5.0: Fast and Flexible Elementary Modes Analysis," *Bioinformatics* **22**(5), 1930–1931. Epub May 26, 2006.

A.2 Software Requirements for Microbial Scientific Objective 1: *Reconstruct and Predict Metabolic Networks to Manipulate Microbial Function*

Summary of Scientific Objective

The scientific objective is to provide a knowledgebase that could be used to evaluate the metabolic potential of an organism, predict the phenotypic outcomes of specific metabolic or environmental interventions, and establish metabolic kinetics and fluxes. This knowledge will lead to the informed modification of enzymes or the introduction of entirely new enzymes and

pathways. This would allow the community to better determine strategies for carbon flow manipulation.

This objective is concerned with reconstructing metabolic networks, predicting the growth phenotypes of organisms from their metabolic networks, understanding the metabolic potential of an organism, and providing scientists with software tools to interrogate and visualize metabolic networks. The goal is to move beyond the current state of the art in all these respects, such as increasing the speed and automation with which metabolic networks can be reconstructed and to increase the accuracy of metabolic network predictions.

Kbase should provide access to a variety of data. Such data would include metabolic maps, both stoichiometric and regulatory; enzyme concentration and activity levels; qualitative data on enzyme regulation and known substrate, product, and cofactor dependencies; enzyme kinetic data if available; suggested kinetic rate laws or reasonable approximations; metabolic flux maps (predicted or measured) and metabolite levels; sensitivity data (rate limitingness and control coefficients) if available; time-course data on metabolite changes and enzyme concentration changes; and relevant thermodynamic data (computed or measured) on individual metabolic reactions.

Resulting Requirements

IMPACT FACTOR: _x_ HIGH ___ MEDIUM ___ LOW

A variety of interoperating software tools are needed, including tools to generate metabolic reconstructions; allow scientists to query, visualize, and curate metabolic models; and predict growth phenotypes from metabolic models.

Process of the Science (Including Workflows)

A timeline for proposed requirement development follows.

Within 6 months

- Decomposition of all existing microbial SBML and CellML kinetic models (many hundreds of models exist) into individual reaction steps and rate laws. This will provide a database of published rate laws that could be used in future models. This data should be cross-referenced to metabolic maps. For example, by selecting a reaction on a metabolic map, a user will be provided with all published rate laws associated with that reaction step.

Within 1 year

- Conversion of many constraint-based models and stoichiometric maps to be stored in standard formats such as annotated SBML and SBML FBA extension format (Bergmann and Olivier 2010). Many tools already read SBML, so it would be a natural format to use. Conversion to other formats (e.g., Matlab, COBRA, and OptFlux) can be achieved easily. It is already possible, for example, to convert COBRA format to SBML (PySCeS).
- Automatically generate genome-scale metabolic reconstructions for all DOE-relevant organisms and make them available in SBML format. This could involve a combination of

existing reconstructions from BioCyc (670 models to date), Model SEED (130 reconstructions to date), and others generated from various methods (Palsson), and including aspects of model generation not currently automated.

Within 2 years

- Establish round-trip testing of metabolic models among software tools in Kbase (e.g., COBRA, OptFlux, Pathway Tools, SBW, COPASI, CellDesigner).
- Provide better access to an online metabolic regulatory map. These data would include all modifiers that affect enzymes, including both activators and inhibitors. At the simplest level, these data could be just a list of modifiers for each enzyme but could be expanded later to include mechanisms (e.g., allosteric or covalent modification) and possibly information on Kis, Hill coefficients, and proposed rate laws.
- Create tools to compare metabolic reconstructions generated from different sources that may rely on different genome annotations. For example, do BioCyc and Model SEED agree or disagree regarding the presence of reactions and enzymes in a given organism?

Within 3 years

- Develop a series of gold-standard manually curated metabolic reconstructions for about 20 organisms of importance to the DOE mission. These reconstructions will serve as important resources and will enable automated reconstruction systems to be calibrated and assessed for their quality.

Within 5 years

- Improve fully automated metabolic reconstruction systems, increasing their speed, comprehensiveness (meaning the types of information they can infer), and accuracy beyond current levels.
- Integrate gene functional annotations and genome-scale metabolic reconstruction and simulation capabilities within the Kbase environment enabling iterative improvement of both layers of information.
- Example workflow:
    - Genome annotations are used for automated inference of initial metabolic network reconstructions.
    - A reconstruction and simulation engine automatically generates a list of gaps (missing enzymes or transporters), potential gene candidates that may resolve them, and inconsistencies (functions without context or “dangling” compounds).
    - This list by itself is of huge scientific value, as it points scientists to open research problems and missing knowledge.

- This list would also be used by existing and newly developed software tools to attempt gap filling and impose consistency on the annotations (e.g., negate "weak" functional assignments not supported by the functional context).
- Examine existing carbon 13 isotopic flux prediction software (e.g., FiatFlux) and improve it to enable better predictions for fluxes in all pathways in the cell, not just central metabolic fluxes. These software tools require metabolic network reconstructions, atom mappings between substrates and products, and experimental measurements (^{13}C labeling distributions on metabolites, biomass composition, and cellular uptake and secretion rates) and should provide estimates for intracellular fluxes (net and exchange fluxes) and confidence intervals for the estimated fluxes.
- Develop methods for determining metabolic fluxes and their confidence intervals based on time-dependent carbon 13 isotope measurements as a function of time after carbon 13 addition (before an isotopic steady-state has been reached).
- Generate methods to integrate metabolic and regulatory models and automate refinement of such integrated models. Using the previously developed gold-standard metabolic reconstructions, develop integrated metabolic and regulatory models, which will leverage regulatory network reconstructions arising from other Kbase efforts.

Mid term (3 to 5 years)

- Create a repository of growth data for organisms of importance to DOE for use in validating metabolic prediction algorithms. The repository must also include metadata associated with the experiments (e.g., media composition, temperature, and pH).
- **Model validation.** Five suggested levels for validating a model are proposed, with each level more demanding than the previous ones. The first level of validation and possibly the easiest to achieve is to compare growth and no-growth phenotypes for wildtype and mutant strains. Related to this is to compare flux balance analysis predictions with isotopic flux measurements as a way to further validate the flux balance models. The next level is to compare predicted steady-state flux and metabolite levels against experimentally measured fluxes and metabolites. Level 4 validation will test the ability of the model to predict the effect of "small" perturbations in enzyme activity levels and environmental conditions. Finally the most demanding validation test in this sequence is to compare time-course changes due to major environmental changes such as shifts in nutrients or O_2.

  Validation levels

  - Level 1: Compare growth and no-growth predictions for wildtype and mutant strains against experimental data (within 2 years).

- Level 2: Compare flux balance predictions against isotopic flux measurements (within 3 years).
- Level 3: Steady-State Validation. Compare predicted steady-state metabolite concentrations and fluxes to predicted values.
- Level 4: Perturbation Validation. Perturb enzyme levels by specified amounts (e.g., a 50% change or less) and recompute fluxes and metabolite changes that result. Compare changes with equivalent experimental perturbations (within 4 years).
- Level 5: Time-Course Validation. Apply environmental changes and significant enzyme perturbations to track time-course changes in metabolite and flux values. Validate against equivalent changes in the kinetic model (long term, 5 to 10 years).
- Measure or predict computationally or theoretically the distribution of the **degree** of rate limitingness in metabolic pathways under different conditions and in relation to the activity of negative and feedback loops. This would complement the flux, enzyme activity, and kinetic data and also be related to the validation procedures (level 4).

Long term (5 to 10 years)

- Create integrated metabolic and regulatory network models, in a largely automated fashion, that pass moderate levels of validation. After additional manual curation, these models pass high levels of validation.

User Interfaces

User interfaces should be intuitive and easy to use for a broad range of users who are not bioinformatics experts. Users will be bench biologists, computational biologists, model builders, and theoreticians. Interfaces, where possible, should be interactive and cross-platform. Visualization interfaces for locally installed software should perform as well as modern computer games and, where possible, exploit graphics processing unit (GPU) technology to enhance interactivity.

Programmatic Interfaces

Interoperation of diverse software and databases is a critical part of Kbase in general and of this objective in particular. Modern computing techniques provide a diverse set of tools for accomplishing such interoperation.

One form of interoperation is the use of web services or other Internet application programming interfaces (APIs) to interconnect databases and software tools. This approach is routinely used, for example, in the systems biology modeling community to resolve ontological terms and to access models on demand from model repositories. Web services are also a suitable method for communicating between mobile devices such as the iPhone or iPad. Web services are widely used by EBI (European Bioinformatics Institute) and other bioinformatics

resources. Modern software development environments make it relatively easy to both expose new web services and use existing ones. One reason Internet APIs are attractive is that they reduce the need to distribute software tools and to install externally developed software. Producing robustly installable software for multiple platforms can be a large burden because of incompatibility of hardware, operating system, and other system software.

In some cases, it will be appropriate to distribute algorithms and analysis methods, such as in the form of open-source reusable libraries, which can then be included in tools developed by other third-party developers. This has been an extremely useful approach used by DOE in the past for widely used numerical analysis algorithms.

Data

- **Develop gold-standard manually curated metabolic reconstructions for about 20 organisms of importance to the DOE mission** (satisfies Subobjectives 2–7). These reconstructions will serve as important resources and will enable automated reconstruction systems to be calibrated and assessed for their quality.
    - **Existing reconstructions:** *E. coli* (EcoCyc, Palsson group), ShewCyc, BeoCyc, PlantCyc, FungiCyc, YeastCyc, and Geobacter. The existing reconstructions are of variable quality; some are qualitative only, and some have received relatively little curation.
- **Develop lower-quality metabolic reconstructions for the majority of sequenced genomes** (satisfies Subobjective 1). Existing large-scale reconstruction sites such as BioCyc are likely to be extended to most genomes in the future. The BioModels.net database includes hundreds of kinetic models from a wide variety of systems; this is an important source of prebuilt curated models from the literature.

Software

Software packages needed are as follows.

- **Model management** (satisfies virtually all subobjectives). Store, query, and edit a joint metabolic, genomic, or regulatory model.
    - **Existing:** Pathway Tools has extensive capabilities in this area and should be leveraged.
- **Metabolic reconstruction** (satisfies Subobjective 1)
    - **Existing:** Pathway Tools has extensive capabilities currently focused on inference of qualitative reconstructions, although work is under way to extend its capabilities to include inference of flux balance models.
    - Model SEED has capabilities of generating metabolic reconstructions from genome annotations generated by rapid annotation using subsystem technology (RAST) and performing computational analysis of the resulting models.
- **Constraint-based analysis tools** (satisfies Subobjective 7). Predict fluxes through metabolic pathways that would lead to optimal states (e.g., maximal growth rate or

maximal ATP production), predict wildtype and mutant behaviors, generate sampling of feasible metabolic flux distributions, metabolic engineering tools, and calculation of elementary modes or extreme pathways.

  - **Existing:** COBRA, OptFlux, Metatool

- **^{13}C metabolic flux analysis** (satisfies Subobjectives 3 and 8). Estimates fluxes and confidence intervals from carbon labeling experiments, where cells are grown on ^{13}C–labeled substrates and the abundance of ^{13}C at positions in downstream metabolites is measured and used to estimate intracellular fluxes.

  - **Existing:** FiatFlux, OpenFLUX, 13CFlux

- **Network visualization** (satisfies Subobjectives 2, 3, 6, and 7). Allows visualization of metabolic networks and projection of predicted fluxes and experimental data onto these networks. Ability for users to customize the layout of networks and improvement of automated layouts would be useful.

  - Existing: Pathway Tools, Cytoscape, KEGG

- **Simulation; model analysis and data fitting; and sensitivity analysis and parameter estimation from experimental data** (satisfies Subobjectives 7–8).

  - **Existing:** Matlab, SUNDIALS (SUite of Nonlinear and DIfferential/ALgebraic equation Solvers)

- **Debugging of metabolic networks** (satisfies Subobjective 7). Identify dead-end metabolites, missing enzymes, reachability analysis.

  - **Existing:** Pathway Tools performs all these functions for qualitative metabolic models.

    Model SEED can perform some of these functions for quantitative metabolic models.

- **Gap filling for metabolic networks** (satisfies Subobjective 7). Hypothesize which genes code for missing enzymes in metabolic networks.

  - **Existing:** Pathway Tools performs this function.

    Model SEED can perform some of these functions.

    A variety of context-based and inference methods also can be used to find gene candidates, such as Phylogenetic Profiles, Gene Neighbors, Gene Clusters, and Rosetta Stone.

- **Exchange and alignment of metabolic models** (satisfies Subobjective 2).

  - **Needed:** Tools for exchanging metabolic models among software platforms. Such tools should perform syntactic conversion among platforms and semantic alignment to facilitate comparisons. Semantic alignment means establishing correspondences among the identifier spaces used by different platforms, such as establishing correspondences between the compounds and reactions

produced by the Palsson platform and those produced by the Pathway Tools platform.

- **Comparison of metabolic models and simulation results** (satisfies Subobjective 2).
  - **Existing:** Pathway Tools provides extensive tools for comparison of metabolic models. More tools are needed, or capabilities should be supplemented.
  - **Needed:** A tool that reports on differences among metabolic network models, both at the reaction and gene levels.

    A tool also is needed that summarizes differences among simulation results or flux states of a model, perhaps with results sortable by magnitude of difference so users can focus on important changes first and sortable by pathway to allow differences to be grouped in understandable ways.
- **Computational tools for metabolic engineering and pathway design** (satisfies Subobjective 6).
- **Computational tools for predicting rate limitingness in metabolic networks** (satisfies Subobjectives 4 and 8).

Standards

A number of standards can be used in metabolic studies. Largely missing are well-established standards for storing experimental data. Standards such as SBML and BioPAX can be used to store stoichiometric information, modifiers (e.g., inhibitors), parameter values, and rate laws, but very few data standards are available to store flux data, enzyme levels, time-course data, and such. Standards for storing experimental data need to be developed. One possibility is to use SBRML (Dada et al. 2010) for data storage together with supporting formats and ontologies that have arisen in recent years.

Existing standards (defined as community-agreed formats implemented in more than one software tool) are:

- SBML: Extensible representation of stoichiometric and regulatory models (www.sbml.org) including extension
- FBA-SBML: SBML Flux Balance Analysis extension package (Bergmann and Olivier 2010)
- BioPAX: Annotation standard for biological pathway data; could be used with SBML (www.biopax.org)
- SED-ML: Standard for describing simulation experiments (www.biomodels.net/sed-ml/)
- SBO: Systems Biology Ontology used to annotate the kinetic aspects of a model (www.ebi.ac.uk/sbo/)
- TEDDY: Ontology for the dynamics of biomodels (www.ebi.ac.uk/compneur-srv/teddy)
- KiSAO: Kinetic Simulation Algorithm Ontology (sourceforge.net/projects/kisao/)

- SBGN: Systems Biology Graphical Notation for unambiguous visualization of pathway (www.sbgn.org)
- SBRML: (www.comp-sys-bio.org/tiki-index.php?page=SBRML). Can be used to store data and simulation results
- CellML: Math-based representation of models (www.cellml.org)

SBML extensions packages might be developed to accommodate these kinds of data, using the core SBML format to anchor the metabolic network itself. An absent significant standard is a method for annotating a model with the assumptions used to build it and with the data sources used to parameterize it. In addition, there is currently no easy way to create a history of a model-building process so that previous models can be easily retrieved and so that model development can be traced.

Governance

Where possible, newly developed software should be open sourced under a suitable license, preferably a Lesser General Public License, Berkeley Software Distribution, or other nonrestrictive license.

To propose new data standards and maintain existing ones would require regular focused workshops, preferably twice a year, with the stakeholder community. These meetings would serve as technical discussion groups to develop a strong community structure. Specification documents would need to be drawn up, test data developed, and supporting software libraries written to read and write the proposed data formats. A governance structure would need to be devised to manage specifications, open-source licensing, and revisions and to assess proposals. The entire process should be as open as possible and involve the community at every stage.

Governance should be handled in a mixed top-down and bottom-up process. Ideas and new proposals should originate from the grassroots community so users have a stake in development. This is critical to ensure buy-in from the community. However, a top-down process also should be in place to ensure adequate development of documentation, organization of regular meetings, maintenance of websites and wikis, quality control of any developed libraries to support standards, and a mechanism to allow the community to vote for editors on a 3-year rotation. Finally, there should also be a mechanism for organizing interoperability jamborees (or Hackathons) at least once a year. Interoperability should include loading and reading the standard correctly and, most important, ensure error-free round-tripping from one tool to the next.

Definitions and References

Definitions

Qualitative metabolic models: Metabolic models that define the reactions, enzymes, and (optionally) regulatory properties of an organism's metabolic network but do not describe its flux rates, concentrations, or kinetics.

Quantitative metabolic models: Metabolic models that define some combination of flux rates, concentrations, and kinetics of a metabolic network.

Metabolic reconstruction: A qualitative or quantitative description of metabolic genes, enzymes, and transporters in an organism and the reactions these proteins catalyze.

References

Bergmann, F., and B. G. Olivier. 2010. "SBML Level 3 Package Proposal: Flux," *Nature Precedings*. (precedings.nature.com/documents/4236/version/1).

Dada, J.O., et al. 2010. "SBRML: A Markup Language for Associating Systems Biology Data with Models," *Bioinformatics* **26**, 932–938.

Kacser, H., J. A. Burns, and D. A. Fell. 1995. "The Control of Flux," *Biochemical Society Transactions* **23**, 341–366.

A.3 Microbial Scientific Objective 2: *Define Microbial Gene Expression Regulatory Networks*

The ready availability and evolution of genome-scale expression data (e.g., microarray, RNA-Seq, and single-molecule RNA measurement) and the rapid extension into new datatypes [e.g., transcription start site determination by high-throughput sequencing (HTS), RNA polymerase and regulator binding by ChIP-chip and ChIP-Seq, and quantitative regulator sequence-specificity determination by HTS methods) make the definition of microbial gene expression regulatory networks an attractive goal of the Kbase project (Bonneau et al. 2007). In the short term, inference of regulatory networks from just genome sequence and expression profiles under varied cellular conditions is possible and would be of very valuable general utility to researchers. In the longer term, validation and refinement of these models using various functional data types will allow robust correlation of regulatory networks to genome features. Interconnection of the regulatory networks with metabolic reconstructions and multidimensional annotations (two other high-priority objectives identified by the Kbase microbial group) would greatly facilitate development of microbial systems biology (Koide et al. 2009a).

A range of phylogenetically diverse microbes should be selected for initial efforts, ranging from well-characterized microbes for which extensive data exist to enable the most-informed analyses (e.g., *E. coli, S. oneidensis, G. sulfurreducens, H. salinarum,* and *D. vulgaris*), to those less well characterized (e.g., *Z. mobilis* and *C. thermocellum*), to those for which little information exists. Prioritization should be given to organisms of central relevance to the DOE mission. For the microbes chosen, the finished genome sequence would be expected to be available, together with those for a few phylogenetically related organisms. In addition, genome-wide transcriptomic data (RNA-Seq, tiling array, or spotted array) for multiple growth states would be expected. Ideally, a multilevel annotation would be advantageous. Data should focus on regulatory paradigms of greatest relevance to the microbe in question and the DOE mission. For instance, for facultative anaerobes, a focus O_2 and C regulation would be appropriate; for *D. vulgaris*, a focus on sulfur regulation might be appropriate; for *G. sulfurreducens* and *metallireducens*, the focus could be on radiation and pollutant stress; and, for *H. salinarum*, a focus on salinity and radiative and oxidative stress would be most interesting. The remainder of this scientific objective document focuses on the case of O_2 and C regulation to describe one possible focus in greater detail, but similar descriptions could be generated for the other possible foci.

Background Information

The advent of genomics and global transcriptomics has identified an ever-increasing diversity of functions regulated by O_2, including not only the expected metabolic functions (e.g., fermentation, respiration, and photosynthesis) but also many unexpected activities (e.g., cell surface proteins, sugar transport genes, nucleotide metabolism genes, transcription factors, and virulence factors) as well as gene products of unknown function. In *E. coli*, O_2 deprivation alters gene expression to promote a decrease in carbon flux through the citric acid cycle and a redirection of carbon and reducing equivalents from aerobic respiration to either fermentative

pathways or anaerobic respiratory pathways; these anaerobic pathways sustain ATP synthesis via substrate level phosphorylation or in the presence of alternative electron acceptors (e.g., NO_3^-) via oxidative phosphorylation. Genome-wide expression profiling by several groups has established that 5 to 10% of the open reading frames (ORFs) in *E. coli* K12 change expression more than twofold (Constantinidou et al. 2006; Kang et al. 2005; Partridge et al. 2007; Partridge et al. 2006). These data indicate that the functions regulated by O_2 are surprisingly varied and that our understanding of anaerobic growth of even a well-studied organism like *E. coli* remains quite limited.

In *E. coli* and related bacteria, multiple transcription factors collaborate to control gene expression in response to changes in O_2. At the center of the O_2 regulatory network lies the triumvirate of FNR, ArcA, and IscrR that collaborate with other regulators like NarL, NarP, DcuR, and PdhR to target a more limited set of genes. The complete set of regulators, their promoter target sites, and the patterns with which they interact (the regulatory network) remain incompletely defined. ArcA and its cognate two-component regulator ArcB independently control a large number of genes, although some genes are regulated by both FNR and ArcA. IscR activates a large number of genes when O_2 is present, represses others, and collaborates with FNR, ArcA, or both to control many of these genes. Further, the roles in O_2 regulation of small RNAs, transcriptional attenuation, and synergy with nucleoid structure are almost wholly unexplored, although involvement of small RNAs and nucleoid structure has recently been documented (Durand and Storz 2010; Oberto et al. 2009; Teramoto et al. 2010).

Of particular importance to the DOE mission due to its centrality to biofuel production, carbohydrate utilization appears to be altered in the presence or absence of O_2. For instance, diverse sugar transporters are among the genes whose expression is controlled by the O_2 regulatory network. *E. coli* and most bacteria encode a large number of diverse transporters for various mono-, di-, and derivatized saccharides. Some of these and other transporters are downregulated during anaerobiosis, raising key questions: (1) Are transport systems that depend on the proton motive force less favored anaerobically than aerobically, which could easily be explained by the lack of electron transport-driven proton translocation?; (2) Are new transport proteins upregulated anaerobically to replace perhaps more energetically costly aerobic transporters?; (3) Is transport of substrates that can only conserve energy via respiration selectively repressed anaerobically? The known players in carbon source regulation and their binding sites, at least under aerobic conditions, are relatively well known and consist of the catabolite activator protein (CAP or CRP) that upregulates a large number of genes when glucose is low or absent as signaled by its coactivator cAMP, FruR (a regulator of central carbon metabolism), DsgA (another global regulator of carbohydrate metabolism, particularly of the PTS system), and specific activators and repressors that control catabolic genes for various sugars (e.g., LacR and AraC). Even the relatively simple paradigms that regulate specific sugar catabolism operons lead to complex individual behaviors that are incompletely understood even without the complication of O_2 regulation (Kaplan et al. 2008). The complete extent of this network is not worked out, a fact highlighted by the existence of up to 1000 CRP-binding sites detected by genome-scale ChIP-chip (Grainger et al. 2005).

Ongoing Experiments (a very incomplete list)

- Expression profiling of both coding and noncoding RNAs is at increasingly higher levels of resolution (e.g., RNA-Seq).
- Regulatory network modeling is ongoing in a variety of laboratories (Ernst 2008; Gianchandani 2009; Michoel 2009; Huttenhower 2009; Lemmens 2009; Faith 2007; Bonneau et al. 2007).
- Visualization approaches are being explored by the Karp group at SRI International (e.g., EcoCyc and Pathway Tools); RegulonDB; the NIH EcoliHub project; MicrobesOnline (Dehal et al. 2010); the Baliga group (Bare et al. 2007; Shannon et al. 2006; Gehlenborg et al. 2010); Integrated Microbial Genomes (IMG) at the DOE Joint Genome Institute (JGI); Integrated Genome Browser at the University of North Carolina, Charlotte (Nicol et al. 2009); Galaxy at Pennsylvania State University; and recently by the Wilkerson group that is part of the Great Lakes Bioenergy Research Center (GLBRC) at Michigan State University (MSU). This is an incomplete list.
- Transcription unit definition experiments (mapping the 5' and 3' ends of transcription units) to locate promoters precisely and possible attenuation sites in the absence and presence O_2 is ongoing in several laboratories [e.g., Palsson lab at University of California, San Diego (UCSD); Kiley and Landick labs at University of Wisconsin, Madison (UW Madison); Morett lab at UNAM Biotech Institute, Cuernavaca, Mexico; and the Baliga lab at the Institute for Systems Biology (ISB), Seattle]. See recent publication from Palsson lab (Cho et al. 2009) and Baliga lab (Koide et al. 2009b).
- ChIP-chip and now ChIP-Seq experiments to map regulator and RNA polymerase occupancy on promoters and genes are ongoing in several laboratories (e.g., Palsson lab at UCSD, Kiley and Landick labs at UW Madison, and the Baliga lab at ISB).
- Metabolomics and metabolic flux experiments on the flow of carbon under aerobic and anaerobic conditions are ongoing at the GLBRC. These include studies using pure sugars and sugar mixtures, as well as complex sugar mixtures present in real biomass hydrolysates that are inputs in the developing cellulosic biofuel industry. Metabolomics experiments are ongoing at UW Madison, and metabolic flux experiments are ongoing at MSU in the Schachar-Hill lab.

Prioritization

PRIORITY: _X_ HIGH ___ MEDIUM ___ LOW

The broad goal of enabling creation of gene expression regulatory networks was recognized as a high priority in the microbial breakout group at the June 1–3 Kbase workshop. There were mixed opinions as to the particular microbes or networks that might make the best initial targets from the DOE perspective. These are reflected in the overview statements for both this scientific objective and the resulting requirements. As noted above, this objective concentrates on O_2 and C regulation as an illustrative example. Providing a broadly applicable tool for generating gene regulatory networks from RNA expression data is a high initial priority. The

network modeling capability should then be extended to additional data types, both to refine the models and to test their predictions against experimentally validated identification of transcription units, promoters, regulator binding sites, regulator binding specificity, protein-protein interactions, genetic interactions, metabolomics, and metabolic flux measurements.

Potential Benefits

As described above, O_2 regulation of carbon metabolism is a central issue for engineering biofuel-producing microbes. A complete understanding of the regulatory networks that mediate this regulation will allow researchers to specify the patterns and extents to which expression of different genes turns on as cells are shifted from aerobic to anaerobic growth conditions. Further, gaining complete control over gene regulation during anaerobiosis is essential to optimize the conservation of reducing equivalents into biofuels and may allow the efficient production of advanced biofuels like isopentanol or alkanes in anaerobic conditions where loss of reducing equivalents to O_2 can be avoided (at present only fermentation products like ethanol or butanol can be produced anaerobically with significant yields). Finally, elucidation of the regulatory network by which O_2 influences carbon metabolism is important to the advancement of science in general. Until we know the roles and interactions of the different regulatory modalities involved (e.g., repression, activation, small RNAs, and attenuation) and how these networks have evolved among microbial lineages, we will lack understanding of fundamental components in the evolution of life on Earth.

Feasibility of Success: Near, Mid, and Long Term

Several data sources are relatively straightforward with current technology (RNA-Seq, ChIP-Seq). Data tracking, manipulation, and visualization challenges must be overcome, but these align well with the central objectives of Kbase. The remaining annotation work for well-characterized microbes also is relatively straightforward and is mostly a matter of collecting information from experts in the field. Thus, building a regulatory network model of the effects of O_2 on carbon source utilization for well-characterized microbes should be achievable in the 1- to 3-year time. Extension to other microbes like *Zymomonas* will be more of a challenge at the annotation end because less is known about TU structure, O_2-responsive, and C source–responsive regulators. This may require the 5-year time frame. Likewise, extension to include other data types such as metabolics, metabolic flux, or quantitative definitions of regulator binding sites will require the 5-year time frame. Full extension of these regulatory models to enough microbes to build a picture of the evolution of O_2–responsive and C source–responsive regulation could be achieved in the 5-year time frame with a massive effort, but this is more likely an appropriate goal for the 4- to 10-year time frame.

Relevance to the Kbase Project

Defining the regulatory network for O_2 control of carbon metabolism in different bacterial lineages offers an outstanding match to the goals of the DOE Genomic Science program's Kbase project. From the Kbase perspective, it will require organization of genome-scale datasets for transcription regulator distributions, expression profiles, and potentially quantitative regulator binding assays, proteomics, and metabolomics. This will drive development of data manipulation and visualization tools that will aid microbial systems biology, both at the cutting

edge and in extension into everyday use in microbiology laboratories. From the perspective of the central mission of DOE Genomic Science, it will provide crucial information to aid in engineering microbes for biofuel production.

Synergies and Leverages: Potential Overlap with Other Projects or Funding Agencies

The project should synergize with tool development by the NIH Pathway Tools, EcoCyc, and EcoliHub projects as well as various DOE efforts such as MicrobesOnline and the JGI IMG project. It will also synergize with efforts at the DOE Bioenergy Research Centers, such as work from the GLBRC to accumulate, visualize, and model data related to O_2 regulation and sugar utilization in *E. coli* and other bacteria, work from the Joint BioEnergy Institute on *Zymomonas*, and work from the BioEnergy Science Center on microbial RNA-Seq. Given the scale of the problem, these overlaps are more likely to generate synergies than conflicts, provided adequate attention is given to coordinating efforts.

Specificity

The specific scientific question to be answered is definition of the mechanisms by which the presence or absence of O_2 influences C source utilization, which C metabolism and C transport genes are upregulated or downregulated by O_2 presence, and the nature of the regulatory network that controls these effects. More broadly, the long-term goal is to define how these regulatory networks have evolved among bacterial lineages. In the shorter term, the goal is to define this network fully in a model, well-characterized microbe and then extend to other bioenergy-relevant microbes like *Zymomonas*, *Cellvibrio*, and *Cellulomonas*. By initially concentrating on annotation, RNA-Seq, and ChIP-Seq and extending to metabolics, metabolic flux, and possibly proteomics only as the initial goals are achieved, a suitable level of specificity will be ensured.

Details

Scientific Discovery Process (Workflows)

An incomplete workflow to create a knowledgebase of O_2 regulation of carbon utilization is described below (for each microbe, steps must be pursued in parallel).

Data Collection

- Complete annotation of genes and regulators involved in O_2 regulation and C metabolism in the model microbes chosen. For well-characterized microbes, this objective can be achieved fairly easily by collecting information from researchers in the field to update existing knowledge in resources such as EcoCyc.
- Strand-specific "tiling" RNA-Seq profiles of RNA levels in different growth conditions (vary O_2 tension and with different sugar carbon sources). A strand-specific RNA-Seq workflow is available from the Sanger Center (Croucher et al. 2009).
- Define transcription start sites (TSS) used in different growth conditions (vary O_2 tension and with different sugar carbon sources). Workflows exist for these experiments (Mendoza-Vargas et al. 2009) but are imperfect and still in a stage of refinement and

improvement owing to the difficulty of distinguishing real TSSs from RNA degradation products and the failure of existing HT methods to detect some known TSSs.

- ChIP-Seq genome-scale binding profiles of RNA polymerase, sigma factor, and major regulators of carbon utilization of O_2 sensing in different growth conditions (vary O_2 tension and with different sugar carbon sources). ChIP-chip workflows exist (Mooney et al. 2009; Cho et al. 2008, 2009), and these can be readily adapted to Chip-Seq experiments, which are under way in several laboratories (e.g., UW Madison; Palsson group).
- (Optional depending on scale of project; for instance, this could be added in the 3- to 5-year time frame.) Metabolomic and metabolic flux measurements in different growth conditions (vary O_2 tension and with different sugar carbon sources). Workflows for metabolomics (Bennett et al. 2009) and metabolic flux (Martin 2010) analyses exist.
- (Years 3 to 5 or 5 to 10, or optional.) Generate data on binding-site affinities as a function of binding-site sequence for transcription factors (TFs) involved in O_2 and carbon utilization regulation. Workflows exist for these types of experiments (Hauschild et al. 2009; Jolma et al. 2010; Zhao et al. 2009; Zhu et al. 2009; Zykovich et al. 2009), but they are in early stages of adaptation to use of high-throughput sequencing.
- (Years 3 to 5 or 5 to 10.) Define transcription termination sites (TTS) used in different growth conditions (vary O_2 tension and with different sugar carbon sources). Workflows do not exist for these experiments. Although poly(A)-tailing followed by high-throughput sequencing should map RNA 3' ends, methods to distinguish legitimate TTSs from degradation products must be developed.

Bioinformatic Analyses

- Generate complete binding-site profiles for regulators under growth conditions tested. Algorithms exist to generate this information from ChIP-Seq data (Jin et al. 2009; Kuan et al. 2008), although they remain in developmental stages of continual improvement.
- Generate from RNA-Seq data a complete catalog of small RNAs binding-site profiles for regulators under growth conditions tested (vary O_2 tension and with different sugar carbon sources).
- Generate from RNA-Seq, TSS, and potentially TTS data a complete catalog of transcription units (TUs) used under the growth conditions tested (vary O_2 tension and with different sugar carbon sources). Workflows to identify TUs exist (Cho et al. 2009) but will need to be extended when TTS data becomes available.
- Generate from data a complete catalog of transcription units (TUs) used under the growth conditions tested (vary O_2 tension and with different sugar carbon sources). Workflows to identify TUs exist (Cho et al. 2009) but will need to be extended when TTS data become available.

Regulatory Network Modeling

- Convert data on transcription factor (TF) occupancy to a predictive model of TF and RNA polymerase occupancy on promoters under different growth conditions (vary O_2 tension and with different sugar carbon sources).
- Infer and predict regulatory roles of small RNAs in O_2 and carbon utilization regulation under different growth conditions (vary O_2 tension and with different sugar carbon sources).

Inputs

Inputs are largely noted above (for each microbe, steps must be pursued in parallel).

Outputs

The output will be a staged set of regulatory network models.

- Initially, a regulatory network model based only on expression data will be produced. This is a reasonable goal for the 1- to 2-year time frame as long as the data inputs are limited to expression data from RNA-Seq or related methods. At this stage, a basic network modeling tool for general use by microbial researchers should be validated and hosted.
- A second-stage output will be extension of these models to include a well-defined set of RNA-Seq, ChIP-Seq, DNA binding-site predictions, and possibly metabolic, proteomic, protein interaction or genetic interaction data. A corollary of this will be testing of the network predictions from the RNA expression data against experimentally determined knowledge of promoters and transcription units.
- A third-stage output (5-year time frame) will be to extend this to additional carbon sources so that the synergy of O_2 regulatory and sugar utilization can be understood. Additionally, models for other organisms can be produced.
- The fourth-stage output (10-year time frame) will be an extension of these network models to include multiple bacterial lineages and additional types of data inputs (quantitative proteomic data, complete metabolomic and metabolic flux data). Concentration on bioenergy-relevant microbes is appropriate, and thus this 10-year time frame could logically include gram-negative bacteria such as *Klebsiella*, *Zymomonas*, and *Cellvibrio*, and gram-positives such as *Clostridium thermocellum*.
- With complete regulatory models from anchor-point species in place, comparative genomics can be used to generate models that explain how O_2 regulation of carbon source utilization varies among bacterial lineages. See Hittinger et al. (2010) for a recent example of such models for yeast. These evolutionary models should enable rapid generation of new regulatory models for novel (newly isolated) bacteria with desirable bioconversion properties.

Tools

Several new types of software tools will be required.

- A sophisticated and facile genome browser able to visualize multiple types of genome-anchored data (e.g., RNA-Seq, ChIP-Seq, TSS, TU, gene annotation, and nucleic acid sequence) and to allow at least rudimentary calculations on these data (track averaging, ratioing, normalization, peak identification using multiple types of algorithms), storage and recall of workflows and notes, and generation of figures.
- A metabolic pathway viewer that illustrates quantitative expression levels for different enzymes. Importantly, such a viewer must be able to distinguish isozymes differentially used for growth with and without O_2 and to visualize sugar transport processes. Extension of this visualization tool to include metabolomic and metabolic flux data would be desirable if such data are included in the scope of the project.
- A regulatory network viewer that illustrates the relationships among different regulators, including small RNAs and small molecules.
- An extensible knowledgebase of regulator binding sites for each gene and TU in the regulatory and metabolic networks. This tool should allow continual improvement of Kbase by user editing and incorporation of improved regulatory network models. A good example of such a knowledgebase is EcoCyc (Keseler et al. 2009).

References

Bennett, B. D., et al. 2009. "Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in *Escherichia coli*," *Nature Chemical Biology* **5**, 593–599.

Bonneau, R, et al. 2007. "A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell," *Cell* **131**(7), 1354–1365.

Bare, J. C., et al. 2007. "The Firegoose: Two-Way Integration of Diverse Data from Different Bioinformatics Web Resources with Desktop Applications,"*BMC Bioinformatics* **8**(1), 456.

Cho, B. K., et al. 2009. "The Transcription Unit Architecture of the Escherichia coli Genome," *Nature Biotechnology* **27**, 1043–1049.

Constantinidou, C. 2006. "A Reassessment of the FNR Regulon and Transcriptomic Analysis of the Effects of Nitrate, Nitrite, NarXL, and NarQP as Escherichia coli K12 Adapts from Aerobic to Anaerobic Growth," *Journal of Biological Chemistry* **281**, 4802–4815.

Croucher, N. J., et al. 2009. "A Simple Method for Directional Transcriptome Sequencing Using Illumina Technology," *Nucleic Acids Research* **37**, e148.

Dehal, P. S., et al. 2010. "MicrobesOnline: An Integrated Portal for Comparative and Functional Genomics," *Nucleic Acids Research* **38**, D396–400.

Durand, S., and G. Storz. 2010. "Reprogramming of Anaerobic Metabolism by the FnrS Small RNA," *Molecular Microbiology* **75**, 1215–1231.

Gehlenborg, N., et al. 2010. "Visualization of Omics Data for Systems Biology," *Nature Methods* **7**, S56–68.

Grainger, D. C., et al. 2005. "Studies of the Distribution of Escherichia coli cAMP-Receptor Protein and RNA Polymerase along the E. coli Chromosome," *Proceedings of the National Academy of Sciences USA* **102**, 17693–17698.

Hauschild, K. E., et al. 2009. "CSI-FID: High Throughput Label-Free Detection of DNA Binding Molecules," *Bioorganic and Medicinal Chemistry Letters* **19**, 3779–3782.

Hittinger, C. T., et al. 2010. "Remarkably Ancient Balanced Polymorphisms in a Multi-Locus Gene Network," *Nature* **464**, 54–58.

Jin, V. X., et al. 2009. "W-ChIPMotifs: A Web Application Tool for De Novo Motif Discovery from ChiP-Based High-Throughput Data," *Bioinformatics* **25**, 3191–3193.

Jolma, A., et al. *2010.* "Multiplexed Massively Parallel SELEX for Characterization of Human Transcription Factor Binding Specificities," *Genome Research* **20**(6), 862–873.

Kang, Y., et al. 2005. "Genome-Wide Expression Analysis Indicates that FNR of Escherichia coli K-12 Regulates a Large Number of Genes of Unknown Function," *Journal of Bacteriology* **187**, 1135–1160.

Kaplan, S., et al. 2008. "Diverse Two-Dimensional Input Functions Control Bacterial Sugar Genes," *Molecular Cell* **29**, 786–792.

Keseler, I. M., et al. 2009. "EcoCyc: A Comprehensive View of Escherichia coli Biology," *Nucleic Acids Research* **37**, D464–470.

Koide, T., W. L. Pang, and N. S. Baliga. 2009a. "The Role of Predictive Modelling in Rationally Re-Engineering Biological Systems," *Nature Reviews Microbiology* **7**(4), 297–305.

Koide T., et al. 2009b. "Prevalence of Transcription Promoters Within Archaeal Operons and Coding Sequences," *Molecular Systems Biology* **5**, 285.

Kuan, P. F., H. Chun, and S. Keles. 2008. "CMARRT: A Tool for the Analysis of ChIP-Chip Data from Tiling Arrays by Incorporating the Correlation Structure," *Pacific Symposium on Biocomputing*, 515–526.

Martin, H. 2010. "Metabolic Flux Analysis via Isotopic Labeling," In DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop Report (Crystal City, Virginia, U.S. Dept. of Energy), pp.18–28.

Mendoza-Vargas, A., et al. 2009. "Genome-Wide Identification of Transcription Start Sites, Promoters, and Transcription Factor Binding Sites in E. coli," *PLoS ONE* **4**, e7526.

Nicol, J. W., et al. 2009. "The Integrated Genome Browser: Free Software for Distribution and Exploration of Genome-Scale Datasets," *Bioinformatics* **25**, 2730–2731.

Oberto, J., et al. 2009. "The HU Regulon is Composed of Genes Responding to Anaerobiosis, Acid Stress, High Osmolarity, and SOS Induction," *PLoS ONE* **4**, e4367.

Partridge, J. D., et al. 2007. "Transition of Escherichia coli from Aerobic to Micro-Aerobic Conditions Involves Fast and Slow Reacting Regulatory Components," *Journal of Biological Chemistry* **282**, 11230–11237.

Partridge, J. D., et al. 2006. "Escherichia coli Transcriptome Dynamics During the Transition from Anaerobic to Aerobic Conditions," *Journal of Biological Chemistry* **281**, 27806–27815.

Shannon, P, et al. 2006. "Gaggle: An Open-Source Software System for Integrating Bioinformatics Software and Data Sources," *BMC Bioinformatics* **7**, 176.

Teramoto, J., et al. 2010. "A Novel Nucleoid Protein of Escherichia coli Induced Under Anaerobiotic Growth Conditions," *Nucleic Acids Research* **38**(11), 3605–3618.

Zhao, Y., D. Granas, and G. D. Stormo. 2009. "Inferring Binding Energies from Selected Binding Sites," *PLoS Computational Biology* **5**, e1000590.

Zhu, C., et al. 2009. "High-Resolution DNA-Binding Specificity Analysis of Yeast Transcription Factors," *Genome Research* **19**, 556–566.

Zykovich, A., I. Korf, and D. L. Segal. 2009. "Bind-n-Seq: High-Throughput Analysis of *In Vitro* Protein-DNA Interactions Using Massively Parallel Sequencing," *Nucleic Acids Research* **37**, e151.

A.4 Software Requirements for Microbial Scientific Objective 2: *Define Microbial Gene Expression Regulatory Networks*

Summary of Scientific Objective

The scientific objectives can be broadly divided into two components. The first is to enable automated inference of gene expression regulatory networks relying principally on expression profiling data. The second is to extend these inferred networks to include additional data types, both to refine the network predictions and to test them.

Enable Automated Inference of Gene Regulatory Networks. Measurements of gene expression are becoming increasingly accurate with improvements in technology. The same is true for measurement of transcriptome structure, protein interactions, and modifications. In addition to advances in technologies for making systems biology measurements, analysis of these data is rapidly progressing at multiple scales—from discovery of operons and regulons to inference of systems-scale regulatory networks. While these rapid advances in experimental and computational methodologies pose significant challenges in standardization, they present a spectacular opportunity to infer high-quality gene regulatory networks through the integration of these data and interoperation across computational tools.

Extension and Testing of Inferred Gene Expression Regulatory Networks. In the intermediate to long term, Kbase should archive a collection of diverse systems biology datasets (e.g., transcript profiles, protein interactions, precise transcriptome structures, regulator binding sites, regulator binding specificity, and small-molecule concentrations) collected using best practices and archived in a standardized manner along with metainformation on how the experiments were conducted. In the intermediate time frame, Kbase should develop algorithms

that enable refinement and validation of regulatory networks using these comprehensive genome- and organism-scale validated datasets. In addition, Kbase should be a repository for algorithms and software tools with open and standardized APIs. Access to such data and tools across DOE-relevant organisms would enable automatic inference of gene regulatory networks and would be an extremely valuable resource to advance microbial research.

Resulting Requirements

Process of the Science (Including Workflows)

Selection of microbes on which to focus modeling efforts. A variety of phylogenetically diverse microbes should be selected for initial efforts, ranging from very well characterized organisms for which extensive data exist to enable the most informed analyses (e.g., *E. coli*, *S. oneidensis*, *G. sulfurreducens*, *H. salinarum*, and *D. vulgaris*) to those less well characterized (e.g., *Z. mobilis* or *C. thermocellum*) to those for which little information exists. Prioritization should be given to organisms of central relevance to the DOE mission. For the microbes chosen, the finished genome sequence should be available, together with those for a few phylogenetically related organisms. In addition, genome-wide transcriptomic data (RNA-Seq or tiling array) for multiple growth states would be expected to be available. Ideally, a multilevel annotation would be advantageous. Data should focus on regulatory paradigms of greatest relevance to the microbe in question and the DOE mission. For facultative anaerobes, for instance, a focus on O_2 and C regulation might be appropriate; for *D. vulgaris*, sulfur regulation; for *G. sulfurreducens* and *metallireducens*, radiation and pollutant stress; and for *H. salinarum*, salinity, radiative and oxidative stress.

Short-term, intermediate-term, and long-term tasks (*still a partial list*).

- 3 months: Identify target list of microbes and regulatory paradigms on which project will focus. These would be based on high-priority, high-value projects.
- 6 months: Complete definition of regulatory network reconstruction workflows. This implies a modifiable workflow that would have to be re-evaluated periodically to update new understanding.
- 6 months: Identify specific network inference algorithms for top-down vs. bottom-up methods. For example, the Baliga lab uses methods of the top-down approach. Use inference based on existing or expression datasets.
- 6 months: Collate existing data from microbes of interest for which sufficient data are available. (This relates to expression data.) Improve metadata associated with them. Build and use automation tools.
- 1 year: Make available for general use a capability for inference of regulatory networks from expression data (RNA-Seq, tiling array, or possibly ORF-specific array data; if generalized as an n x m matrix, any technology that generates such data could serve as input). Evaluate network inference tools and select them. Ultimately this task is an attempt to build an environment that does all network inference in an integrated way.

- 1 year: Create and make available inferred regulatory network from existing expression datasets.
- 1 year: Create a controlled vocabulary for metainformation to capture experiment design including perturbed environmental and genetic variables, media compositions, and growth conditions. Need to work with GEO or ArrayExpress to get this information captured better to be sure they have the required controlled vocabularies.
- 1 year: Provide a user interface for import and display of existing datasets, inferred transcription regulatory networks (TRN), and predicted binding sites (e.g., Pathway Tools, Cytoscape, and BioTapestry). Import inferred networks into these tools to provide better data exchange standards to allow interoperability among these tools. This would be a short -term, quick-and-dirty standard development.
- 3 years: Standardize interfaces and APIs for interoperation across selected data repositories, algorithms, and visualization software (using middleware such as Gaggle).
- 1 year: Develop web-based tutorials and demos for user "training" and Kbase dissemination. Some existing tools are not easy to use and need tutorials to help users.
- 3 to 5 years: Generate standards for regulatory network representations. This is about description of the network itself.
- 3 to 5 years: Incorporate other data types into regulatory network models (TSS, ChIP-Seq, proteomic; genome-anchored or unbiased determinations of regulator binding site specificity) for a bottom-up definition of regulatory networks. This is experimental validation of predictions made based on expression and is for validation and reconstruction. These data measure the regulatory network and currently are difficult to integrate. We need methods for capturing evidence and quality and to use them for making further predictions.
- 3 to 5 years: Extend regulatory networks to other types of regulation (transcription elongation, translation, RNA-based regulation, 3D genome architecture).
- 5 to 10 years: Extend regulatory networks to enough organisms to build a knowledgebase of the evolution of selected regulatory networks as well as network motifs through comparative network analysis capabilities such as multiple network alignment.
- 5 to 10 years: Develop a capability for coupled regulatory network models, metabolic network models, and annotation so that information is updated and exchanged.
- Develop high-quality regulatory network models by combining inference tools and bottom-up data and data from the literature. Note: Calibrate expectations appropriately.

Specific workflow for regulatory network generation by inference ("bottom-up" approach). For the organism of interest, the genome is assumed to have been completely sequenced and fully annotated and that RNA-Seq or tiling array data are available for a minimum of 10 growth curves with 6 time points and 3 biological replicates on biological conditions relevant to O_2 and C use regulation.

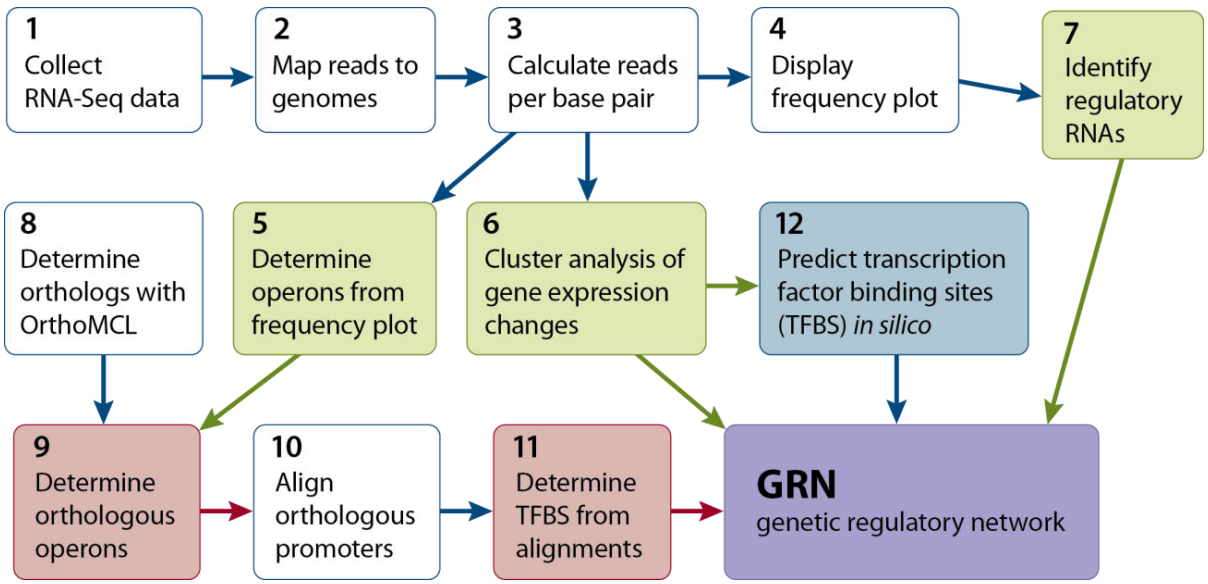


Fig. A.1. Transcriptome Analysis Pipeline for Gene Regulatory Network Prediction. White boxes are procedures we already know how to do. Green boxes are procedures that have not been determined but are expected to be fairly easy to construct (year 1). Red boxes are procedures that will be more difficult to construct (year 2). Blue boxes are optional techniques that would increase analysis accuracy. The purple box is the final product (year 2).

1. Collect RNA-Seq data and the **accompanying metadata** for each growth curve. The metadata could include optical density, substrate consumption, metabolites, temperature, and incubation condition (as comprehensive as possible). Although some data could be collected manually, Kbase would need to have the ability to store it in conjunction with RNA-Seq as an experimental project.
2. Map RNA-Seq data to genome.
3. Calculate reads per base pair (normalize and calculate expression levels of each gene and operon).
4. Display frequency plot for visual inspection and rule development for algorithms to identify the operons and regulatory RNAs in steps 5 and 7.
5. Determine operons from mapped reads (generate a separate list for each growth curve). These should include all genes, transcription initiation sites (TIS), and termination sites with the accuracy of a few base pairs for each operon.

6. Perform cluster analysis on the calculated gene expression levels to determine co-regulated operons.
7. Identify regulatory RNAs (unknown riboswitches and small regulatory RNAs) based on analysis-derived rules identified in step 4 with expert guidance.
8. Determine orthologs from multiple related genomes using OrthoMCL or some other software tool.
9. Determine orthologous promoters from multiple related genomes.
10. Align orthologous promoters using Muscle or ClustalW.
11. Determine transcription factor binding sites (TFBS) from alignments.
12. Use *in silico* TFBS prediction tools together with co-regulated operons to predict additional TFBS (import known TFBS from a database such as RegTransBase).
13. Predict genetic regulatory network.

Specific workflow enabling inference of regulatory networks for which limited information exists. This could be implemented initially with a focus on O_2 and C source regulation, but the intent is to make this available as a tool for general use via a web portal.

Inputs. Depending on the specific type of network inference analysis a user has in mind, a different combination of the following data might be necessary. Minimally, these seven types of information cover most of what is available today.

1. Measurements of transcription (with confidence values if available) in the form of an n x m matrix with n genes and m conditions (microarray or sequencing).
2. Measurements of fitness associated with systematic gene knockouts or over-expression (maybe these last two can be condensed in measures of genome-scale gene function with confidence in the form of an n x m matrix. This could be generalized as phenotype and also have associated confidence depending on how it is measured.
3. Gene interaction networks = nodes (genes), edges (interactions and type), confidence values or weights for edges.
4. Gene locations on genome; with RNA-Seq this is becoming extremely precise with direct measurement.
5. Genome sequence or individual upstream sequences (for motif detection).
6. List of predictors (transcription factors and environmental factors including metabolites).
7. Machine-readable descriptions of conditions, specifically time series information with standardized measurements of environmental factors.

User will specify an organism and import (or broadcast) the above data items. Many of these data types are stored in existing databases and can be loaded automatically

through interoperability with these data sources. Many data types (items 2 through 4) can be obtained automatically, given the organism. Item 1 may be obtained automatically from expression databases such as GEO or MicrobesOnline. Item 3 can be obtained from STRING, and some information is also accessible in MicrobesOnline. Items 4 through 6 can be obtained from NCBI, MicrobesOnline, or other databases.

As these workflows are being developed and have increasingly precise data such as RNA-Seq and can have associated confidence measures that can be carried through the analyses, this is providing a basis for comparing the precision of results between methods and laboratories that would help to improve quality and would benefit existing systems such as GEO if applied consistently.

Apply clustering and network inference. Group the genes into putative regulatory modules whose transcription is correlated over a set of conditions (e.g., cMonkey). Select a subset of known transcription factors and environmental factors that best predict the transcription levels of each module (e.g., Inferelator, CLR). Additional inputs such as motifs or protein interactions may be statistically integrated in the clustering step or the network inference step, and shared regulatory motifs can be computed. Several algorithms have been devised for clustering and discovery of regulatory influences; some are available in R and MatLab.

Outputs

- Clusters of putatively co-regulated genes or biclusters containing genes putatively co-regulated under subsets of conditions.
- cis-regulatory motifs.
- Regulatory network mapping: Influences of predictors on genes within clusters and biclusters directly or through ”and” as well as “or” operations. Confidence values for edges.

Results can be exported as raw data or presented to the user in a searchable and browsable form (e.g., in Cytoscape or BioTapestry). Users may want to start with a set of genes or a metabolic process and ask which factors are its regulators, or they may want to take a given regulator and ask what its targets are. Subnetworks can be displayed graphically along with graphical views of expression profiles and regulatory motifs and the gene content of individual clusters. Useful output also will include the ability to compute and present predictions (and confidence estimates on predictions) of the effect of TF deletions and overexpressions or environmental changes.

Scope. A user should be able to select an organism; upload, broadcast, or import expression data from public repositories or their own data; and submit a request for network inference. Metainformation on experiment design should be automatically parsed from public data, or the user should be prompted to upload this information. All other data types can be automatically parsed from public repositories. An advanced user should have privileges to change or override default settings by changing the sources of information and threshold of significance. A user should be given options for

choice of algorithms based on the amount and type of available data; the user should have access to published citations for the algorithms and basic information on workings of the algorithm in nontechnical jargon-free language. It should be possible to store a session with default or user-edited settings so the entire analysis can be recreated.

Workflow for data-driven regulatory network generation and testing and for network analysis ("bottom-up" approach). Further work after the initial implementation (years 1 to 2) would include evaluation of additional technologies and experimental verification to improve the process (5' RACE to identify additional TSSs, ChIP-Seq, microfluidic TFBS determinations, and transcription factor regulatory ligand determinations). As these data begin to accumulate from experimental efforts, possibly driven in part by solicitations from the Kbase project, the specific predictions of inferred regulatory networks should be compared to experimental data that will test TFBS predictions, transcription unit predictions, and promoter predictions. The TF-binding information can readily be combined with sigma-factor networks and genome-wide promoter architecture that are available now or currently being reconstructed by using either computational or experimental methods.

Integration of multiple datasets should be based on genome coordinates. The various types of data available will influence the confidence level of users' interpretation. Therefore any changes and interpretations need to be documented (e.g., as a comment function) with genome coordinates (base pair resolution).

The next step following regulatory network inference in a broad range of organisms implemented during years 1 to 2 of the project is to analyze and compare their topological structure and to attempt to reconstruct their evolutionary history.

The workflow for this phase of the project is as follows:

1. Implement Kbase software tools allowing users to analyze and visualize the genome-wide architecture of a regulatory network.
2. Calculate the distribution of regulon sizes and the number of regulatory inputs.
3. Perform the hierarchical layout of TRNs using a variety of algorithms (e.g., breadth-first and depth-first).
4. Minimize the number of bottom-up links. Use this layout for network visualization.
5. Identify feed-forward network motifs of different types depending on the combination of signs of regulatory interactions (activation and repression).
6. Identify and characterize cross-talk and regulatory overlap between different functional pathways.

The fraction of transcriptional regulators in bacterial genomes has been shown to systematically increase with genome size (van Nimwegen 2003). This has to be reflected in the architecture of transcription regulatory networks. To investigate these trends, one needs to analyze the topological characteristics of TRNs (e.g., average number of inputs, regulon size, and number of hierarchical layers) as a function of the organism's genome size, its lifestyle (free-living vs. parasitic), and evolutionary group. Tools need to be developed for comparing

regulatory networks in different species. These would allow users to align regulatory networks in different species using information about orthologous proteins and to trace and visualize phylogenetic profiles for network topological properties in a group of genomes selected by the user.

TFBS determination methods should be incorporated into this workflow. Examples of existing methods for this evolving technology are listed below.

References

Van Nimwegen, E. 2003. "Scaling Laws in the Functional Content of Genomes," *Trends in Genetics* **19**(9), 479–84.

Hesselberth, J. R., et al. 2009. "Global Mapping of Protein-DNA Interactions in vivo by Digital Genomic Footprinting,"*Nature Methods* **6**, 283.

Hauschild, K. E., et al. 2009. "CSI-FID: High-Throughput Label-Free Detection of DNA Binding Molecules," *Bioorganic and Medicinal Chemistry Letters* **19**(4), 3779–82.

Jolma, A., et al. 2010. "Multiplexed Massively Parallel SELEX for Characterization of Human Transcription Factor Binding Specificities," *Genome Research* **20,** 861–73.

Zhao, Y., D. Granas, and G. D. Stormo. 2009 "Inferring Binding Energies from Selected Binding Sites," *PLoS Computational Biology* **5**(12), e1000590.

Zhu, C., et al. 2009. "High-Resolution DNA-Binding Specificity Analysis of Yeast Transcription Factors," *Genome Research* **19**, 556.

Instruments to Support Achievement of the Scientific Objective

An important caveat here is that technologies change rapidly and in unanticipated ways. It is important to separate instrumentation data (raw data) from the processed data type (gene expression). For example, technologies involving array (hybridization intensities) and sequencing (sequence reads) can both generate gene expression data as an n x m matrix, where n represents genes and m represents environmental conditions over which those measurements are made. Also, these measurements are associated with statistical significance values. Herein there might be some challenges in reconciling differences in statistical methods used to analyze different datasets. We should also make executive decisions regarding the level at which data need to be archived. Technology maturity would be a consideration. For instance, there is no sense in archiving array images, but these data were important when the technology was in its infancy.

Kbase should support data from old (legacy) as well as new technologies including arrays and RNA-Seq as well as ChIP-chip and ChIP-Seq data from Solexa, ABI Solid, and 454. For the future, additional machines such as PacBio or Ion Torrent may need to be supported. These instruments produce data of particular types and sizes that will need to be stored and managed within the context of Kbase and are further described in the Data section below.

There will be potential for use of automated or high-throughput instruments for generating phenotypic data. Metadata such as optical density may be recorded manually or in spreadsheet

output from instruments, and Kbase will need to have capabilities for manual input or upload of such electronic data, which would then be integrated within the experimental project.

User Interfaces

The anticipated users will include biologists who wish to analyze their data; bioinformaticists who want to analyze data, contribute or improve methods, and use existing methods; and scientists requiring information and visual representations for scientific publications. Users are expected to come from the academic, government, and industry communities. Users below the university level are not anticipated.

Interfaces will be needed for specifying an experimental project and locating the relevant RNA-Seq and associated experimental metadata. Users will expect to have a login space where they describe their experiment, save it, and return at a later time.

Scientific data visualization is needed that renders genome annotation, gene expression information, operons, alternative transcriptional starts, and multiple sequence alignments. User interfaces will be needed for visualizing frequency plots that show depth of coverage (relative expression levels) for genes and operons and the resulting gene regulatory network model.

Programmatic Interfaces

Kbase will need to have programmatic interfaces to support specific queries such as to return a list of all experimental conditions that an organism has been exposed to for which there are gene expression data.

Also, software that determines expression levels or predicts operons or refines their prediction will need access to genome annotation. Therefore data interfaces to NCBI's Sequence Read Archive, GEO (Gene Expression Omnibus), and GenBank (bacterial genomes) will be needed and perhaps application interfaces to IMG, RAST, or DOE JGI–Oak Ridge National Laboratory annotation systems. We will also need to import known TFBS from a database such as RegTransBase.

The results of the workflow to predict gene regulatory networks will also produce data that would be output to data interfaces such as to all systems mentioned above to supply new data, support publication submission, or update annotation.

Data

In the near term, we expect to see for a given experiment several hundred files from short read sequencing technology. These files, if based on Solexa or other next-generation sequence technology, will range in size from 100 megabytes (MB) to 100 gigabytes (GB) for the next couple of years. Current size ceiling is at about 4 GB compressed for one run. Total data storage required is based on coverage and number of replicates, conditions, and time steps and therefore would be a multiplicative factor of 4 GB (180 X minimum as proposed). For the first 1 to 3 years, 30 to 100 datasets are expected per year (each dataset corresponding to studies on one microbe); the number would grow to 100 to 300 per year in the 3- to 5-year time frame when these data will be coming from many laboratories.

Database and storage resources in the terabyte to petabyte range are needed. Data reduction will play a role in keeping storage resources manageable. Online backup capabilities are needed for disaster recovery and long-term archival.

Data types that cover high-throughput technologies to interrogate the transcriptome are required for this scientific objective.

Genome sequences and a full complement of annotation features are also required. The data representation model as characterized by a GenBank record is probably not sufficient. New data models that capture gene annotations and their relationships to other annotations will be required. Annotation can exist remotely as in the case of taxonomy information housed in the NCBI taxonomy database and other NCBI Entrez data for which stable access exists through NCBI web services.

The gene regulatory network from a data structure perspective is the collection of operons, transcription factor binding sites, sigma factor binding sites; and parameters that affect kinetics. These would benefit from representation based in semantic web technology.

Relational database technology is expected to play a limited role insofar as perhaps providing structured storage of ontologies and Resource Description Framework (RDF) tuples.

Providing easy-to-use mechanisms for data download and upload is important to meet the best practices of data collection and recording of experimental protocols as close to the instruments' raw data as possible yet usable for data processing and analysis algorithms. For example, the matrix form of gene expression data can be more suitable than the raw images. Ideally, very little or no normalization should be applied to data exposed to the user, and end users would decide how they want to preprocess or normalize data before feeding it into analysis software. The mechanisms of data access and upload also should comply with governance rules, such as ones specifying policy on how soon data should be available to the community and how data should be credited or acknowledged if being used for other publications. Similar requirements hold true for algorithms and software tools, which both should provide simple-to-use examples of how to use them. Ideally, software should be accompanied with base-level tutorials on how to use it.

Software

Software should be treated at four loosely coupled distinct levels:

- Data and metadata management level.
- Middleware level (e.g., Gaggle or workflow systems such as Kepler, DagMan, Taverna).
- Algorithm and software tools level [e.g., Context Likelihood of Relatedness (CLR) and Inferelator algorithms].
- User interface level (e.g., Cytoscape, BioTapestry, MeV, DMV). Refer to *Nature Methods*' March 2010 special issue on Biological Visualization.

Note that the elements within each layer should be pluggable and replaceable. They should be capable of addressing individual layers without being imposed to go through any specific layer.

For example, an algorithmic layer should be able to directly access the data management layer. Likewise, the data flow across multiple algorithms could be orchestrated by the middleware layer, and the output of individual algorithms could be visualized by the corresponding element in the user interface layer.

It is important that data provenance information (such as versions of algorithms and tools with algorithm parameters) has been executed on the data that must be recorded in the data- and metadata-management layer for the purpose of reproducibility and interoperability of results.

Here are some representative examples of algorithms and software tools for different elements in the workflow described above:

- Clustering software: Biclustering such as cMonkey and SAMBA (Statistical-Algorithmic Method for Bicluster Analysis) as well as R's package in Bioconductor
- TFBS prediction software

Software for performing transcriptome analysis will be needed as part of the workflow and for visualization. It will integrate existing available genome annotation and provide measures of confidence. Annotation quality will be accessed based on confidence. A specific module will focus directly on improved identification of transcription factors.

Improved annotation with confidence and evidence codes will be sent back to repositories if possible.

Clustering software will be needed to group genes and operons into clusters based on patterns of regulation. Whether a part of the clustering software or of a different package, software that focuses on fine details of the operon, such as alternative transcriptional starts and stops, will be needed.

Clustering algorithms will be compute-intensive. Other methods are manageable with mid-range servers.

Data visualization software that spans genome annotation, transcriptome analysis, and clustering also will be needed.

Microbial 2: *Define Microbial Gene Expression Regulatory Networks*

Table A.1 Software Requirements for Microbial 2

[Note: "Compute" means significant processor resources are required (>100 cores), and "storage" indicates significant storage resources are required (>1 Terabyte).]

Software Purpose	Availability	Improvements Needed?	Resource Impact
Maps RNA-Seq data to genome	Few	Probably not	Storage
Cluster analysis of gene expression changes	Many	Probably	Compute, Storage
Operon determination	Few	Yes	
In silico TFBS prediction	Many	Yes	Compute
Ortholog determination	Few	Probably not	
Orthologous operon determination	None		
Promoter alignment	Few	Yes	
Promoter prediction	Few	Yes	
Gene regulatory network prediction	Few	Yes	Compute, Storage

Standards

Standards are effective only when they are adopted widely.

- Gene regulation ontology (GRO) for terms related to gene expression.
- Gene ontology (GO) for terms related to biological processes, cellular location, and gene function.
- NCBI Sequence Read Archive xml schemas for sequence read metadata.
- Genomic Contextual Data Markup Language (GCDML) xml schema for genome metadata.
- Minimum Information about a Microarray Experiment (MIAME) regards gene expression arrays but may be relevant to RNA-Seq.
- A new schema for archiving experimental metainformation. Please note that MIAME does not capture this.

Governance

A data release policy will need to be in place, most likely the current DOE policy. Kbase is expected to enforce it, implying a private login that maps to System Architecture.

Summary and Prioritization of Requirements

In silico prediction of TBFS can be postponed until other elements of the workflow are complete (mid term). Support for microarray data was considered but has not been included for the sake of simplicity. If it were part of the requirements, it might be lower priority because we believe it is phasing out. Other requirements for possible inclusion would be various kinds of validation such as 5' RACE and TFBS verification (mid term).

System Architecture Attributes

Users will expect that data they submit will be secure in accordance with the governance model. This would be the highest priority.

There could be some performance issues resulting from the choice of clustering algorithms and the amount of input data. Performance and security are architecture issues considered of highest importance for this objective.

Kbase Key Services

- Mapping RNA sequence reads to a genome
- Identifying operons and transcriptional units
- Identifying alternative transcription starts and stops
- Identifying transcription factor binding sites
- Improvements to genome annotation based on the services above
- Data structures for representing gene regulatory networks
- Query services for retrieving gene regulatory network models
- Query services for retrieving all experimental conditions to which an organism has been exposed and for which there are gene expression data

Risk Analysis and Mitigation Strategies

- Disagreement among stakeholders over objectives and approaches could undermine the project's ability to produce tools that will find widespread use. This risk is higher because various stakeholders have been involved in different stages of the Kbase initiative's development, with not all being present at a single forum that would allow generation of consensus. This is a frequent Achilles heel in large-scale bioinformatics projects. Potential for this in the Kbase project, at least among the microbial contingent, is evident. Possible mitigation would be continued efforts to achieve consensus and careful selection of goals that will achieve the widest buy-in among stakeholders.

- Unanticipated changes in technology (sequencing, microarray) that would significantly change the requirements or implementation plan. Mitigate by anticipating changes and adjusting requirements and implementation plan as soon as possible.
- Inadequate data or poor data quality that precludes a productive workflow as currently designed. Mitigate by testing typical datasets for adequacy and quality. Modify experimental protocol to correct and change minimum standards.
- Cluster analysis on these datasets requires more resources than currently anticipated. Mitigate by modifying algorithm to accept some additional error in return for performance speed. Allow clustering on subsets to manually find the optimum with reduced error.

APPENDIX B

Supporting Scientific Objective and Software Requirement Documents for Near-Term Plant Science Needs

This appendix provides the working documents for the two selected plant scientific objectives and requirements. These documents were the core output from the final DOE Systems Biology Knowledgebase (Kbase) workshop held June 1–3, 2010 (see Appendix D for the workshop report). The scientific objectives must answer the question, "What is the scientific or research goal that needs to be solved?" The related "requirements" establish workflows and provide details on the needs to accomplish these objectives. The process of identifying the scientific objective, determining its software requirements, and then developing an implementation plan from the two was described in Chapter 1. These working documents are provided as backup to their implementation plans, which are described in Chapter 3 and contain the final revised judgment by the community concerning the tasks needed for each objective.

In research, a scientific objective is satisfied by creating hypotheses and doing one or more experiments depending on the scope of the objective. For every experiment, there are rationales, protocols to be executed, a number of data inputs (data sources) and outputs (results), and analysis tools. Workflows describe this information. They are sequential procedures that describe the envisioned steps to answer questions. Workflows are the bioinformatic equivalent of an experimental protocol. Detailed workflows are bridges between the experimental research and computing communities and thus are key to translating research objectives into computing requirements that will most effectively advance the science. Workflows were developed for these objectives. From these workflows and the underlying objectives, the requirements could be defined that lead to the articulation of an implementation plan with tasks and scope to achieve the scientific and technical goals.

The first objective in the plant science area is to establish the capability to predict alterations in plant biomass properties caused by genetic or environmental changes. This capability would be based on the mining of data that reflect the complex relationships among the physical properties of plants, their genetic makeup, and the environment in which they grow. The second objective is to develop the ability to organize and analyze regulatory "omics" data to improve understanding of how plants (particularly species relevant to DOE missions) regulate gene expression. This capability will be critical for understanding genes, their action, and regulation—knowledge required to engineer plant growth and development and, in particular, biomass accumulation.

B.1 Plant Scientific Objective 1: *Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype*

The ability to gain an understanding of the genetic underpinnings of desirable plant biomass properties relevant to Department of Energy (DOE) missions (e.g., biomass yield, conversion efficiencies to biofuels, and the sequestering of soil carbon or contaminants) depends in turn on the ability to conduct co-relational assessments between molecular and phenotypic data. Identification of genes underlying the expression of desired phenotypes depends upon association of multiple genotypes contributing to a trait of interest (forward genetics) or, if a candidate gene is being investigated, an understanding of the impact that a gene product's gain or loss of function has on an extended phenotype (reverse genetics). In most cases, the complexity and plasticity of plant growth and development make it difficult to predict the impact of a perturbation in one specific gene because this phenotypic impact is rarely confined to the pathway in which the gene product operates. A key objective for the DOE Systems Biology Knowledgebase (Kbase) is to facilitate understanding of genotype-phenotype relationships by providing a platform for integrating information on genotype, extended phenotypes, and the metadata associated with field and greenhouse growth conditions. The long-term goal is to support a systems biology approach for the prediction of biomass properties by new combinations of alleles through breeding strategies or by targeted modification of genes or genetic features.

Analysis of experimental data describing the chemical and physical phenotypes of plants is a complex endeavor not within the scope of any single researcher or laboratory. Acquisition, retrieval, processing, mining, and analysis of these data depend critically on robust and standardized methods to record and track key metadata used to maintain the coherence and context of observational data. Semantic technologies, such as formal ontologies, provide vocabulary control and specify the meanings of core concepts and key features of these methods. Current semantic infrastructure, however, is inadequate to support and contextualize experimental plant phenotypic data well enough to achieve the long-term goal of predicting changes in physical properties of biomass that occur as a result of a specific environmental change, genetic modification, or allelic variation. This goal will be enabled by the formal representation of community knowledge regarding relationships among phenotype, genotype, and environment as a basis for inferring the logical implications of diverse experimental datasets.

Achievement of this ambitious goal depends on the creation of a robust semantic infrastructure for collection, annotation, and storage of diverse phenotypic and environmental datasets. These data include measurements such as those from imaging and mass spectrometry technologies that capture visible phenotypes (e.g., images) and chemotypes (e.g., analytical spectra) that are fundamentally related to yield and physiological performance and sustainability. Specifically, this infrastructure will be used as a basis for software applications that extract, quantify, and catalog phenotypic features from the data for the purpose of data mining and further analysis. This involves the association of data with relevant metadata to enable querying, modeling, clustering, and comparison of data from diverse datasets generated by different platforms. For example, it will be used to identify genotypes, haplotypes, or

mutants that have a particular phenotype and to identify all phenotypes related to a specific genotype. It also will be used as the basis for tools that generate graphical representations of the data that intuitively illustrate the important relationships of data to their biological context.

The near-term goal is to establish a controlled vocabulary that allows description and representation of genotypic, phenotypic, and environmental data; associations or correlations among them; and the resulting facilitation of quantitative comparisons of spatiotemporally varied plant phenotypic data across diverse experimental contexts. One immediate implementation of such a controlled vocabulary is the development of software tools that aid the researcher in designing experimental protocols that automatically return the appropriate semantically contextualized data and metadata. Implementation of these experimental designs will be facilitated by linkage to other applications that support the actual data collection using mobile devices such as an iPhone, a laptop peripheral device, or Laboratory Information Management System (LIMS) scanner.

Background Information

The DOE Bioenergy Research Centers and laboratories funded through research activities such as the DOE and US. Department of Agriculture (USDA) Feedstock Genomics program are engaged in, for example, understanding the basis of cell-wall recalcitrance to enzymatic hydrolysis for biofuels production. Part of this effort is the generation of hundreds of transgenic lines in which genes with functional annotations related to cell-wall biogenesis are misexpressed. In the broader plant community and through the USDA, an emphasis on sustainable growth of bioenergy feedstocks will generate perhaps thousands of target genes for which functional data will be acquired in areas such as field performance, plant growth, and environmental interactions. As high-throughput phenotyping platforms develop (e.g., saccharification screens measuring release of sugars from hydrolysis of biomass genotypes or mass spectroscopy based on sample pyrolysis), these diverse datasets need to be associated with individual genotypes. In addition, some low-throughput data with high information content are related to individual genotypes, such as structural data contained in light, confocal, or electron microscopy images. No infrastructure or platform currently allows all phenotypic data to be first collated and correlated with a specific genotype and, in the long term, allows inferential learning for new knowledge.

Another limitation to associating genotypic and phenotypic variation is the expense and inconsistency among groups in phenotyping large populations. The interplay of genotype and the environment leads to extensive variation in the physical, morphological, and biochemical properties of the plant. Interpretation of phenotypic data thus requires capture of the experimental design and metadata that formally represent the experiment's controlled components and data provenance. This ultimately can enable individual datasets to be related to each other, which can be accomplished within a single species if methodologies, genetic stocks, and reagents are in common. Between-species comparisons require a formalized vocabulary that captures not only the data but how the experiment was accomplished. Therefore, a Kbase function that serves as an interface between community-established standards and a researcher's experimental plans would be essential to enabling effective

downstream cross-comparisons between diverse experimental datasets to be performed using other recommended Kbase features.

Prioritization

Since there is an immediate need for high-level integration of both high-throughput genetic datasets and different phenotypic datasets in plants across and within experiments using a variety of different analytical approaches, the development of a rich controlled language and basic experimental standards is a high priority. This is true particularly in the diverse realms of phenotypic data, experimental design components, and associated metadata. A tangible benefit to users will be a simpler infrastructure for implementing many types of statistical analysis that may be contributed by the community or that already exist for purposes such as feature extraction from raw data, application of genetic models, and drawing associations from unrelated experiments. Incorporating them into Kbase should receive high priority. This objective will benefit directly from improved genome annotation methods (see Section 5.5), genetic diversity databases, and crop-specific databases that may not even exist for target species.

PRIORITY: __X_ HIGH ___ MEDIUM ___ LOW

Potential Benefits

The process of feedstock development and redesigning feedstock properties from the level of plant architecture and yield to biomass recalcitrance would benefit from having a unifying semantic infrastructure from which to draw inferences and organize diverse datasets. These benefits may take the form of mobile applications for actually acquiring data, experimental design tools, statistical analyses that were previously inaccessible, and high-level modeling. Success would allow adoption not only within the six target species but within the larger plant science community. Tools generally will be valuable within the larger community and particularly outside the few crops that are targets of large investments in technology. For bioenergy crops and model species, integration of data from both high- and low-throughput phenotyping experiments across species and with other omics datasets, although not a short-range goal, is nonetheless critical to refining definitions of gene function, high-level model building, interpreting orthologies, and understanding the genetic architecture of traits. This goal is dependent on being able to relate diverse datasets in a broader biological context that can then be interpreted and used for inferencing.

Feasibility of Success: Near, Mid, and Long Term

This is a high-level objective that relies on development as well as adoption of standards by the plant research community.

TERM: NEAR (1–3 years) _X_ MID (3–5 years) ___ LONG (5–10 years)

Implementation of assistive technologies for phenotyping applications can begin in a relatively short time frame for mobile applications, once an appropriate semantic structure has been created. Also, in the near term, the incorporation of functionality to perform a basic set of statistical analyses would provide some value to the user and would drive adoption.

For allowing feature extraction from complex biological data, appropriate analysis will depend on the type of data. Images at various scales are one of the most ubiquitous types of data, and allowing feature detection and extraction from images is feasible; some software already is available for performing these functions. Integrating these images into Kbase in a manner that can be used productively is critical for success. Such a capability is recognized as being feasible over the longer term but should begin in a time-sensitive manner.

The prospects for success in combining genotypic datasets with phenotypic datasets will depend on implementing algorithms for marker-assisted selection, association, and mapping of quantitative trait loci (QTL) that can be achieved with appropriate statistical genetic packages and rely on contributions from the community. Success will need to be tied to community-driven donation of these algorithms, integration of existing packages, and the willingness of individual groups to share data. This should be feasible with data-availability requirements of publicly funded programs and creative incentivizing of Kbase contributions (e.g., active contributors can be active users). Adoption of standards by individual communities is an issue that needs to come through consensus but may be hastened if incentives such as tools and analyses are provided. With these incentives, success is defined by adoption within the community and growth in the size of standardized experimenting. Data entry is expected to be steady but long term in nature.

Relevance to the Kbase Project

In preceding Kbase workshops and in the 2009 report (*Systems Biology Knowledgebase for a New Era in Biology*), significant emphasis has been placed on crops within the context of DOE's biomass research program. The activities of basic and applied plant biologists working on the six target species will amass phenotypic data from functional genomics projects, plant improvement programs, and environmental samples. In conjunction with controlled language structures, genotypic data, transcriptomics, and detailed metadata, these phenotypic data can be reused and combined in new ways to the improve overall predictive value of models envisioned in the plant scientific objective described in Section 5.3, Construct, Simulate, and Validate Plant Life Models.

Synergies and Leverages: Potential Overlap with Other Projects or Funding Agencies

Other programs, departments, and governmental agencies—such as the National Science Foundation (NSF), Plant Genome Initiative, and the Cooperative State Research, Education, and Extension Service within USDA's National Institute of Food and Agriculture—also should be involved in this activity because the underlying analytical software and modeling capability generally will be applicable to all crops. Other initiatives related to this area are oriented toward defining trait ontologies for individual groups of crops and developing database models to handle phenotypic, genotypic, and provenance data. Significant overlap within these groups' objectives should be resolved into individual contributions. In most cases, these activities are synergistic in that they have already laid much groundwork. No current efforts are ongoing in providing rapid analysis through integration of phenotypic and genotypic data within the context of plant improvement programs.

The following projects relate to this objective.

- **Existing LIMS systems and germplasm management systems.** Germinate and Agrobase (proprietary) and the International Crops Research Institute for the Semi-Arid Tropics. These provide some germplasm management tools, statistical analysis, and experimental design templates appropriate for self-fertilizing clonal crops; crops with cytoplasmic male sterility; cross-pollinating crops; open-pollinated crops; and polycrosses and synthetics common to some grass breeding programs.
- **Phenotyping projects incorporating smart phone technologies that engage the larger research community similar to those required for Kbase applications.** Epicollect (www.spatialepidemiology.net) and Phenomap (www.appstorehq.com/phenomap-iphone-113872/app/).
- **Trait indices, inventory management systems, and taxonomical data.** International Crop Information System (www.icis.cgiar.org/icis/index.php/ICIS_Concepts/).
- **Standards, protocols, and descriptors lists** (used by curators of germplasm stock centers). International Plant Genetic Resources Institute (IPGRI).
- **Gramene plant ontologies and trait ontologies.** (www.gramene.org/plant_ontology/).
- **Gene ontologies.** (www.geneontology.org). Gene ontology (GO); protein ontology (PO); and phenotype, attribute, and trait ontology (PATO).
- **Standard definitions, Crop Science Society of America, and others.** Specific systems for developmental staging.
- **Diversity databases.** Genomic Diversity and Phenotype Data Model (GDPDM) and Genomic Diversity and Phenotype Connection (GDPC) (www.maizegenetics.net/gdpc/).

Specificity

This is a higher-level objective that can be broken down into component parts such as incorporation of statistical models for determination of QTL or breeding value by different methods appropriate to the desired goal; integration of sequence and genotype standards; mobile applications; and structured ontologies. Adoption of these tools will be gradual as users perceive value to their research programs and produce appropriate plant populations and datasets. Thus, the necessary work also can be gradual, starting with the ability to integrate and assemble short reads derived from large numbers of individuals within a population and following with single-nucleotide polymorphism (SNP) calling.

Details

How is information stored for further related datasets in subsequent experiments or rounds of selection? What privacy policies would be acceptable to user groups to encourage data submission? The data should be leveraged for future use and should be connected to physical stocks to facilitate reuse. How are existing software tools and other groups' work best integrated into the context of bioenergy crops and models?

Scientific Discovery Process (Workflows)

Development of a robust semantic infrastructure for plant phenotyping research will streamline the acquisition, annotation, archiving, retrieving, processing, and mining of data that reflects the complex relationships among plants' physical properties, their genetic makeup, and the environment in which they grow. A typical workflow that would be supported by this infrastructure is described below.

Biologists, geneticists, and breeders set up field trials or controlled greenhouse experiments to collect data. The type of analysis to be performed is based on an interview process or similar interface to help identify the parameters to be measured (data) and key metadata based on prior assumptions, knowledge of biologically interesting features, previously collected data, and relationships among parameters inferred from this data. A formal, semantic model of the data and associated metadata is developed using semiautomatic methods that incorporate the parameters to be measured and relevant metadata. These methods involve user interfaces that allow the experiment designer to enter the parameters in response to relevant questions regarding the experimental setup, providing lists of appropriate (semantically defined) responses. Armed with such data models, applications that facilitate the collection and uploading of datasets (along with metadata) from the field and greenhouse are implemented using mobile devices such as an iPhone or laptop computer. These applications automatically record positional and temporal data and also could integrate local weatherstation data. The collected data are packaged and automatically uploaded to a server. Sequence data, including barcoding or similar methods for identifying specific plants and their treatment, is included in the uploaded data in relatively raw form. The uploaded data are collected and stored by a server for subsequent processing. For example, SNP data are generated for a specific population based on uploaded sequence data. Data are checked for consistency, for example, by identifying outliers that may arise due to human error. Using automated methods the data, along with associated metadata, are semantically annotated and archived in a database. Subsequent data processing leverages semantic annotations to populate specific models used to identify QTL and other biologically relevant information implied by the data.

Inputs

Datasets include the following.

- **Metadata.** Experimental design; plant starting source (seed, stem cutting, propagule); age; replication; field or greenhouse layout; position [global positioning system (GPS)]; climate and environmental conditions; log history of major environmental variations, date and time, laboratory, and institution; references; SNP indices; technology platforms; individual DNA sequences; and barcoding keys.
- **Population data.** Pedigrees, ploidy level, breeding type (outcrossing and inbreeding estimates), kinship estimates, population structure analysis, parents, and seed lots (more metadata).
- **Sequence datasets with associated metadata.** Genotype data from various platforms or from genetic diversity databases, nature of genetic modifications (mutation, transgene over- or underexpression, random or site specific), Illumina, and other platforms.

- **Phenotypic datasets.** Derived from individual traits; may include images and chemotypes.

Outputs

Results would be analysis dependent. For example, in genome-wide selection, the output would be the estimated breeding value of individual genotypes and haplotypic effects. A second example is the capability to perform graphical clustering of related phenotypes and phenotypic attributes when co-analyzing phenotypic datasets obtained from distinct characterization platforms.

Tools

1. **Phenote.** Software for recording some types of phenotypic ontology information (phenote.org).

- **MausDB.** Links mouse phenotypic data with genotypic data, metadata, and external data such as public web databases, a prerequisite for comprehensive data analysis and mining.
- **Mouse Phenome Database.** (phenome.jax.org).
- **Tools developed and listed above under related projects.**

References

Maier, H., et al. 2008. "MausDB: An Open Source Application for Phenotype Data and Mouse Colony Management in Large-Scale Mouse Phenotyping Projects," *BMC Bioinformatics* **9**,169.

Calder, R. B., et al. 2007. "MPHASYS: A Mouse Phenotype Analysis System," *BMC Bioinformatics* **8**, 183.

International Plant Phenomics Network (www.plantphenomics.com). Aims to develop, integrate, and provide novel technologies to analyze plant phenotypes.

B.2 Software Requirements for Plant Scientific Objective 1: *Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype*

Summary of Scientific Objective

Improvements in computational infrastructure are required to support and contextualize experimental plant phenotypic data to predict changes in the physical properties of biomass that occur as a result of environmental changes and genetic diversity or manipulation. Achieving this ambitious goal depends on the creation of a robust semantic infrastructure for collection, annotation, and storage of diverse phenotypic and environmental datasets. These data include measurements such as photographic images and analytical spectra that capture visible phenotypes and chemotypes fundamentally related to yield and physiological

performance and sustainability. Specifically, this infrastructure will be used as a basis for software applications that extract, quantify, and catalog phenotypic features from data for data mining and further analysis. This involves association of data with relevant metadata to enable querying, modeling, clustering, and comparison of data from diverse datasets generated by different platforms. For example, it will be used to identify genotypes or haplotypes that have a particular phenotype and to identify all phenotypes related to a specific genotype. It also will be used as the basis for tools that generate graphical representations of data that intuitively illustrate the important relationships of data to their biological context.

In the short term, computational tools are required to aid the researcher in designing experimental protocols that provide semantically contextualized data and metadata. Implementation of these experimental designs will be facilitated by software applications that support the collection of semantically contextualized data using mobile devices such as iPhones or laptop computers. The long-term goal is the formal representation of community knowledge regarding relationships among phenotype, genotype, and environment as a basis for inferring the logical implications of diverse experimental datasets.

Resulting Requirements

IMPACT FACTOR: __ X __ HIGH ___ MEDIUM ___ LOW

Attaining the scientific objective requires appropriate vocabulary standards for a wide variety of data and metadata that describe phenotypes, chemotypes, genotypes, and experiments designed to collect these data. Although several such standards and ontologies exist, they require additional expressiveness to achieve the objective. To share relevant experimental data and ensure its completeness (e.g., in terms of associated metadata), a community-approved standard for the Minimum Information for A Plant Phenotyping Experiment (MIAPPHE) would be helpful. However, such a standard does not currently exist. The development of all these standards demands a long-term committed collaboration between computer scientists and plant scientists.

Appropriate standards for semantic description and exchange of primary data (physical measurements, images, and spectroscopic data) are not available. For example, standards are required to specify plant form, morphology, anatomy, coloration, development, and function. Development of such standards may involve the extension of existing standards after identification of their shortcomings. Some measurements are species specific, so customizing standards to the target plant species may be necessary in some cases. That is, currently available standards may be too species specific, not quantitative, or outdated. In other cases, entirely new standards may be required.

Initial testing of data structures and semantic annotation protocols would be facilitated by phenotypic and genomic datasets that could be analyzed retrospectively. Conclusions obtained via the newly developed Kbase infrastructure and tools would be compared to results previously obtained by manual methods.

For target species, there are no genomic databases that support specification of genetic diversity (e.g., SNPs) within the germplasm of existing stocks. These are necessary to identify

useful correlations between genetic and phenotypic variations. Populating these genomic databases requires pipelines for calling SNPs *de novo* in the absence or presence of annotated genomes. Such pipelines exist but have not been validated for the target species, leading to high false-positive rates and low rates of validation.

Methods that detect and quantify defined features from complex data (such as photographic images or spectroscopic data) are required to facilitate correlation of data within or among datasets. Comparison of vast amounts of raw data that will be generated is not practical. Furthermore, conceptualization of correlations embedded within the diverse datasets will require representation of identifiable features rather than raw data patterns. As an example, the shape of a leaf can be parameterized and represented as a value chosen from a set of possible "leaf shapes" enumerated as specific, semantically defined entities.

Statistical methods are required to assess the consistency of data, identify correlations, and provide metrics describing the confidence of any conclusions inferred from the data (e.g., genetic or environmental causality of a phenotypic variation). The general statistical framework for such analysis largely exists but is evolving. Currently, implementation of statistical methods that incorporate both phenotypic and genotypic data (e.g., for parent selection in plant breeding experiments) is extremely slow and cumbersome, and methods tailored for processing plant phenotypic data are needed. Methods to rapidly identify and implement such statistical methods are required, for example, for efficiently selecting breeding stocks within the context of a plant improvement program.

Process of the Science (Including Workflows)

A broad range of experimental protocols will be implemented, and a typical experimental workflow is described below.

Scientists will enter relevant information describing the experimental setup directly into a LIMS system or onto a PC application. This information will then be used to develop an experiment-specific data model that will be used to automatically configure an application implemented on a mobile device used for data acquisition in the field. Complementary data for the same set of plants will be generated using a broad range of instruments but will be integrated using semantic annotation and made conformant with community standards for representation and content (e.g., the proposed MIAPPHE standard). During acquisition, the experimentalist will have the opportunity to eliminate artifactual data and impute missing data using automatic, semiautomatic, and manual methods implemented on the mobile device. Data then would be uploaded along with metadata using representations that reflect the relevant experimental data model. This data model will suggest certain types of analysis that could be run automatically or prompted via an interview process. Data processing may be as simple as performing analysis of variance (ANOVA), although complex experiments might require comparing varieties or individuals whose phenotypes were recorded in different years, experimental groups, and locations, in combination with genotypic data in a genome-wide association study and archival environmental (e.g., weather) data. This will make it possible to evaluate the impacts of temperature and moisture variation across years and locations and how they affect the identification of candidate QTL or estimated breeding value.

Instruments to Support the Achievement of Scientific Objective

A wide variety of instruments is constantly evolving for this purpose, including

- Smart phones
- Barcode scanners
- GPS systems
- Plate readers
- Image recorders (cameras)
- Physiological measurement systems (e.g., water potentiometers and fluorimeters)
- Analytical chemical analysis tools (e.g., mobile pH meters, spectrophotometers, and mass spectrometers)

User Interfaces

Users will include primarily plant scientists who develop and implement traditional forward genetics experiments (e.g., plant breeding) that are performed under a broad range of environmental conditions (i.e., in the field). Users also may include experimentalists who are performing detailed studies on gene up and down regulation. Data collection in the field may be performed by field laborers with little training or relevant experience.

Programmatic Interfaces

Programmatic interfaces include

- Data-uploading protocols that interface mobile devices with Kbase servers.
- Data integration interfaces that enable association of uploaded data with relevant metadata, semantic annotation of data, and archiving annotated data in the database.
- Interfaces that allow remote applications to execute semantically enabled queries for retrieval of annotated data.

Data

A very broad range of data will be collected and processed. This includes physical measurements (e.g., size, weight), location, color, temperature, moisture, various other environmental parameters, and analytical spectroscopic data. Much of this data may be in the form of digital images that will have to be processed for feature extraction and noise reduction. In addition, genomic and genetic data—including genetic background, ancestry, and known genetic variations (e.g., SNPs or engineered genetic modifications)—also will be required. The core challenge of this scientific objective is the collection, annotation, and integration of these diverse data.

Software

Plant 1: *Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype*

Table B.1 Software Requirements for Plant 1

Software Purpose	Availability	Improvement Needed?	Resource Impact
Data collection and uploading	Yes	Yes	
Experiment design and data model generation	Yes	Yes	
Graphical visualization	Yes	No	
Data feature extraction	Yes	Yes	
Statistical evaluation	Yes	Yes	Community Contributed
Data integration			
Knowledge inferencing			

Standards

Standards for maintenance of controlled vocabularies describing various data types and possible measurement values (e.g., units of measurements) are of primary importance. In addition, standards for exchanging diverse primary data would be helpful because they would minimize the amount of time required for data collation and integration at the server level. Finally, a global standard that specifies the content and form of a plant phenotyping experiment would be extremely useful. Such a standard (MIAPPHE) could be based loosely on the MIAME (Minimum Information About a Microarray Experiment) standard for microarray datasets.

Governance

Data privacy would be a concern among many users, and, in many cases, data would need to be kept until publication. Extensive community engagement is required for standards and encouraging data and algorithm contribution.

Summary and Prioritization of Requirements

Of primary importance is the development (within the first year) of the basic semantic infrastructure to support controlled vocabularies and representation of relevant concepts related to plant phenotype, chemotype, genotype, and growing environment. This effort includes the identification, extension, or *de novo* development of ontologies (or other formal

descriptions of relevant concepts and nomenclature), along with primary methods that can query these formal descriptions (e.g., to check data for proper classification or representation).

Within the first 2 years, development of prototype versions of a data collection platform that leverages the controlled vocabularies described above will be necessary. Tools for remote data collection and server-side tools for aggregation, annotation, and archiving of raw data will be needed.

The ability to collect and archive well-annotated data is a central requirement for success of the scientific objective. Support for a broad range of data types (e.g., spectroscopic and other high-volume data) will depend on the adoption or development of appropriate data-exchange standards.

Also required within the first 2 years are methods for assessing data consistency and confidence levels as well as statistical methods for correlation, clustering, multiple regression, and data dimension reduction (e.g., principle component analysis) that facilitate data mining and analysis. Methods for identifying, extracting, and quantifying features present in the raw data (e.g., spectroscopic peak ratios and morphological features in digital images) will be necessary to take full advantage of the data collection, annotation, and processing infrastructure described above.

Evaluation of experimental data and correlation of phenotypic and genetic variation will require genomic databases populated with information describing the genetic diversity (e.g., occurrence of SNPs) in breeding stocks or other sources of plant germplasm. Populating such databases requires a significant investment of resources, although pipelines for automated population are likely to become more available in the near future. Thus, acquisition of required data will be an ongoing process performed over the entire 10 years of the research project.

Ultimately, success of the scientific objective will depend on providing scientists with intuitive graphical representations of raw and processed data in the context of agronomy, plant biology, and genetics. The most powerful (i.e., contextualizing) graphical interfaces will depend on the availability of a robust semantic infrastructure. These graphical interfaces are expected to evolve over time, with prototype visualization tools being developed during the first year. These resources should improve considerably as general tools for semantic querying, browsing, and data representation are developed by computer scientists outside the Kbase team. Again, the challenge to Kbase will be implementation of the most powerful technologies and their adaptation to phenomics, genomics, and environmental data of interest.

Risk Analysis and Mitigation Strategies

Failure to implement at least prototype versions of the semantic infrastructure at an early stage will result in this project's failure, although development of ontologies and associated semantic methods will be an ongoing process aimed at increased expressiveness and robust performance. To avoid duplication of code development, the most general of methods supporting semantic annotation should be implemented as part of the Kbase global infrastructure so they can be applied in different contexts within Kbase. The risk is that parallel ontology development by groups outside the Kbase team will result in competing vocabularies

describing the same concepts. This is a well-recognized issue in ontology development, and methods (i.e., ontology alignment and mapping) are being developed continuously to maintain consistency among related ontologies and allow semantic disambiguation of data annotated using diverging ontological representations. Development of the basic infrastructure should be straightforward, but semantics currently is recognized as an area of intense study. Therefore, flexible infrastructure should be designed so that is capable of adapting to and incorporating new developments in this field.

Provided that robust mechanisms for vocabulary control (as described above) are available, development of the data acquisition platform should be a straightforward implementation of well-established technology, and the risk of failure is low. Data representation and exchange standards may evolve over time, most likely impacting the scope and efficiency of the data collection platform. That is, development of the platform's advanced functionality will be limited by our ability to efficiently and unambiguously represent raw and partially processed data using defined standards. This risk can be mitigated by allocating modest additional resources to development of data exchange standards, assuming that interest in the plant phenomics community is sufficient to support such standards.

Development of methods for assessing data consistency and confidence levels is likely to be straightforward with little risk of failure. In addition, a wide range of statistical methods is currently available, and adaptation of these methods to plant phenomics probably will be straightforward, provided that appropriately annotated raw data is generated using the methods described above.

A vast body of literature describes feature detection methods, and the major challenge will be selecting the most appropriate computational methods and their application to plant phenomics. Implementation of several of these methods is expected to be straightforward during the first 3 years, and more robust or specialized methods will be developed at later times. Failure to identify and adapt such methods is unlikely, although predicting their selectivity and sensitivity is not possible when applied to phenomics data.

The availability of appropriately indexed genomic diversity data cannot be predicted. Although a considerable amount of such data probably will be generated for biomass crops in the near future, much of it undoubtedly will remain proprietary. A considerable risk is that, although methods to populate genomic diversity databases using high-throughput sequencing data are in development, the availability of genetic diversity data is likely to remain limited without support from scientific journals, funding agencies, and the scientific community at large. For example, a requirement for depositing genetic diversity data as a condition for publication will lead to a considerable increase in the available amount of such data.

Development of intuitive visualization tools is unlikely to represent a significant risk. However, their effectiveness will depend on adequate allocation of resources for tool development.

B.3 Plant Scientific Objective 2: *Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Annotation, Comparisons, and Modeling*

This objective seeks to collect several key types of regulatory omics data and associated quality metadata for six target plant species: *Brachypodium*, *Chlamydomonas*, poplar, sorghum, switchgrass, and *Miscanthus*. Such information will support other plant scientific objectives related to annotation, comparison, and modeling. RNA levels as measured by expression arrays or RNA-Seq are no longer sufficient to evaluate mechanisms and networks that regulate plant transcriptomes. Kbase also must include available small RNA and target RNA information, differential RNA processing and decay information, and epigenetic marks such as DNA methylation and histone modifications. This information is important for data integration and for filling in important missing links in gene regulatory networks within a species and facilitating their comparison across two or more species. In the short 1- to 3-year term, classical transcriptomic data (microarrays and mRNA-Seq) as well as small RNA and basic proteomic data will be assembled. Epigenetic data, small RNA target data and RNA degradome data, other types of RNA processing data, and additional proteomic data will be assembled after year one, with the most developed genomes such as *Brachypodium's* beginning first. The data will be made publicly accessible with user-friendly web interfaces and will be downloadable for power users.

Background Information

Understanding which genes are regulated during growth and development and under various conditions is critical for elucidating gene function and regulatory networks. Massive amounts of genome-wide gene expression data are accumulating in plant systems that can be used to evaluate these controls at the transcriptional and post-transcriptional levels during development and in response to stimuli such as adverse environmental conditions. RNA abundance levels have been assayed routinely using microarrays and, more recently, mRNA-Seq, which is the current state-of-the-art approach (Wang et al. 2009). Also accumulating since 2005 are small RNA data from deep sequencing that report on miRNA and siRNA abundances and gene silencing potential (reviewed in Chen 2010). Other types of emerging data and data analyses are providing insights about miRNA targets and the RNA degradome (German et al. 2008; Addo-Quaye et al. 2008) as well as other aspects of RNA processing such alternative and regulated splicing and polyadenylation (Licatalosi and Darnell 2010). Beyond RNA data, proteomic data from shotgun mass spectrometry is available for some species that allows protein levels to be evaluated to examine translational control. All these data are required to effectively evaluate gene expression, and they provide essential support for the other plant scientific objectives.

Prioritization

PRIORITY: __X_ HIGH ___ MEDIUM ___ LOW

Potential Benefits

To date, we have limited understanding of how plants regulate gene expression and how this is manifested in the cell. Critical to understanding and then engineering plant growth and development for DOE missions is an informed understanding of genes, their actions, and their regulation. Our early understanding of gene regulation was focused on upstream promoters and mRNA expression levels. We are now aware of entirely new pathways of regulation involving small RNAs, post-transcriptional control, the epigenome, and more. Deep research in understanding multiple types of regulation at the DNA, RNA, and protein levels is occurring in plant, mammalian, yeast, *Caenorhabditis elegans*, and fly systems. Arabidopsis is the most studied plant with respect to the regulatory pathways affecting genes and their products.

Feasibility of Success: Near, Mid, and Long Term

Achieving the foundation of this objective in the 1- to 3-year time frame is feasible.

TERM: _X__ NEAR (1–3 years) ___ MID (3–5 years) ___ LONG (5–10 years)

Existing and new methods for generating and analyzing large-scale omics data from experiments with differential regulation need to be applied to the target species and the results analyzed and integrated. While a portion of regulatory omics data has been generated on select target species, a comprehensive effort has not been made to characterize complete datasets of regulatory omics data. Further lacking has been a funded mechanism to collate and integrate the multiple datasets and data types. These mechanisms do exist and, if applied through a core set of funded projects and funneled into Kbase, the scientific objective will have a high likelihood of success.

Relevance the Kbase Project

This objective is essential to the DOE systems biology mission. Without key data, including the acquisition of datasets, coupled with analysis of their interactions, no informed prediction of biological systems can be attempted. Naive attempts at networks are certainly possible with co-expression data, but they are highly limited and represent neither the full spectrum of what can be accomplished with current technology nor what should be completed if the mission is to understand plant species on a systems level.

Synergies and Leverages: Potential Overlap with Other Projects or Funding Agencies

Systems biology is an immature field in plant biology (Coruzzi and Gutiérrez 2009). Certainly, large-scale datasets are being generated in an array of plant species. The focus of this objective on key species relevant to the DOE mission will deepen and expand these resources. Additional major advances relevant to this objective—such as improvements in cost and throughput funneled into genome centers—will arise from the genome technology field. Algorithmic and computational improvements in network prediction and visualization are under way in model

organisms. These typically are made available to the greater research community via publications, open-source software, and collaborations. Partnering with DOE microbial systems biology scientists who have experience in constructing regulatory networks would provide great synergy. The types of datasets may overlap with iPlant and other resources, although the focus on bioenergy crops and models is unique to DOE and USDA. Therefore, synergy is expected.

Specificity

Defining the species, datasets to be collected, and species to target as well as the timetable for generating, analyzing, and integrating data is essential. While generating data is straightforward, analysis and integration will be a challenge. Thus, we recommend that only five species be undertaken for analysis and integration during the first 1 to 3 years. Omics datasets would be limited to classical transcriptomic data (microarrays and mRNA-Seq) as well as small RNA and basic proteomic data. This would provide an appropriate dataset for resolving any bottlenecks prior to generation of omics datasets for post-transcriptional regulation and epigenetic regulation, beginning first with *Brachypodium*.

Details

Scientific Discovery Process (Workflows)

Workflows exist for large-scale regulatory omics data generation. These are straightforward and can be adapted to the target species with little effort. Analysis and integration tools can be leveraged from model organism research but will need to be adapted slightly for the specified plant species.

Inputs

For a target plant species, Kbase will need an annotated genome sequence (draft or finished) as well as omics data that are or become available in the public domain, including RNA-Seq, microarray, small RNA, and basic proteomic data initially, with methylation, RNA degradome, and RNA processing subsequently.

A genome sequence in the form of an assembly and accompanied by metadata and quality information will be downloaded from public repositories [e.g., GenBank, and the DOE Joint Genome Institute (JGI)] and stored in Kbase as a flat file and in a relational database. Genome annotations will be obtained from the same sources or from other efforts to improve plant annotations and their accessibility as described in Section 5.5. Data standards exist for genome assemblies and annotation.

Omics data will be downloaded from public repositories [e.g., DOE JGI; NCBI's GenBank, Expressed Sequence Tags Database (dbEST), Short Read Archive (SRA), Gene Expression Omnibus (GEO); and others] and stored in Kbase as a flat file and in a relational database. Such data include

- Conventional expressed sequence tags (EST) and cDNA sequences (Sanger cDNAs and ESTs). Data standards exist.

- Microarray data (Affmetrix, Agilent, NimbleGen, others). Data standards exist.
- High-throughput transcriptomic sequences (e.g., Illumina, SOLiD, 454). Data standards exist.
- Small RNA data. Data standards exist.
- Proteomic data. Data standards exist.

Subsequently:

- Epigenomic data (DNA methylation, histone modification). Data standards under development.
- RNA degradome and other RNA processing data. Data standards under development.
- Additional proteomic data.

Outputs

Functional annotation of genome and gene models with respect to the following.

- Transcript levels and differential regulation.
- Small RNA levels and differential regulation.
    - Protein coding genes.
    - Noncoding RNA genes.
- Gene and transcript associations.
- Protein levels and differential regulation.

Subsequently:

- Genome epigenetic marks.
- Transcript start sites and polyadenylation sites.
- RNA decay products, some of which identify miRNA targets.
- Alternatively spliced and alternatively polyadenylated transcripts.
- Additional protein and protein modification data.

Tools

Available software:

- Conventional transcript alignment algorithms.
- High-throughput sequencing (HTS) transcript alignment algorithms.
- Reference-guided transcript assembly algorithms.
- *De novo* transcript assembly algorithms.
- HTS gene expression analysis algorithms for transcript abundance.
- Microarray analysis algorithms.
- Within platform (e.g., microarrays) expression profile clustering algorithms.
- Proteomic data analysis algorithms.
- Algorithms for analysis of HTS-derived epigenetic data (DNA methylation, ChIP-Seq for chromatin modifications).
- Algorithms for analysis of alternative splicing and polyadenylation.
- Data visualization (Cytoscape, Genome Browser).
- Co-expression analysis software.

Software needed or needing improvement:

- Small RNA analysis algorithms.
- Algorithms for analysis of HTS-derived RNA degradome data.
- Hybrid empirical (HTS) and *de novo* gene finders.
- Between platform (microarrays vs. RNA-Seq) expression profile clustering algorithms.

B.4 Software Requirements for Plant Scientific Objective 2: *Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Annotation, Comparisons, and Modeling*

Summary of Scientific Objective

This objective seeks to collect several key types of regulatory omics data and associated quality metadata for six target plant species: *Brachypodium*, *Chlamydomonas*, poplar, sorghum, switchgrass, and *Miscanthus*. This information will support other plant scientific objectives related to annotation, comparison, and modeling. RNA levels as measured by expression arrays or RNA-Seq are no longer sufficient to evaluate mechanisms and networks that regulate plant transcriptomes. Kbase must also include available small RNA and target RNA information, differential RNA processing and decay information, and epigenetic marks such as DNA methylation and histone modifications. This information is important for data integration and for filling in important missing links in gene regulatory networks within a species and facilitating

their comparison across two or more species. In the short 1- to 3-year term, classical transcriptomic data (microarrays and mRNA-Seq) as well as small RNA and basic proteomic data will be assembled. Epigenetic data, small RNA target data and RNA degradome data, other types of RNA processing data, and additional proteomic data will be assembled after year one, with the most developed genomes such as *Brachypodium's* beginning first. Data will be made publicly accessible with user-friendly web interfaces and will be downloadable for power users.

Resulting Requirements

- **Pipeline for access to omics datasets, genome sequences, and genome annotations from external sources (see Sections 3.1, 3.2, and 5.5).** This includes traditional transcript data (cDNAs, long read ESTs); RNA-Seq data generated from high-throughput sequencing platforms; small RNA data; microarray data; epigenetic data (DNA methylation, chromatin modifications), and proteomic data. The acquired data will include sequences, quality information (e.g., Q values), and associated metadata. Sources will include NCBI (GenBank, GEO, SRA), DOE JGI, ArrayExpress, and the Plant Expression Database (PLEXdb). This must leverage other efforts to improve plant annotations and their accessibility as described in Section 5.5.
- **Analysis of data assembled by the pipeline outlined in first bullet above.** Analysis will include genome mapping; normalization (across datasets and platforms); association to annotated genome features (e.g., genes, exons, and splice junctions); *de novo* assembly of applicable HTS data; clustering of expression profiles; clustering and special analysis for small RNAs; and summarization for linkage to genome annotation pipelines. Analyses will be prioritized according to current and anticipated data availability—transcriptomic (conventional, RNA-Seq, microarray) > small RNA > RNA processing and degradome > proteomic.
- **User Interface.** A user-friendly web-based interface will enable members of the community to mine omics datasets and associated annotations. Large datasets will be available for download by power users.

IMPACT FACTOR: _X__ HIGH ___ MEDIUM ___ LOW

Process of the Science (Including Workflows)

Plant biologists want to access high-quality, well-documented omics datasets associated with relevant plant gene annotations. Three main deliverables are the following:

- Consolidated platform for access to omics datasets, genome sequences, and genome annotations acquired from external sources.
- Platform for precomputed and on-the-fly analysis of plant omics datasets.
- Web-based interface enabling users to mine plant omics datasets and associated annotations.

Plant biologists want to access omics datasets in a single location (Kbase) and traverse between plant species, while being confident that the underlying data analysis and annotation methods are comparable and of consistent high quality. In addition, they will want to process new or custom omics datasets with the same tools and pipelines used to analyze the data already summarized in Kbase. To achieve these goals, Kbase will need to feature a user-friendly interface for the general user, providing summaries of gene and protein expression profiles and clusters and links to functional genomics resources (genome browser, descriptive annotations, and publications). Kbase will also need to make the analyzed and summarized data available to power users as downloadable genome-scale datasets and associated metadata. Workflows that enable the analysis of user-supplied data in Kbase will be needed. These workflows must be easy to use and made up of well-defined and -described pipeline modules.

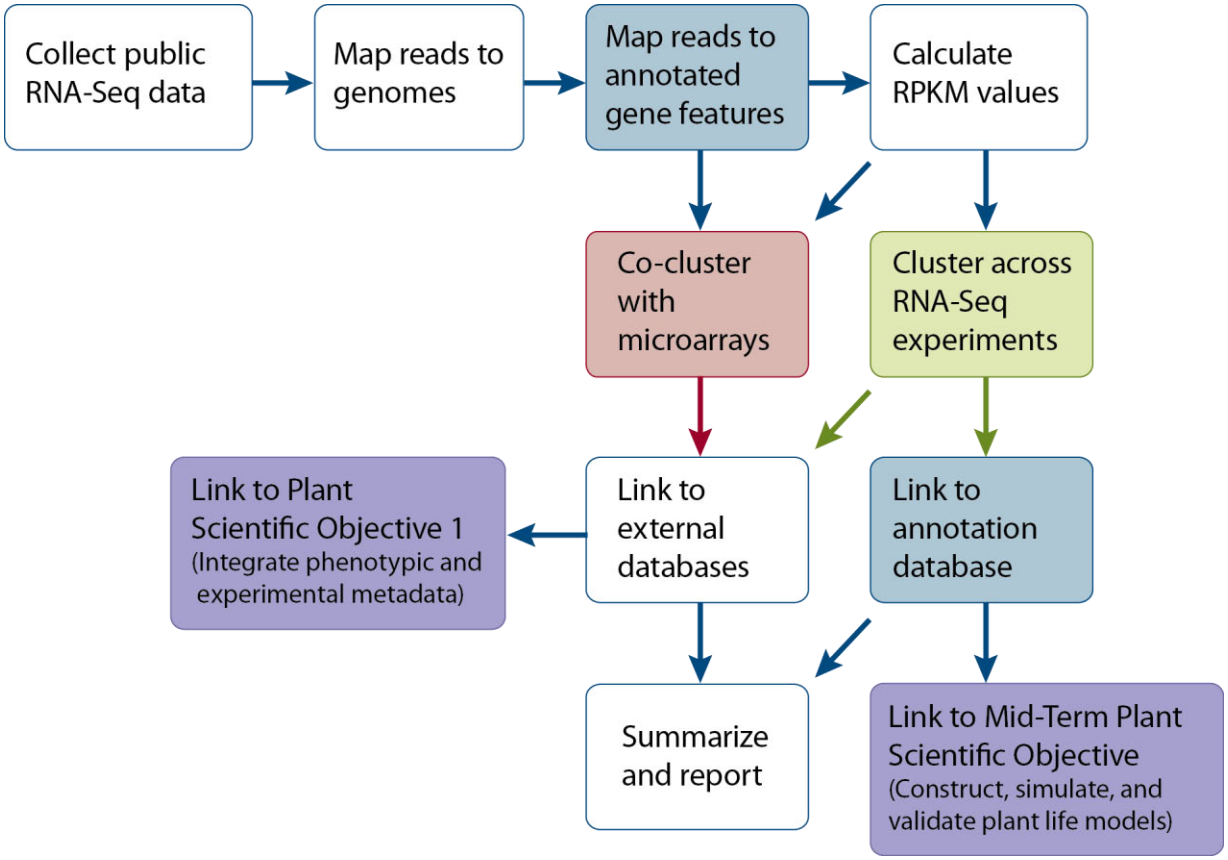


Fig. B.1. Transcriptome Analysis Pipeline for RNA-Seq Data. White boxes are established procedures. The green box is a procedure that has not been developed but is expected to be fairly easy to construct. The red box is a procedure that will require research efforts. Blue boxes depict linkages to existing and improved annotation sources (see Section 5.5). Purple boxes depict linkages to other plant scientific objectives (see Section 3.1, Integrate Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype, and Section 5.3, Construct, Simulate, and Validate Plant Life Models).

The following workflow is an example for RNA-Seq data in particular. Other types of omics data will require similar but distinct workflows.

1. Collect RNA-Seq data from NCBI SRA.
2. Map the reads to a genome (or perform *de novo* assembly of RNA-Seq data and then map assembled contigs to a genome).
3. Use genomic coordinates to associate reads with annotated gene features.
4. Calculate RPKM (reads per kilobase of transcript target per million mapped reads) values.
5. Cluster expression profiles across experiments.
6. Cluster expression profiles across platforms (RNA-Seq vs. microarray).
7. Summarize and provide linkage to external genome annotations (see plant objective described in Section 5.5) and network modeling efforts (see plant objective described in Section 5.3).
8. Enable data mining via a public web-based interface.

Instruments to Support Achievement of Science Objectives

Currently and in the near term, transcriptomic data from the Sanger, Illumina, SOLiD, and Roche 454 platforms should be supported. Microarray platforms including Affymetrix, Agilent, and Roche NimbleGen should be supported. Technology improvements and new platforms (e.g., PacBio high-throughput sequencing) will necessitate expanded platform support in the future.

User Interfaces

Users will be plant community researchers with an interest in genetics, biochemistry, molecular biology, functional genomics, comparative genomics, evolutionary biology, plant systems biology, and ecology. Users who are biologists prefer web-based graphical interfaces such as genome browsers, report pages, and simple search tools (text and sequence based). These users require downloads of small, custom datasets as well as workflows to perform custom queries and analyses. “Power users” typically are bioinformatics-savvy scientists interested in obtaining genome-scale datasets to enable genome-scale analyses. These users require large, well-documented files in standard formats appropriate for downstream analytical pipelines.

Kbase will need a user-friendly web-based interface that will enable members of the community to mine omics datasets and associated annotations. Search options should include inputs such as locus identifiers, gene symbols, GO terms, protein domains, sequences, and free text. Datasets to be mined should be definable by the user. For example, a user should be able to choose subsets of data based on interest, and metadata associated with transcriptomic datasets should be searchable by key words.

Programmatic Interfaces

Kbase will need interfaces to external data sources for genomes, annotations, and transcriptomic and proteomic datasets (e.g., from the DOE JGI and NCBI's SRA, GEO, GenBank; see Section 5.5). Software required for mapping omics data to annotated genomes typically is open source and can be placed into a workflow. Standard practices exist for (1) summarizing the output of such sequence mapping efforts into gene feature–level expression estimates, (2) normalizing results between datasets, and (3) summarizing results for visualization.

Data

For a target plant species, Kbase will need an annotated genome sequence (draft or finished) as well as existing omics data available in the public domain (e.g., RNA-Seq, small RNA, DNA methylations, histone modification, and proteomics).

Genome sequence in the form of an assembly accompanied by metadata and quality information will be downloaded from public repositories (e.g., GenBank and the DOE JGI) and stored in Kbase as a flat file and in a relational database. Genome annotations will be obtained from the same sources or from other efforts to leverage improved plant annotations (see Section 5.5).

Omics data will be downloaded from public repositories (e.g., NCBI GenBank, dbEST, SRA, GEO, DOE JGI, and others) and stored in Kbase as flat files and in a relational database. Such data include:

- Conventional EST and cDNA sequences (Sanger cDNAs and ESTs).
- High-throughput transcriptome sequences (e.g., Illumina, SOLiD, 454).
- Small RNA data.
- Microarray data (Affymetrix, Agilent, NimbleGen, and others).
- Epigenomic data (DNA methylation and histone modification).
- Proteomic data.

Software

Plant 2: *Assemble Regulatory Omics Data for Target Plant Species in Common Platforms to Enable Analysis, Comparisons, and Modeling*

Table B.2 Software Requirements for Plant 2

Software Purpose	Availability	Improvements Needed?	Resource Impact
Conventional transcript alignment algorithms	Yes	No	High
HTS transcript alignment algorithms	Yes	No	High
Reference-guided transcript assembly algorithms	Yes	Yes	High

De novo transcript assembly algorithms	Yes	Yes	High
Hybrid empirical (HTS) and *de novo* gene finders	No	Yes	High
HTS gene expression analysis algorithms	Yes	Yes	High
Microarray analysis algorithms	Yes	No	High
Within platform (e.g., microarrays) expression profile clustering algorithms	Yes	No	High
Between platform (microarrays vs. RNA-Seq) expression profile clustering algorithms	No	Yes	High
Small RNA analysis algorithms	Yes	Unknown	High
Proteomic data analysis algorithms	Yes	Unknown	High
Algorithms for analysis of HTS-derived epigenetic data (DNA methylation; ChIP-Seq for chromatin modifications)	Yes	Yes	High
Algorithms for analysis of HTS-derived RNA degradome data	Yes	Yes	High

Standards

Standards are well defined for some omics data (MIAME for microarrays) and conventional EST and cDNA sequences but are either emerging, poorly defined, or nonexistent for other types of omics data. NCBI SRA and NCBI GEO standards may be acceptable surrogates for RNA-Seq and other HTS data.

Summary and Prioritization of Requirements

Support for and analysis of HTS-derived RNA degradome and epigenetic (DNA methylation and chromatin modification ChIP-Seq) datasets and proteomic datasets can be postponed until other workflows (RNA-Seq, small RNA, microarray) are complete. (There currently is comparatively little RNA degradome, epigenetic, and proteomic data available in the public domain for the target species.) RNA-Seq data simultaneously provides information about transcript expression and structure, so RNA-processing workflows should be developed along with other workflows related to RNA-Seq.

Risk Analysis and Mitigation Strategies

- **Unanticipated slow adoption of one or more target species by the plant biology community and limitations or delays in the availability of genome-scale datasets for one or more target species.** Mitigate by prioritizing the target species for funding on genome-scale resource development as well as communication and collaboration with other funding agencies to ensure adoption and support of genome-scale research in these species.

- **Unanticipated changes in omics technology (namely, high-throughput sequencing and proteomics) that would significantly change the requirements or implementation plan for this scientific objective.** Mitigate by anticipating changes and adjusting requirements and implementation plan as soon as possible.
- **Inadequate omics data or poor data quality that prevents productive workflows as currently designed.** Mitigate by assessing available datasets for adequacy and quality and by modifying the platform by adjusting workflows to conform to available and projected datasets.
- **Anticipated algorithm and software improvements for several aspects of the project (reference-guided and de novo assembly of transcripts, analysis of RNA-Seq data for gene expression profiling, cross-platform expression clustering, and analysis of HTS-derived epigenetic and RNA degradome data) require more resources (software engineering) than currently anticipated.** Mitigate by anticipating improvements in open-source algorithms used as components of workflows and by adjusting requirements and the implementation plan as soon as possible.
- **Bioinformatic analysis on these datasets requires more computational resources (RAM, cores) than currently anticipated.** Mitigate by modifying algorithms or workflows to improve performance in terms of speed or hardware requirements, while possibly accepting increased error or other negative performance characteristics.

APPENDIX C

Supporting Scientific Objective and Software Requirement Documents for Near-Term Metacommunity Science Needs

This appendix provides the working documents for the two selected metacommunity scientific objectives and requirements. These documents were the core output from the final DOE Systems Biology Knowledgebase (Kbase) workshop held June 1–3, 2010 (see Appendix D for the workshop report). The scientific objectives must answer the question, "What is the scientific or research goal that needs to be solved?" The related "requirements" establish workflows and provide details on the needs to accomplish these objectives. The process of identifying the scientific objective, determining its software requirements, and then developing an implementation plan from the two was described in Chapter 1. These working documents are provided as backup to their implementation plans, which are described in Chapter 4 and contain the final revised judgment by the community concerning the tasks needed for each objective.

In research, a scientific objective is satisfied by creating hypotheses and doing one or more experiments depending on the scope of the objective. For every experiment, there are rationales, protocols to be executed, a number of data inputs (data sources) and outputs (results), and analysis tools. Workflows describe this information. They are sequential procedures that describe the envisioned steps to answer questions. Workflows are the bioinformatic equivalent of an experimental protocol. Detailed workflows are bridges between the experimental research and computing communities and thus are key to translating research objectives into computing requirements that will most effectively advance the science. Workflows were developed for these objectives. From these workflows and the underlying objectives, the requirements could be defined that lead to the articulation of an implementation plan with tasks and scope to achieve the scientific and technical goals.

The first objective in the metacommunities science area is to determine the metabolic role of each organism residing in a community and understand which community features provide robustness to environmental change. This will lead to improved characterizations of microbial community physiology, which is necessary to design strategies to accelerate or ameliorate microbial activity for environmental remediation.

The goal of the second objective to discover novel functions and genes within microbial communities. Data from metagenomics projects can provide the information necessary to better understand the function of poorly characterized genes. The resulting data provide actionable hypotheses about the function of many genes that have yet to be studied in detail. Additionally, scientific efforts associated with this objective will lead to the discovery of new genes that perform useful biological functions of relevance to DOE priority areas such as energy production and environmental remediation.

C.1 Metacommunities Scientific Objective 1: *Model Metabolic Processes within Microbial Communities*

Our overall scientific objective is to achieve a predictive understanding of the role of microbial communities in environments relevant to the DOE missions in bioenergy production, environmental remediation, and carbon cycling. Here, we will focus specifically on modeling the metabolic processes within the community, since this topic most directly ties into developing metagenomic workflows and single-organism systems biology tools. This predictive understanding of communities will progress in three stages. (1) Understanding: Descriptive models that provide insight into the metabolic role of the members within the community and their interactions. (2) Prediction: Predictive models that allow us to simulate the metabolic processes in the community and the response of community activity or composition to environmental conditions. (3) Manipulation: Eventually, these models will allow us to not only predict but actively drive changes in the community into desired directions (e.g., to accelerate environmental processes such as environmental remediation, cellulose degradation, or carbon sequestration in photosynthetic mats).

Background

A number of studies have been done to characterize the structure and activity of natural microbial communities (Tyson et al. 2004; Belnap et al. 2010; Ram et al. 2005; Turnbaugh et al. 2008; Wang et al. 2009; Biddle et al. 2008; Warnecke et al. 2007). Experimental technologies exist for determining the structure of microbial communities using sequencing (e.g., shotgun sequencing, pyrosequencing, and 16S rRNA sequencing reviewed in Hamady 2009) and microarray approaches (e.g., PhyloChip, DeSantis et al. 2007). However, these datasets do not always capture the complete metabolic capabilities of the microbial communities. Other experimental techniques can be used to determine the activity of the microbial community by determining the expression of genes and proteins in the community using metatranscriptomics (He et al. 2010a; Vila-Costa et al. 2010), functional gene arrays (e.g., GeoChip; He et al. 2010b), or metaproteomic methods (Belnap et al. 2010; VerBerkmoes et al. 2009; Lacerda and Reardon 2009). These datasets are useful for determining organisms and pathways that are biologically or metabolically active in the microbial community. Computational models of microbial species can be rapidly developed from annotated genomes (Thiele and Palsson 2010), and recent studies have begun to model the interactions among two and three microorganisms (Stolyar et al. 2007; Taffs et al. 2009; Miller et al. 2010). Scaling up these methods to characterize environmental microbial physiology can proceed through two broad approaches. These are (1) a bottom-up (microbes-to-communities) approach, where the microbes are isolated and cultured in the laboratory and integrated, evaluated, and modeled in a defined community and (2) a top-down (community-to-microbes) metagenome-based approach, where the DNA from environmental samples is directly sequenced for understanding the metabolic potential through bioinformatics and pathway reconstruction.

The first approach is intricately linked to the microbial systems biology workflows discussed in other breakout groups. However, such models are valuable for developing sophisticated

models of microbial communities. Hence, there is a need to develop workflows that facilitate the analysis of metagenomes as well as for the dynamics of microbial communities.

A wide variety of metagenome studies in diverse environments (e.g., acid mine drainage, enhanced biological phosphorus removal, termite gut, rumen, compost, soil, ocean, and sediment) are already under way. In many of these processes, the biotransformation activity is related to the integrated phenotype of the microbes present in the community. To enhance these biotransformation activities, characterizing the metabolic pathways of constituent members and linking individual organisms to their substrate and product profiles are important.

Prioritization

PRIORITY: _X__ HIGH ___ MEDIUM ___ LOW

Potential Benefits

Single microbial strains rarely, if ever, act alone, and the complex network of interactions among microbial populations drives all the major metabolic processes in the world around us. The proposed objectives will lead to improved characterization of the physiology of microbial communities, and such a characterization is necessary to design strategies to either accelerate biotransformation activity (e.g., uranium bioremediation) or ameliorate the outcome (e.g., acid mine drainage). Understanding metabolic interactions and substrate preferences of relevant organisms is anticipated to assist in developing design strategies to optimize biotransformation activity. If successful, the benefits can be valuable in providing a framework for analysis of microbial physiology in any impacted environment and can lead to lowered treatment costs as well as accelerated removal strategies.

Feasibility of Success: Near, Mid, and Long Term

TERM: ____ NEAR (1–3 years) __X_ MID (3–5 years) ___ LONG (5–10 years)

Clear and achievable short- and medium-term goals can be formulated both for the top-down (metagenomic) and the bottom-up (multispecies models) approaches. A “mock-up” integration of these two approaches can be achieved in the short term, although full integration into a single analysis workflow is a medium-term task. Extensions to this basic modeling paradigm to integrate additional data types, tackle spatial and temporal variation are medium to long term, while fully leveraging the predictive understanding of these communities to guide and control them is a long-term goal. At all stages of this process, the availability of relevant simplified communities (whether artificial co-cultures, low-complexity natural communities, or enrichments) should significantly speed up tool development and allow a gradual buildup to more complex communities.

Relevance to the Kbase Project

This scientific objective directly addresses the systems biology of communities and so will be able to leverage many tools and resources being developed for systems biology of single

organisms. In addition, it will provide an analysis workflow to extract highly useful knowledge from metagenomic sequence data, which probably will become one of the most voluminous data types that Kbase is likely to handle. So, on the one hand, this type of analysis will be an essential component to fulfill the overall goals of Kbase. On the other hand, data-integration aspects and systems biology tools and resources provided by the Kbase platform will be essential to the success of this scientific objective.

Synergies and Leverages: Potential Overlap with Other Projects or Funding Agencies

Existing metagenomic analysis tools such as IMG/M (Integrated Microbial Genomes with Microbiome samples), MG-RAST (Metagenome Rapid Annotation Using Subsystem Technology), or CAMERA (Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis) currently provide some initial preprocessing needed for the analysis presented here, including assembly and functional annotation. However, none currently provides satisfactory phylogenetic binning tools or, more important, the powerful systems biology analysis tools necessary to take functional analysis to the higher level. Platforms such as Pathway Tools include inference engines to predict pathways from potentially incomplete data (Dale, Popescu, and Karp 2010) or fill holes in predicted pathways (Green and Karp 2004), but they are not adapted to the noise and incompleteness inherent in metagenomic data. Databases funded by federal grants (e.g., BioCyc and KEGG) have some components necessary for the metabolic modeling parts of this workflow; however, there is no clear integrated database or simulation efforts. Leveraging existing databases to accelerate development efforts would be useful. There may be potential overlap with some human microbiome projects at the National Institutes of Health (NIH). This probably would be more on the metagenomics side than on the metabolic modeling side, where the human microbiome project will result in a large amount of data relating to the structure and activity of microbial communities that interact with their human host. Some computational and experimental methods that are developed could be applicable to some datasets and analysis being done under this program.

Specificity

The broad scientific objective of achieving a predictive understanding of the role of microbial communities will necessarily include components other than metabolic modeling. We have decided to focus specifically on the metabolic modeling aspect precisely because it integrates so well with the data, tools, and resources planned for Kbase. Other aspects of the role of microbial communities may require further development of more coarse-grained models or microbial ecology tools not specifically covered in this section, although metabolic models will clearly be of some use there as well. Some of these other requirements are covered under different scientific objectives. For example, the scientific objective focusing on comparative community analysis (see Section 5.1 Analyze Understudied Microbial Phyla) covers a range of tools which that will be invaluable to ecological analysis.

Details

Scientific Discovery Process (Workflows)

Workflows for constructing metabolic models from an individual organism's genome sequences have been developed (Thiele and Palsson 2010). While many steps for generating metabolic models for microbial communities may be similar, missing information (such as missing genes) may be a more challenging problem when dealing with metagenomic datasets. Here, workflows for the microbial community modeling framework for the bottom-up and the metagenome-analysis method for the top-down approach are discussed, and the input, outputs, and tools for each are provided. Such workflows are currently available in a scattered manner, but integrating them into a common framework and database would be useful.

For the analysis of defined communities, the first step will involve incorporation of all individual metabolic models into a common environmental model. This will require information on substrates metabolized and secreted by community members as well as common nomenclature for the exchanged metabolites (Zhuang et al. 2010). In addition, kinetics of substrate uptake and secretion as well as biomass yields will be critical in developing such community models.

Reconstruction of Community Metabolic Models from Metagenomic Data

- **Metagenomic Sequencing and Assembly.** Methods in this area are fairly well established, although further development may be needed to deal with even higher-volume next-generation sequencing technologies. Improvements can be made in assembly of shorter reads and in dealing with systematic sequencing errors (e.g., 454 frameshifts).
- **Phylogenetic Binning.** Several methods have been developed, but further validation is needed. Significant advances could be achieved by combining the best features of multiple binning methods (e.g., use phylogenetic markers where available, then use k-mer frequency-based methods on contigs too short to contain useful marker genes).
- **Pathway Reconstruction from Noisy, Incomplete Data.** Most phylogenetic bins will tend to be incomplete, and the binning itself may be of differing qualities. Pathway reconstruction methods will need to take these sources of noise and uncertainty into account in a systematic manner. Knowledge of reference genome content and pathway content may help guide reconstruction.
- **Metabolic Modeling of Community Members.** Given an inventory of pathways present in each phylogenetic bin, construct a functioning metabolic simulation model for each bin. This problem is identical to metabolic network inference in single genomes, except that some of the probabilistic nature of the previous step may need to be taken into account.
- **Combine Community Metabolic Model.** Integrate component metabolic models into a whole-community model. Simulation frameworks should be capable of integrating models of individual organisms into a community model with the ability to integrate customized workflows for simulation purposes.

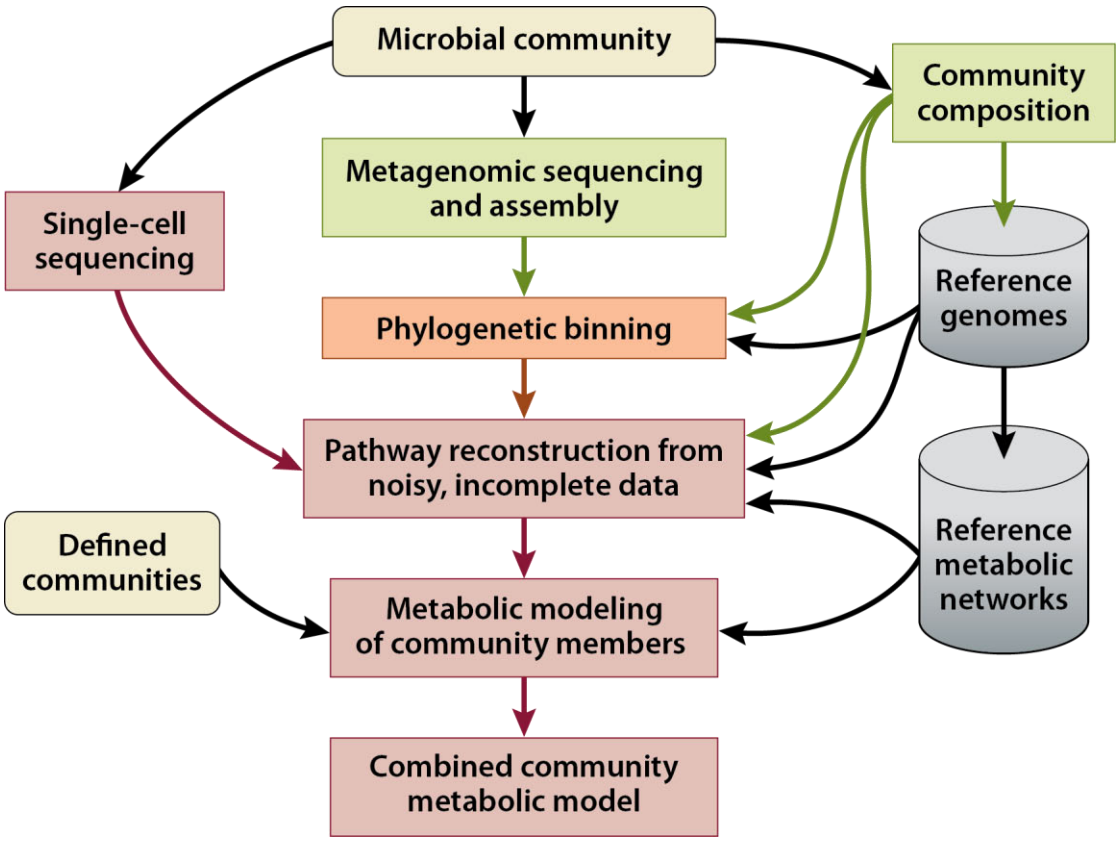


Fig. C.1. Workflow for Reconstructing Metabolic Models from Metagenomic Data. The "bottom-up" approach involves data from individual species or members, and the "top-down" approach relies on whole-metacommunity analysis. In addition to available genomic sequencing technologies, Kbase must prepare to use other data in the near future. Green modules already are reasonably well established. The orange box shows an available method that needs more development. Red boxes show capabilities that still require significant new development.

Inputs

- Metagenomic sequence data.
- Community composition (e.g., based on 16S libraries, pyrotags, or PhyloChips).
- Sequenced genomes and reconstructed enzyme and pathway repertoire (e.g., BioCyc pathway or genome database) for a wide range of reference organisms from all phyla, including poorly sequenced or unsequenced phyla.
- Consistent, nonredundant database of biochemical reactions with metabolite stoichiometries and reaction names, including unique identifiers and standard naming conventions.
- Both elementally and charged-balanced biochemical reactions.

- Thermodynamic information such as Gibbs free energy for all biochemical reactions.
- Reconstructed metabolic networks of individual organisms.
    - Standard format including Systems Biology Markup Language (SBML) file with functional information (e.g., which pathways are involved, amino acid metabolism, and glycolysis).
    - Consistent naming scheme.
- Biological data of individual organisms.
    - Kinetics of substrate uptake rates.
    - Thermodynamics.
    - Growth physiology.
- Specialized environmental data about the niche in which the community lives.
    - Geochemistry.
    - Spatial heterogeneity.
    - Surface chemistry.

Additional data that can be mapped to these models or help guide the modeling effort.

- Metatranscriptomics.
- Metaproteomics.
- Spatial localization of small molecules (e.g., nutrients, metals, and metabolites), proteins, and cells.
- Microscopy, fluorescence *in situ* hybridization (FISH), nano-secondary ion mass spectrometry (SIMS).

Outputs

- Overview of metabolic repertoire and compartmentalization into component species.
- Community predictions of
    - Dynamic shifts in community composition.
    - Metabolic interactions between the community and the environment.
    - Interspecies metabolic interactions.

Tools

- One curated, extremely well-maintained encyclopedia of biological data that includes all genomic, biochemical, physiological, experimental (transcriptomic, metabolomic) data with standard naming conventions for genes, reactions, metabolites, and pathways.

- Open-source automated network reconstruction tools with integrated mathematical toolbox, such as COBRA (constraint-based reconstruction and analysis; Becker et al. 2007) for simulating metabolism and analyzing network properties.
- Tools or algorithms for automated prediction of localization and compartmentalization of biochemical reactions (e.g., cytosolic, periplasmic, and mitochondrial).
- Develop methods to represent uncertainties associated with the (1) reaction network inference from metagenomic data and (2) uncertainty in compartmentalization associated with binning algorithms.
- Open-source tools for integrating metabolomics and gene-expression data with model-predicted flux data so that any conflict among them can be identified.
- Tool for predicting the metabolic role of any member in the community.
- A high-quality reconstructed network in a standard format with a consistent naming scheme.
- Modeling methods and tools for coupling the network with additional biological data of individual organisms.
- A computational framework that integrates individual organism models. Must be highly adaptive, allowing current and future models to be integrated.
- Visualization tools for community composition and metabolic activities.

References

Belnap, C. P., et al. 2010. "Cultivation and Quantitative Proteomic Analyses of Acidophilic Microbial Communities," *ISME Journal* **4**, 520–30.

Biddle, J. F., et al. 2008. "Metagenomic Signatures of the Peru Margin Subseafloor Biosphere Show a Genetically Distinct Environment," *Proceedings of the National Academy of Sciences of the United States of America* **105**, 10583–588.

Dale, J. M., L. Popescu, and P. D. Karp. 2010. "Machine Learning Methods for Metabolic Pathway Prediction," *BMC Bioinformatics* **8**, 15.

DeSantis, T. Z., et al. 2007. "High-Density Universal 16S rRNA Microarray Analysis Reveals Broader Diversity than Typical Clone Library when Sampling the Environment," *Microbial Ecology* **53,** 371–383.

Garcia Martin, H., et al. 2006. "Metagenomic Analysis of Two Enhanced Biological Phosphorus Removal (EBPR) Sludge Communities," *Nature Biotechnology* **24**, 1263–69.

Green, M., and P. D. Karp. 2004. "A Bayesian Method for Identifying Missing Enzymes in Predicted Metabolic Pathway Databases," *BMC Bioinformatics* **5**, 76.

Hamady, M., and R. Knight. 2009. "Microbial Community Profiling for Human Microbiome Projects: Tools, Techniques, and Challenges," *Genome Research* **19**, 1141–52.

He, S., et al. 2010a. "Metatranscriptomic Array Analysis of *'Candidatus Accumulibacter phosphatis*'-Enriched Enhanced Biological Phosphorus Removal Sludge," *Environmental Microbiology* **5**, 1205–17.

He, Z., et al. 2010b. "GeoChip 3.0 as a High-Throughput Tool for Analyzing Microbial Community Composition, Structure, and Functional Activity," *ISME Journal* **4**, 1167–79.

Lacerda, C. M., and K. F. Reardon. 2009. "Environmental Proteomics: Applications of Proteome Profiling in Environmental Microbiology and Biotechnology," *Briefings in Functional Genomics Proteomics* **8**, 75–87.

Miller, L. D., et al. 2010. "Establishment and Metabolic Analysis of a Model Microbial Community for Understanding Trophic and Electron Accepting Interactions of Subsurface Anaerobic Environments," *BMC Microbiology* **10,** 149.

Ram, R. J., et al. 2005. "Community Proteomics of a Natural Microbial Biofilm," *Science* **308**, 1915–20.

Stolyar, S., et al. 2007. "Metabolic Modeling of a Mutualistic Microbial Community," *Molecular Systems Biology* **3**, 92.

Taffs, R., et al. 2009. "*In Silico* Approaches to Study Mass and Energy Flows in Microbial Consortia: A Syntrophic Case Study*," BMC Systems Biology* **3**, 114.

Thiele, I., and B. O. Palsson. 2010. "A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction," *Nature Protocols* **5**, 93–121.

Turnbaugh, P. J., and J. I. Gordon. 2008. "An Invitation to the Marriage of Metagenomics and Metabolomics," *Cell* **134**, 708–713.

Tyson, G. W., et al. 2004. "Community Structure and Metabolism Through Reconstruction of Microbial Genomes from the Environment," *Nature* **428**, 37–43.

VerBerkmoes, N. C., et al. 2009. "Systems Biology: Functional Analysis of Natural Microbial Consortia Using Community Proteomics," *Nature Reviews Microbiology* **7**, 196–205.

Vila-Costa. M., et al. 2010. "Transcriptomic Analysis of a Marine Bacterial Community Enriched with Dimethylsulfoniopropionate," *ISME Journal* **314**(5799), 652–54.

Wang, F., et al. 2009. "GeoChip-Based Analysis of Metabolic Diversity of Microbial Communities at the Juan de Fuca Ridge Hydrothermal Vent," *Proceedings of the National Academy of Sciences United States of America* **106**, 4840–45.

Warnecke, F., et al. 2007. "Metagenomic and Functional Analysis of Hindgut Microbiota of a Wood-Feeding Higher Termite," *Nature* **450**, 560–65.

Zhao, J., et al. 2010. "Modeling and Sensitivity Analysis of Electron Capacitance for *Geobacter* in Sedimentary Environments," *Journal of Contaminant Hydrology* **112**, 30–44.

C.2 Software Requirements for Metacommunities Scientific Objective 1: *Model Metabolic Processes within Microbial Communities*

Summary of Scientific Objective

The scientific objective is to be able to integrate different types of experimental measurements relating to the metabolic activity of different microbial communities in microbiomes relevant to DOE missions in bioenergy production, environmental remediation, and carbon cycling. The purpose is to (1) generate hypotheses about the nature of interactions among community members and interactions between the community and the local environment; (2) generate hypotheses about the organisms and pathways responsible for community metabolic activities; and (3) be able to predict how the community will respond to environmental changes or introduction of new microorganisms. The ability to understand and compare communities, including those that vary spatially and temporally, will also be essential to building community metabolic models and require tools for comparative community analysis.

Resulting Requirements

IMPACT FACTOR: _X_ HIGH ___ MEDIUM ___ LOW

Process of the Science (Including Workflows)

Characterization of environmental microbial physiology can proceed through two broad approaches, namely, the (1) bottom-up approach, in which the microbes are isolated and cultured in the laboratory and integrated, evaluated, and modeled in a defined community; and (2) top-down metagenome-based approach, in which DNA from environmental samples is directly sequenced for understanding metabolic potential through bioinformatics and pathway reconstruction.

For analysis of defined communities, the first step will involve incorporation of all individual metabolic models into a common environmental model. This will require information on the substrates metabolized and secreted by community members, as well as common nomenclature for exchanged metabolites. In addition, kinetics of substrate uptake and secretion as well as biomass yields will be critical to develop such community models. This first step is a key requirement before more complex communities can be studied.

For the top-down approach, the metagenome sequence after assembly and annotation can proceed through the workflow described below. (See Fig. C.1 for an illustration of this process.)

- **Phylogenetic Binning.** Several methods have been developed, but further validation of these methods is needed. Significant advances could be achieved by combining the best features of multiple binning methods (e.g., use phylogenetic markers where available, then use k-mer frequency-based methods on contigs too short to contain useful marker genes).
- **Pathway Reconstruction from Noisy, Incomplete Data.** Most phylogenetic bins will tend to be incomplete, and the binning itself may be of differing quality. Pathway

reconstruction methods will need to take these sources of noise and uncertainty into account in a systematic manner. Knowledge of reference genome content and pathway content may help guide reconstruction.

- **Metabolic Modeling of Community Members.** Given an inventory of pathways present in each phylogenetic bin, construct a functioning metabolic model for each bin. This problem is identical to metabolic model development for single genomes, except that some of the previous step's probabilistic nature may need to be taken into account. This step will leverage automated methods for constructing draft models for individual bins.
- **Combine Community Metabolic Model.** Integrate component metabolic models into a whole-community model. Simulation frameworks should be capable of integrating models of individual organisms into a community model with the ability to integrate customized workflows for simulation purposes. Ultimately, these models should be able to integrate different types of data (transcriptomic, proteomic, and metabolomic, as well as microbial composition).

Instruments to Support Achievement of the Scientific Objective

- Next-generation sequencing for metagenomes, including single-cell sequencing, 16S libraries, RNA-Seq, proteomics, and metabolomics.
- Functional gene arrays.
- Spatial patterning of cells in communities, FISH.
- New methods to isolate and perform single-cell measurements (transcriptome, proteome) would be beneficial.
- Need access to instruments that enable high-throughput culturing of defined communities in different environments to enable rapid physiological characterization.

User Interfaces

Interfaces are needed for uploading data, integrating and visualizing metabolic pathways, and performing simulations via web interfaces. Also needed are tools to predict, visualize, and compare community response to experimental data and to conduct queries for model and network comparison and queries across metabolic models and pathways such as reachability analysis from one metabolite to another across species boundaries. In addition, developing tools for visualizing simulation and experimental data simultaneously would be valuable, as would methods to flag conflicts among these datasets. Comparisons among community models will include clustering representations (such as trees and PCA plots) and systems representations (such as systems of metabolic maps). Other interfaces are needed to visualize predicted fluxes through metabolic networks and compare them with genome-wide "omics" data.

Programmatic Interfaces

Need the same interfaces required for modeling microbial metabolism in individual organisms, including:

- Access to genome (GenBank, Swiss Prot) and metabolic (KEGG, MetaCyc, and BRENDA) databases.
- Interfaces to automatically download models and model variations.
- Interface to inputs and outputs of binning algorithms to enable the exploration of alternative binning solutions.

Data

Needed data include metagenome sequence data; information on environmental conditions (e.g., available nutrients, extracellular metabolite profiling, carbon, nitrogen, phosphorus, oxygen, pH, temperature, and light); temporal and spatial measurements; transcriptome, proteome, metabolome, and microbial physiology data; and stable isotope probing that allows us to focus on highly active pathways and organisms. On the hardware side, we need to enable parallel computing because some of the simulation methods can be easily parallelized.

Software

Tools should work with missing and noisy data and should integrate data across multiple platforms (genomics, transcriptomics, proteomics, and metabolomics). Developed tools and databases also need to be managed and maintained over time to keep them relevant and functional. In addition, tools relating to solving differential equations with embedded linear programming problems could be needed for dynamic simulations of defined communities. These tools will overlap with microbial group requirements for the science objective.

Metacommunities 1: *Model Metabolic Processes within Microbial Communities*

Table C.1 Software Requirements for Metacommunities 1

Software Purpose	Availability	Improvements Needed?	Resource Impact
Assembly and gene calling and annotation pipeline			
Binning	Few	Yes	Computing
Metabolic pathway inference from noisy and incomplete data	Very few	Yes	Computing
Pathway hole filling	Few	Yes	
Network debugging for simulation	Few	Yes	
Methods for modeling incomplete networks			Computing

Metacommunities 1: *Model Metabolic Processes within Microbial Communities*

Table C.1 Software Requirements for Metacommunities 1

Software Purpose	Availability	Improvements Needed?	Resource Impact
Community diversity analysis (e.g., Mothur)	Some	Probably	
High-resolution functional annotation	Only gene-family membership and nearest homology-based (e.g., BLAST) approaches	Yes. Would prefer gene trees	High-resolution functional annotation
Comparisons of gene content and pathways	Some	Yes, including better statistical methods	Comparisons of gene content and pathways
Linkage between taxonomic profile and metabolic network reconstruction (e.g., distinguishing systems and their components in closely related subtypes. Also filling in gaps from confident core systems in reference genomes.	None (for either need?)		
Build community regulatory networks	No		
Identify interactions between organisms	Limited	Yes	
Incorporation of metadata (e.g., environmental parameters) into comparisons of community models	Limited	Yes	
Selection of communities and isolates for comparison by numerous criteria (e.g., environmental parameters, physiology, taxonomy, gene content)	Limited	Yes. Also need more complete metadata	

Standards

Models and networks need to be developed in an SBML or BioPax format that will allow visualization and simulation in a variety of software platforms.

We also need:

- Standard set of metabolites and reactions (at least the reactions that describe the transport of metabolites across cell boundaries) to effectively integrate individual models of community members.
- Thermodynamic properties of metabolites and reactions (potential overlap with microbial group).
- Metric that identifies the extent of reliance on gap filling to overcome patchy input data.

Since many tools and analyses will be used for comparative analyses, standardized data formats (input and output data) will be crucial for "plug and play" ability.

Governance

While the open-access policy that allows users to incorporate improved modules for analysis should be encouraged, some validation of computational methods should be submitted to ensure that data formats are compatible with existing modules. Methods for ranking the confidence of models, portions of models, and methods used will be important for ascertaining reliability and will be needed to highlight tools that are useful and others that need additional development. This ranking should be a requirement for developers of analysis methods submitted to Kbase.

Maintenance and updating of reference databases will require ongoing stakeholder input.

Summary and Prioritization of Requirements

Tools must account for noisy and incomplete data, with a wide range of community complexity (but usually very diverse and nonuniform), with corresponding issues including fragmentary sequence information (incomplete genes and assembly) and sparsely sampled omics data in general.

- **Short-Term Essential Components.** Essential components include the establishment of methods for modeling a defined community, a necessary first step before complex communities can be modeled. A number of phylogenetic binning methods have been proposed. They need to be more fully validated and integrated to combine the best features of each.
- **Short-Term "Nice to Have" Components.** Standards for data exchange and improved binning and modeling methods will be nice to have. We would also like to identify missing regions of taxonomic space for reference genome data, furthering efforts like GEBA (Genomic Encyclopedia of Bacteria and Archaea) but using environmental sources instead of isolates. So requirements include Kbase making linkages between metagenomic data and reference genome sources as well as tools to determine the

available reference genomes relevant to a particular environment and identify the taxonomic gaps.

- **Mid-Term Essential Components.** Another essential component is a basic genome annotation pipeline that can be integrated with the binning and network reconstruction pipeline. This workflow should be followed by a basic descriptive metabolic model of natural communities. In addition, user interfaces to import sequence data and visualize pathways in the context of multi-omic data are important. New methods that can deal with noisy and incomplete data for network reconstruction will need to be developed.
- **Mid-Term "Nice to Have" Components.** These include improved modeling methods, comparison tools for basic incomplete models at the level of ecotype and higher, and resolution comparison tools for metabolic and systems capabilities.
- **Long-Term Essential Components.** These include the ability to integrate other omic datasets with the metagenome-based metabolic network and to use such data to improve the model and develop the predictive capabilities of models. Also needed is incorporation of other networks including regulatory, signaling, and intercellular interaction networks and integration of the predictive metabolic models with models that incorporate spatial and temporal distribution of metabolic activity. Multispecies interacting metabolic models predictive of response to perturbation are required for the purpose of environmental remediation or other desirable functional behavior.

System Architecture Attributes

Need exportable formats for models that are consistent to allow model exchange

Kbase Key Services

- Methods to predict missing enzymes from sequenced datasets**.**
- Methods and models to predict metabolic behavior of a community.
- Models to serve as a framework to integrate other experimental datasets.

Risk Analysis and Mitigation Strategies

- Single-cell genome sequencing may render obsolete the need for metagenome sequencing and phylogenetic binning because metabolic models can be developed from such sequences, but species interaction modeling and behavior of the model in community context will continue to require community analysis.
- Relevant communities may be too complex to even permit a descriptive model due to the metabolic networks' incomplete nature. Consequences of incomplete DNA sequence datasets may block completion of tasks. Gap filling using other existing datasets may overcome this problem. Gap filling also may be addressed by additional experimentation or by obtaining additional reference genomes.

- Metadata associated with metacommunity (metagenomic) samples will vary widely depending on research groups and data generators, which will complicate efforts to compare samples and datasets collected by the scientific community. For example, studying the resiliency of soil communities to perturbations will require availability of metadata on perturbation conditions.

C.3 Metacommunities Scientific Objective 2: *Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function*

Metagenomic projects have the potential to dramatically expand the number of genes known to scientists, overpowering current efforts aimed at determining the function of genes. At the same time, data generated in large-scale metagenomics projects can provide the information necessary to better understand the function of poorly characterized genes. As metagenomic data are rapidly coming online, a critical scientific objective is to develop approaches for (1) mining the data to identify previously unknown genes and (2) leveraging the wealth of metadata associated with metagenomic datasets, as well as gene and organism co-occurrence information, to identify testable hypotheses about the function of newly identified or poorly characterized genes.

Background

Roughly one-third of all genes in the *E. coli* genome have no known function (Hu et al. 2009), despite the fact that this bacterium is among the best-studied organisms. Although scientists are slowly elucidating the function of some of these genes (Weber et al. 2010), their efforts cannot keep up with the wealth of data being generated in both traditional genomics projects and through large-scale metagenomics efforts. The magnitude of the problem is perhaps best exemplified by the number of novel protein sequences identified by the Global Ocean Sampling expedition (Yooseph et al. 2007). The authors of this study identified over 1700 genes with no similarity to any known protein families. Efforts to understand the function of these genes cannot effectively be conducted without first prioritizing the genes on the basis of their importance to pressing biological questions. But how can we know which genes are important if we do not even know what they do? The key to this chicken-and-egg problem lies in the metagenomic datasets themselves. Specifically, metagenomic data are not comprised simply of DNA sequences but also contain a rich set of metadata, information linking the sequences to location (e.g., latitude and longitude, height, or depth), to physical characteristics of the environment (e.g., temperature, pH, and salinity), and to time. Also, information is available that links together multiple metagenomic datasets (e.g., multiple data generated from the same location at different points in time). Prioritization of experimental and annotation efforts as well as possible hypotheses about the function of a gene or group of genes can be derived from available metadata. For example, a particular gene might be found only in samples derived from communities known to perform a particular biological process (e.g., a gene or group of genes found only in oil-contaminated water, implying their possible role in hydrocarbon metabolism). In addition, some genes might be found only in conjunction with genes whose function is known, thereby implying their involvement in similar biological processes.

Prioritization

PRIORITY: _X__ HIGH ___ MEDIUM ___ LOW

Potential Benefits

The development of methods for extracting information about gene function from metagenomic datasets and associated metadata can have far-reaching impacts on biological research in general and on the DOE mission in particular. The resulting data will provide actionable hypotheses about the function of many genes that have yet to be studied in detail. Also, scientific efforts associated with this objective will lead to the discovery of new genes that perform useful biological functions of relevance to DOE priority areas such as energy production and environmental remediation. In addition, these efforts could lead to the development of sensitive markers of ecosystem health.

Feasibility of Success: Near, Mid, and Long Term

TERM: _____ NEAR (1-3 years) __X_ MID (3-5 years) ___ LONG (5-10 years)

Relevance to the Kbase Project

The availability of reliable functional annotations is a critical prerequisite of a successful research program in systems biology. This objective will potentially accelerate efforts aimed at characterizing the function of currently understudied genes. In addition, tools developed as part of this project will be valuable assets to scientists generating new datasets by allowing them to leverage Kbase-associated datasets in the analysis process and to generate actionable hypotheses.

Synergies and Leverages: Potential Overlap with Other Projects or Funding Agencies

Similar efforts will probably be undertaken in other research fields that are starting to apply metagenomic methods (environmental, agricultural, and health research). Potential overlap thus exists with research funded by a broad range of agencies (National Institutes of Health, National Science Foundation, National Aeronautics and Space Administration, U.S. Department of Agriculture, U.S. Food and Drug Administration). It will be necessary to maintain regular communication between DOE and these agencies and to actively and broadly disseminate the results of work performed by the Kbase effort.

Specificity

This scientific objective can be broadly broken up into two subobjectives: (1) novel gene discovery and identification and (2) functional prediction.

- **Novel Gene Discovery and Identification.** Starting with one or more metagenomic datasets, we need to reliably identify genes (metagenomic assembly and gene finding are important prerequisites to this effort) and ensure they can be tracked across datasets and databases. More precisely, if a gene G is found dataset A, we need to ensure that it is "real" (not a spurious call or possible contaminant), identify all other datasets that contain homologues of G (note: homology detection itself is a challenging

question), and determine whether homologues of G appear in public databases with or without a specific annotation.

- **Functional Prediction (or Hypothesis Generation)**. Starting with a matrix containing genes or gene clusters as rows and metagenomic datasets as columns and metadata associated with or linking together these datasets, we need to develop methods for identifying genes that correlate with environmental features of interest. At the most basic level, we could perform enrichment analyses to identify genes or groups of genes significantly associated with specific environmental characteristics (e.g., genes significantly associated with datasets from oil-contaminated water). Methods for performing such analyses already exist; however, managing the above-mentioned matrix and applying these methods in the context of rich metadata are logistical nightmares. Kbase could provide the tools and infrastructure necessary for "playing" with the data as seamlessly as possible.

In the longer term, more complex analyses could be applied such as using various differential equation models to analyze longitudinal data to understand mechanistic interactions between genes, genes and organisms, and genes and environmental parameters.

Details

Scientific Discovery Process (Workflows)

- Metagenomic sequences are assembled.
- Genes are found within the assembled contigs.
- Genes are compared to other datasets registered within Kbase and to public databases, homologies are detected, and appropriate identifiers are assigned that enable tracking the same gene across datasets.
- A data matrix is constructed from user-selected (or automatically suggested) datasets.
- Statistical computations are performed on the data matrix based on user-defined criteria and column permutations (e.g., "interesting" columns are selected on the basis of a combination of metadata, and genes significantly enriched or depleted in these columns are identified using statistical software).
- The resulting data can feed into new hypotheses or predictive models of gene interactions.

Inputs

- DNA sequences.
- Data matrix as described above in Functional Prediction (or Hypothesis Generation) in the Specificity section.

Outputs

- "Universal" gene identifiers linking together homologues of a same gene across multiple datasets even if the gene is not annotated in any public database.
- Graph of connections between genes, genes and organisms, and genes and environmental parameters annotated with strength of connection and statistical significance.

Tools

- Gene "naming" tool.
- Homology detection tool.
- Tools for detection of statistical correlations within data matrix.

References

Hu, P., et al. 2009. "Global Functional Atlas of *Escherichia coli* Encompassing Previously Uncharacterized Proteins," *PLoS Biology* **7**(4), e1000096, doi:10.1371/journal.pbio.1000096.

Weber, M. M., et al. 2010. "A Previously Uncharacterized Gene, *yjfO* (*bsmA*), Influences *Escherichia coli* Biofilm Formation and Stress Response," *Microbiology* **156**, 139–147.

Yooseph, S., et al. 2007. "The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families," *PLoS Biology* **5**(3), e16.

C.4 Software Requirements for Metacommunities Scientific Objective 2:

Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function

Summary of Scientific Objective

To improve the characterization of microbial environments relevant to DOE missions in carbon sequestration, environmental remediation, and biofuels, the appropriate models must be constructed. To produce such comprehensive models, an extensive collection of linked data is needed allowing for the interrogation of genes, where they are found, and what the metadata are for each of the environments (e.g., co-occurrence of data, taxa, functions, and genes). To accomplish this, we need to create a digital ecosystem that allows researchers to collaboratively manage, explore, and interrogate the deluge of community-based sequence data.

Resulting Requirements

Standards Development Requirements. As has been demonstrated in many large multi-institutional projects, development of standards requires significant investment of resources. Kbase resources should be specifically designated to promote adherence, development, and utilization of standards to describe experimental, expression, genomic, and geophysical data.

These activities should be coordinated with existing international efforts as well as projects derived from other genome-related funding agencies.

Process of the Science (Including Workflows)

To begin creating such a digital ecosystem, a coordinated set of core tools (i.e., modules) will be needed for a community of researchers to generate a collection of community resources that will allow researchers to independently and collaboratively investigate metacommunities:

- Generate sequence (or import legacy data with tool for ingestion).
- Assemble sequence.
- Automate annotation of sequence.
- Generate a database (contribute to existing generated databases).
    - The generated database should allow for community contribution and interaction.
    - Tools will be associated with the database to allow for interaction (e.g., genome browsing).
- Provide a facility that will allow for the annotations and associated data to be "spidered" and made discoverable to the larger community.
- Provide the means to link and incorporate other available data (e.g., proteomic and transcriptomic data).

Instruments to Support Achievement of the Scientific Objective

Instruments related to data collection (samples, sequence, and metadata) will be involved.

User Interfaces

Users targeted by this requirement description are the bench scientists. That is, we are enabling a diverse and growing community of researchers to analyze much larger collections of data. User interfaces will fall into five categories:

- Domain-specific tools for specific analyses (e.g., assembly).
- Collection of tools to easily generate and contribute to extensible data resources.
- Interfaces that allow researchers to interrogate the data in these resources.
- Suite of tools to allow researchers to navigate through appropriate modules (allowing for selection of alternative modules) and subsequent workflows of analysis and visualization modules.
- Search-and-discovery interface allowing researchers to explore the broader collection of linked community data.

Programmatic Interfaces

To enable the pipelining of tools and modules, creation and submission to the database, appropriate data access, and data submission, application programming interfaces (APIs) need to be developed based on the standards described below. In addition, specific linked data APIs will be required to enable interlinking of resources.

These APIs should be made available as RESTful web services that can be easily incorporated into diverse environments.

Data

Data required to meet the scientific objective is standard. However, data format (to enable this modular process and the creation of linked data resources) and needed standards will be critical. Rich data formats will be needed for data interchange to ease plumbing between tools. These would not be minimal, as they would be required to create persistent digital objects with semantic consistency.

- Need for simple data exchange standards.
- Simple ways to describe and link data.
- If data have been used for search, should be able to return the search results.

Through the development of a collection of file formats to which any researcher could write, researchers would be able to upload files to their sites and allow appropriate indexing and discovery via Kbase spiders that can read these formats. This will enable a network model for the creation and publication of data (i.e., large-scale repositories will fail over time).

The system will also need access to up-to-date data resources that are utilized by various tools. For example, the assembly module will require an annotation clearinghouse.

Software

***Metacommunities 2**: Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function*

Table C.2 Software Requirements for Metacommunities 2

Software Purpose	Availability	Improvements Needed?	Resource Impact
Data conversion	Yes	Yes (standard interoperability)	
Data submission tools	Yes	Limited number of tools that work with comprehensive standards	
Generate sequence data	Yes		
Assembly	Yes		

Metacommunities 2: *Mine Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function*

Table C.2 Software Requirements for Metacommunities 2

Software Purpose	Availability	Improvements Needed?	Resource Impact
Automated annotation	Yes	Improvements of annotation via CAFAE	
Generate collaborative database from data	No		High
Genome browser	Yes		
Data crawling and indexing infrastructure	No		High
Workflow management system for guiding researcher through independent and configurable modules	Yes	Many such systems exist, however, such a system needs to be adapted for Kbase	
Allow linkage to other community data	No		High
Social and community tools for interaction with the published data	No		High

Standards

To be able to provide an integrated system, as described above, a core requirement is better definition and incorporation of metadata related to data and processing of data. To accomplish this, a coordinated set of standards needs to be implemented so that:

- The infrastructure can handle diverse types of metadata.
- The standards that exist are extensible.
- A governance structure ensures that people comply with standards.

A wide range of community-level standardization activities are now ongoing in the scientific community. Largely, these activities are done on a volunteer basis and their outputs include a range of checklists, ontologies (vocabularies), file formats, and tools to enable data sharing. These activities have included standards associated with complete genomes, environmental samples, metagenomes, microarrays, proteomics, and phylogeny. Typically, these activities include formation of organized consortia and societies to foster improved cross-community

interactions, with the aim of finally achieving robust data sharing and minimizing the expense of reconfiguring data described in incompatible formats.

The recent publication on "omics data sharing" (Field et al. 2009) reviews a range of data policies from major funding agencies and the kinds of content that good data policies should share. This includes active support for relevant standards and the databases that implement them. To accompany the paper, the authors have created a biosharing website with a listing of many genomic standards.

This website is dedicated to centralizing and giving a higher profile to bioscience data policies and standards. Its offers a focal point for the various stakeholders in data policy by fulfilling two main roles: (1) providing a "one-stop-shop" for those seeking data policy documents and related information (including information about the standards and technologies that support them and (2) encouraging the exchange of ideas and policy components among funders and between funders and potential fundees, ultimately to harmonize policy components where feasible.

A related effort in the domain of genomics and metagenomics is the Genomic Standards Consortium (GSC, gensc.org). GSC is an initiative working toward richer descriptions of our collection of genomes and metagenomes. Established in September 2005, this international community includes representatives from a range of research institutions and major sequencing and bioinformatics centers, including the National Center for Biotechnology Information, European Molecular Biology Laboratory (EMBL), DNA Data Bank of Japan (DDBJ), J. Craig Venter Institute, DOE Joint Genome Institute, European Bioinformatics Institute, Sanger, and the Fellowship for the Interpretation of Genomes. The GSC goal is to promote mechanisms for standardizing the description of (meta)genomes, including the exchange and integration of (meta)genomic data. The number and pace of genomic and metagenomic sequencing projects will only increase as the use of ultrahigh-throughput methods becomes commonplace and standards are vital to scientific progress and data sharing.

The International Nucleotide Sequence Database Collaboration (INSDC)—which includes GenBank, European Nucleotide Archive (including EMBL-Bank) and the DDBJ—partners have recognized the Minimum Information about a (Meta)Genome Sequence (MIGS/MIMS) and the Minimum Information about an Environmental Sequence (MIENS) family of minimum standards. They have recently reserved an official keyword for compliant INSDC sequence records and are working to introduce support for submission of compliant datasets.

Reference

Field, D., et al. 2009. "'Omics Data Sharing," *Science* **326**(5950), 234–36.

APPENDIX D

Individual Reports from the 2009–2010 DOE Systems Biology Knowledgebase Workshops

DOE Workshop on Cloud Computing in Systems and Computational Biology: Workshop Report

November 16, 2009
Portland, Oregon

Convened by

The U.S. Department of Energy Office of Science
Office of Biological and Environmental Research

Organizers: Folker Meyer (Argonne National Laboratory), Susan Gregurick (U.S. Department of Energy), Peg Folta (Lawrence Livermore National Laboratory), Bob Cottingham (Oak Ridge National Laboratory), and Elizabeth Glass (Argonne National Laboratory)

Audience: 130+ people

Speakers: Folker Meyer (Argonne National Laboratory), Dawn Field (Oxford, UK), Eugene Kolker (Seattle's Children Hospital), David Haussler (University of California, Santa Cruz), Simon Twigger (Medical College of Wisconsin), Ananth Kalyanaraman (Washington State University), Michael Schatz (University of Maryland), Sam Angiuoli (University of Maryland), Narayan Desai (Argonne National Laboratory), Lavanya Ramakrishnan (National Energy Research Scientific Computing Center), Kate Keahey (Argonne National Laboratory), Bob Grossman (University of Illinois at Chicago), Judy Qiu, (Indiana University), Thomas Brettin (Oak Ridge National Laboratory), Owen White (University of Maryland), and Deepak Singh (Amazon)

Panelists: Susan Gregurick (DOE), Owen White (University of Maryland), Pete Beckman (Argonne National Laboratory), Kathy Yelick (Lawrence Berkeley National Laboratory), Dawn Field (Oxford Centre for Ecology and Hydrology), David Haussler (University of California, Santa Cruz), Jeff Grethe (University of California, San Diego), Folker Meyer (Argonne National Laboratory), Victor Markowitz (DOE Joint Genome Institute), Eugene Kolker (Seattle Children's Hospital), Bob Cottingham (Oak Ridge National Laboratory)

Introduction

According to a recent National Institute of Standards and Technology document, cloud computing is a model for enabling convenient, on-demand, network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction ("The NIST Definition of Cloud Computing," v15). This new approach is an evolving paradigm for providing and using computational services. Previously, scientists could either build local computing resources, where all aspects of the system hardware and software could be tuned to their application, or they could adapt their application to pre-existing

computing resources. There was no middle ground between these two options. Now, with cloud computing, through the use of virtual machine images (VMs), scientists can carry a full computing environment with their application to new systems, in a simple and portable fashion. A cloud platform provisions both the hardware platforms and a means to install and run a given environment stored in a virtual machine (VM).

Unlike Grid or high performance computing, cloud architectures can have little to no centralized infrastructure and up to now have mostly been characterized as a third party computing service which is rendered under a utility model or a 'pay as you compute' model. The virtue of this model is that the end user does not necessarily have to invest in computational hardware, software, or administration to enable their particular application or science. However, the drawbacks include latency in data transferred to and between a cloud system, potential data security issues, and system resilience. Moreover, users need to learn how to best make use of cloud interfaces and capabilities.

Nevertheless, the 'compute as you go' properties that are inherent to cloud computing make this an interesting platform that may be well suited for the computational needs of the systems biology research community. While other communities have shared data, the large volume of data shared within those communities typically comes from a small number of data sources. In the 'Omics' disciplines of systems biology such as genomics, transcriptomics, proteomics, etc., data generated using a variety of instruments (DNA sequencers, mass spectrometers, micro array readers, etc.) are shared with the community at large and are used as the basis for a variety of integrated research projects, frequently involving large computations. However, the sharing process is not yet very effective, and as more data generators and data consumers enter this arena, more efficient ways of data sharing will be required. The scale of the deployments of next-generation instruments in the biology disciplines mirrors the number of research laboratories working in the Omics disciplines. The result is a striking democratization of Omics data generation and the subsequent need for collaborative research. As a consequence, many high volume data sources exist in biology and with the data volume rising exponentially, a platform that provides low-cost, flexible computational services, like that of the cloud, could be a good match for the needs of the systems biology community.

This community is characterized by its growing creation and consumption of data as well as the inherent rise in computational demands. Two prominent features attracting solution providers in bioinformatics, computational biology, and systems biology to "the cloud" are the "seemingly endless supply of cycles" (P. Beckman) and the ability to "bring your own environment" (O. White).

Cloud Computing for the DOE Systems Biology Knowledgebase

The 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 (genomicscience.energy.gov/compbio/). 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.

A cloud-based computational platform presents a promising opportunity for the DOE Systems Biology Knowledgebase (Kbase). The objective of this workshop was to elicit community input on the feasibility of using the cloud paradigm as a component of Kbase. Specifically, the workshop participants evaluated the requirements for a cloud-enabled Kbase and presented ideas for engaging the High Performance Computing (HPC) community in this effort.

The three charges for this workshop were:

1. What are the characteristics of applications that would be appropriate for effective utilization of cloud architecture?
2. What are the hardware bottlenecks that prohibit cloud architectures from being easily adopted by high-throughput biological data analytics?
3. What are specific tools that need to be developed or enhanced in order to make cloud architectures easily adopted for biological data and bioinformatics algorithms?

This workshop brought together more than 130 computer scientists, bioinformaticists, and computational biologists to discuss the feasibility of using cloud computing for DOE's systems biology Kbase. This workshop was held in conjunction with Supercomputing '09 in Portland Oregon. The one-day workshop consisted of leaders in these fields presenting work in cloud computing for biological research as well as a two-hour round table discussion centered on the charge questions**.**

The recommendations from this workshop are summarized below:

1. Switching to clouds will require re-engineering, at first for scalability and then for fault tolerance.
2. A healthy research community requires open source reference implementations of algorithms and pipelines. While many algorithms are open source, the pipelines tend to be closed source due to tight integration with locally existing environments. Clouds provide a unique opportunity to open up existing pipelines to more scientific scrutiny.
3. There is a need for standards in computational workflows and data, including the sharing of intermediate computational results.
4. There is no consensus for the software appliance operational model in bioinformatics (as demonstrated by an informal vote on preferences to "bring your own VM" vs. "use a provided VM").
5. Wide area data movement is still an active issue in the bioinformatics community; these problems are amplified when dealing with clouds.

6. There is a need for a system enabling the flow of data across different VMs instantiated in a cloud, potentially for VMs from different appliance providers.

Clients for the Kbase Computing Platform

We discussed three types of potential users who would benefit from the Kbase computing platform: computational biologists, bench biologists, and integrators who aim to integrate smaller components into advanced workflows.

From an architectural perspective it is interesting to consider for whom the Kbase computing platform should be tailored. While any platform can support multiple types of users, having an understanding of which users one intends to support the most, or with highest priority, will influence the ultimate design of the hardware platform. We anticipate computational biologists and bioinformatics users being advanced users insofar as they would require more access and more control of their resources hosted by the computing platform.

Bench biologists, on the other hand, would be users of a wider variety of readily available third-party software applications, and therefore the ability to support key third-party applications would influence the architecture.

Need for Reference Data and Standards

Standards are a mechanism for capturing information in a form easily shared and integrated with other data or data types. Using data standards to capture data entities is the foundation for comparative analysis and integration of information.

Increases in biological data production are dramatic. For example, DNA sequencing throughput has grown from 100-500 megabytes per run to 30-90 gigabytes per run in a 12-18 month time frame. As an increasing number of diverse data products (e.g., genome annotation, protein families, pathways, etc.) are transformed and consumed by many different parties, there is a strong need for versioning and controlling production of reference datasets and provenance information. The choice for data transformation also is important for downstream processing (e.g., missing genes impact metabolic pathway predictions, false gene starts impact protein family curation). There is limited sharing of results—especially for compute intensive intermediate results—because data formats have been designed for different purposes. Since using a few centralized, closed pipelines can not meet the communities' dynamic data and information needs, rapidly evolving data exchange standards and data provenance descriptions will be required.

Use of Workflows

The concept of "abstract workflows" seems to be the right level of abstraction for most bioinformatics and computational biology groups. Most HPC providers are focused on parallel implementation of workflows.

In bioinformatics, large-scale computations are often encoded in scripts that wrap and automate the use of standard tools and methods, typically making the pipelines opaque and hard to port between computational platforms. The international Genome Standards

Consortium (GSC) has promoted rich, transparent descriptions of analysis pipelines, such as Standard Operating Procedures (SOPs), which can be published in academic journals and help promote best practices. While these SOPs are fine for describing a process to colleagues, they are not readily interpretable by a computer and are often not rich enough to ensure that results are reproducible at another institution. As a complement to human-readable SOPs, the GSC and its M5 metagenomics working group seek a workflow description language for describing bioinformatics pipelines that can be executed on large computational resources. The format, such as an XML, should be exchangeable between institutes, platform independent, and compatible with existing workflow systems. Of particular interest is portability across grid and cloud architectures, as groups interested in running workflows may have access to a variety of resources. An ideal format would be relatively high level and would describe the process in terms of standard tools or algorithms, such as BLAST. An ideal format would also hide certain complexity required to improve performance, such as partitioning data into batches for batch processing, as these steps may significantly impact the process flow but not the analysis algorithm (such as an readily parallelizable search like BLAST). Also, ideally, such a format already exists within the workflow community and can be adopted and promoted for use by the GSC and M5, although it is not immediately clear which, if any, existing workflow descriptions fulfill the goals outlined here. We also highlight that in cloud computing environments, there are unique opportunities to integrate analysis pipelines and data as part of a single shared resource, the "cloud."

Reference Virtual Machine Images

First introduced in the 1960s on IBM's 370 platform, virtual machines (VMs) enable running multiple operating systems on the same hardware platform using a hardware abstraction layer (nowadays referred to as "hypervisor"). Using VM technology, the process of loading an operating system environment becomes similar to loading an application in today's desktop environments. The VM-image is a single file not dissimilar to existing application programs (e.g., GNU Emacs).

The DOE FastOS program (run out of the DOE Office of Advanced Scientific Computing Research) has clearly demonstrated the value of OS customization for application performance. Clouds (VM instances in particular) provide an unprecedented opportunity to tailor runtime systems to bioinformatics applications. Thus, it seems clear that the ability to run arbitrary VMs is going to be a key component of the Kbase systems architecture.

The creation and maintenance of a set of reference VMs can add significant value, enabling a large number of bioinformatics scientists. A "VM marketplace" with open, maintained reference images and the ability for individual developers to add functions and/or programs to those images will be an important component.

The team providing the computational platform (cloud) should be charged with providing a set of reference VMs and maintaining them over time. Outside developers should be enabled to generate "appliances" using the VMs provided (or using user-provided VMs). The use of a configuration system (e.g., BCFG) should be encouraged to separate images (provided by the resource provider) and the semantics of value provided by third parties.

The provision of reference VMs as an open-maintenance model and a configuration management tool to enhance the VMs was addressed in a discussion about infrastructure as a service (IAAS) vs. software as a service (SAAS). It seems clear that a core team should provide infrastructure as a service enabling a distributed team to provide software services.

Evaluation of Use Paradigms for Clouds

Despite the availability of cloud computing IAAS and SAAS offerings over the last few years, adoption in the computational science community in general and bioinformatics community in particular has been slow. Recently, groups (several of whom presented at the workshop) have begun to evaluate and use cloud resources as a part of their computational platform. Cloud resources provide a set of capabilities and operational properties that are distinct from traditional computational resources, both in terms of computation as well as data storage. Effective use of these resources will likely require explicit adaptation of applications, use policies, and operational practices in order to accommodate these differences.

The key challenges posed by cloud resources are scalability, fault tolerance, lack of locality, and pricing. One of the primary promises of clouds is elastic, on demand scaling. In order to take advantage of this capability, pipelines need to eliminate scalability bottlenecks and support the dynamic addition and release of resources. With the addition of scalable computational resources, fault tolerance quickly becomes an issue. Pipelines must be able to cope with frequent resource failures in a transparent and robust fashion.

Bioinformatics applications tend to be quite I/O intensive. This will be a key challenge with cloud systems, where locality and network topology are not well exposed. Traditional network file systems are not well positioned to solve this problem, as they are usually deployed in fixed configurations and cannot easily be migrated to follow compute resources as they move around inside the cloud. Data aware programming models (like MapReduce/Hadoop) are capable of solving this problem, but require substantial reworking of analysis pipelines.

The final challenge is pricing. Commercial clouds provide compelling services, but are not optimized for computationally bound workloads. Cloud pricing models include raw node hours consumed, as well as data transfer, storage, and use forecasting. These must be taken into consideration when deciding which tasks to run on commercial clouds instead of local resources. This issue may also drive adoption of cloud approaches (such as Eucalyptus, Nimbus, or Hadoop) on local resources.

Several approaches to adapt pipelines to cloud resources were presented at the workshop. Each of these demonstrated pros and cons of particular approaches, and none is likely to represent eventual production cloud architectures. Several groups demonstrated architectures that extended the local computing infrastructure directly into clouds without modification. This has the benefit of being relatively straightforward to implement, but suffers from a number of potential security issues, when exposing local infrastructure to cloud resources. The lack of clear locality information makes this approach suboptimal for highly I/O intensive analysis techniques. Another approach was to build a work management system (AWE) that has knowledge of the semantics of work units. This allows the re-use of intermediate results (that have already been computed), as well as the optimization of task placement based on data

requirements. This approach has been more labor intensive than the previous approach, however, it should provide better scalability, security, and a better fault tolerance model. A third approach was to completely adapt applications to the Hadoop/MapReduce programming model. This programming model has explicit support for data parallel operations, hence it supports very aggressive data locality optimizations. This approach is the most labor intensive of the three discussed here, however, it supports the most aggressive data locality optimizations of the three.

Applications Suitable for Cloud Technology

Applications Suitable for Porting to "the Cloud"

As was demonstrated by a number of presentations during the workshop, not all computations are equally well suited for running on a distributed cloud platform. A cloud-based approach is only one of many conceivable technical choices. For specific tasks, using a large shared-memory machine might be more appropriate. In some instances, a local cluster might offer benefits over a cloud machine (e.g., when the amount of computation is small, and the amount of data is large, as is the case for the image analysis step of DNA sequencing pipelines).

Currently there is no fixed, well understood model for when an existing application is "a good fit" for a cloud-based solution. The weakly defined nature of "cloud computing" is an important contributing factor. Assuming that "cloud" is synonymous with "distributed VM image-based computing," the following factors seem to play a role in determining whether a cloud platform is a good target for a given application:

- What are the communication patterns of the application (client-server vs. intense client-client communication)? And in this context another important consideration is available network bandwidth.
- Does the application rely on a central/global high-performance file system with specific semantics or performance?

Example Applications

Various groups in bioinformatics have already gathered experiences with the use of cloud (or cloud like) platforms. Here we highlight some of the use-cases presented during the workshop.

While the scales of their computational requirements are different, several groups found aspects of cloud computing to be enabling for their data analysis needs. A group from the Medical College of Wisconsin used Amazon's EC2 product to make an existing internal pipeline available to more users (overcoming internal resource limitations that were throttling use of an instrument pipeline). This can be taken as a typical example where relatively small groups are enabled to provide their pipeline to outsiders, without the need for the group to invest in local hardware and/or charging users for computational services. Both the amount of data and the computational resources required for this type of approach are modest.

Examples two and three both showcase metagenomics applications that consume large quantities of resources with "larger" gigabyte sized datasets. Both examples are pipelines that are in constant use and are currently resource limited.

Example four from the University of Maryland is showcasing how the novel computational metaphors coming available within the cloud context alter application development.

Example 1: A Cloud-Enabled Proteomics Workflow at Medical College of Wisconsin. Modern mass spectrometers are capable of generating data many times faster than a typical single desktop computer is able to analyze it. We have brought together two recent developments, open source proteomics search programs and distributed on-demand or "cloud" computing, to allow for the construction of a highly flexible, scalable, and very low cost solution to proteomics data analysis: the Virtual Proteomics Data Analysis Cluster (ViPDAC). On boot, the application sets up the databases, links launch scripts, executes worker daemons, and starts monitoring the running processes. Access to the application is via a web browser to a server name provided by EC2 on startup. Users create a new search job and upload their datafile, which is split into independent chunks that are stored on S3 and distributed to waiting worker nodes. Each worker searches the datafile against a database specified in the job, storing the search results back on S3. When the job is complete, the head node downloads and assembles the result files into an archive suitable for use with other analysis tools.

Example 2: Argonne's MG-RAST Server. Metagenomics applications were among the first to explore the use of cloud computing. These large resource consumers are traditionally implemented as distributed applications, requiring a complex software stack and a central file system. They are also very similar to many of the existing genome analysis pipelines.

Argonne National Laboratory's metagenomics RAST server (MG-RAST) is one example for a recent development in that type of application. More than 120 gigabases of DNA have been analyzed via MG-RAST using a local cluster, TeraGrid, and cloud like resources. While the integration of TeraGrid happened by manually moving datasets and computations to TeraGrid, the integration of cloud resources was facilitated by using a novel workflow system: AWE. AWE (Argonne Workflow Engine) was initially used to run the similarity computation step of the pipeline on a variety of cloud-like resources.

AWE relies on a set of appliances that connect to a scalable fault tolerant server infrastructure for coordination. Both client and servers are lightweight and highly scalable. The server assigns work to clients based on the current workload and client capabilities. Work units are typically a small fraction of the full similarity comparison. AWE understands the structure and semantics of the work that is to be done, and hence can reuse intermediate results as well as scale the size of the work units depending on the speed and capabilities of the client execution environment. Similarly, AWE can use work unit data requirements to route work to locations where needed data is already present. Finally, AWE uses a lease mechanism in work assignment that allows automatic detection and re-routing of failure work units.

AWE provides a lightweight mechanism for distributing work across heterogenous resources, including HPC clusters, clouds, Blue Gene systems, and systems with accelerators (GPUs or FPGAs). Effectively harnessing these resources is a key challenge in order to maximize the analysis progress we can make.

Example 3: JGI's IMG/M. The DOE Joint Genome Institute (JGI) is one of the major sources of microbial genome and metagenome sequence data, currently conducting about 21% of the

reported bacterial genome projects worldwide. Genome and metagenome sequence datasets from JGI and other centers are processed using annotation pipelines and then included in the Integrated Microbial Genome (IMG) system and its metagenome counterpart, IMG/M for comparative analysis. The JGI annotation pipelines support the annotation of genome and metagenome datasets sequenced using any kind of sequencing technology (Sanger, 454 GS0 – 454 Titanium, Illumina). Thus, in the past two years, the metagenome pipeline has processed more than 330 datasets, with variable sizes and distribution of sequence length. This pipeline employs a cluster of 280 CPUs with a processing rate of 24 hours for an average 454Titanium dataset of approximately .5 M reads. For datasets generated by new sequencing platforms (e.g., Illumina) the processing time is estimated to increase several fold with the current computing infrastructure. The integration of such datasets into IMG and IMG/M is also expected to require substantially larger computing capabilities. In order to determine the best solution for meeting this increasing demand for computing resources, a collaboration of researchers from the JGI's Genome Biology Program, Lawrence Berkeley National Laboratory's Biological Data Management and Technology Center, Advanced Computing for Science Department, and the National Energy Research Scientific Computing Center (NERSC) explored the performance and scalability of BLAST on a variety of platforms including a traditional HPC Platform (NERSC's Cray XT4 "Franklin" system), a commercial "Infrastructure as a Service" Cloud (Amazon's EC2), and a shared research "Platform as a Service" Cloud (Yahoo's M45).

The pricing model for cloud services rewards long-term subscription to resources, as shown by the JGI/NERSC group and Wilkening et al. (IEEE Cluster 2009).This aspect of the pricing model limits the ability to dynamically scale their computation in resource limited environments. Additional overhead costs of scientific computing on a commercial cloud include boot up time, data transfer, and loading time.

Hadoop provides an alternative programming model for data intensive computing. It has a number of interesting capabilities including moving computation to the data in distributed environments as well as fault tolerance capabilities not present in traditional HPC systems. While Hadoop can be used with pre-existing applications, replacing workflow systems, its real potential is in new fault tolerant, scalable bioinformatics algorithms. An example of this is shown in the next section.

Example 4: Using Hadoop for Genome Assembly. Michael Schatz from the University of Maryland presented Crossbow, a novel Hadoop-enabled pipeline for quick and accurate analysis of resequencing data for large eukaryotic genomes using clouds.

It combines one of the fastest sequence alignment algorithms, Bowtie, with a very accurate genotyping algorithm, SoapSNP, within Hadoop to distribute and accelerate the computation. The pipeline can accurately analyze an entire genome in one day on a 10-node local cluster or in about three hours for less than $100 using a 40-node, 320-core cluster rented from Amazon's EC2 cloud computing service.

In addition, Schatz presented a new assembly program Contrail (contrail-bio.sf.net), which uses Hadoop for de novo assembly of large genomes from short sequencing reads. Contrail relies on the graph-theoretic framework of deBruijn graphs, similar to other leading short read assemblers (Velvet, Euler-USR, and ABySS). Preliminary results show Contrail's contigs are

similar to those generated by other leading assemblers when applied to small bacterial genomes, but provides superior scaling capabilities when applied to large genomes.

Architecting the Cloud Machine

A number of preferences were clearly visible in the audience comments and in the presentations during the workshop regarding the layout of a possible cloud machine to support Kbase work.

VMs. A cloud machine for the Kbase should support both predefined virtual machine images (VMs) and the capability to run user-provided VMs. This feature will enable easy porting of existing software environments adapted to pre-existing local installations. In addition, however, a number of groups will require help with creating the runtime environments required for their work and will benefit greatly from a set of predefined images. The model suggested by Kate Keahey (Argonne/University of Chicago) was to provide a VM marketplace for the machine, with user-provided (non-supported) and supported VMs.

File System. Since a lot of the computation in genomics will be data driven, a fast parallel file system providing storage for all active datasets is another requirement that was implicit in many of the presentations and became abundantly clear in the discussions during the breaks in the workshop. Because of the large amount of existing code and data, this file system needs to support the typical Linux file system semantics. Hadoop Distributed File System (HDFS) is not an option because it does not integrate with existing code and binaries used in bioinformatics and genomics.

Nodes. As application requirements are vastly different between among machine types, the cloud machine should support a variety of node types, similar to EC2. A user performing BLAST analysis will need many (16) cores with a modest amount of memory (8-16 GB), whereas a single-core sequence assembly program (velvet) benefits from maximizing the available memory (e.g., 256 GB RAM).

Node Interconnect. While many HPC machines use MPI and rely on fast internal interconnects, the vast majority of bioinformatics applications do not benefit from fast interconnects. Instead, most communication is between a node and one or more data servers, not between nodes. For this reason a fast parallel file system will add more value to a Kbase cloud machine than any fast interconnect.

Support Model. An operational model like that for EC2 is well suited for a Kbase cloud machine, as its primary technical customers will be bioinformatics groups providing solutions.

Governance Model for Kbase

Governance. The Kbase cloud infrastructure must include management and oversight as appropriate to serve the needs of the scientific community. The Kbase administrators will be expected to be aware of the user communities accessing the system. They also will continuously address issues of user satisfaction, engage in frequent use-case development, and provide resource allocation mechanisms. Several methods should be considered to ensure appropriate utilization and governance of the system.

Usage Advisory Committee. A usage advisory committee should be established to ensure fair use of the Kbase cloud system. The committee should comprise Kbase IT staff, senior Kbase personnel, and possibly experienced service providers from the external community. The advisory committee should be prepared to meet on a rapid ad hoc basis to resolve contention issues and to review the scientific merit of projects creating a high demand for the system. Other committee issues may include verification of the user's credentials, appropriateness of applications run on the system, general availability of cloud resources, and rapid development of hardware or software solutions. Mechanisms for resource requests for the general scientific community should also be developed and reviewed by the usage committee. The Kbase cloud should establish robust configuration mechanisms for all hardware, software and storage utilization. Resource allocation of cluster size, performance, and scheduling will also be reviewed by the usage committee. This committee will address issues such as the need for service level agreements and will provide recommendations on the type of service provided by the Kbase cloud.

Performance metrics. The Kbase administrators will establish performance metrics for the cloud resource. The metrics will monitor reliability, usage, and general utilization of the system. The Kbase cloud resource should also consider the use of surveys to monitor its utilization by the research and bioinformatics community. Surveys should address whether researchers are aware of the cloud resource, whether users utilize the system for their own research, whether any publications have resulted from the resource, and whether use of the system has contributed to the DOE mission. Usability studies should also be performed during the course of the cloud project. Users and workflow developers should be contacted to address the quality of services, documentation, and APIs (Application Programming Interface) of the cloud system.

Scientific advisory board. The Kbase cloud resource should include a scientific advisory board (SAB) to ensure the successful deployment of this facility and help achieve the scientific goals of the users. The SAB will meet on a regular basis and establish overall policies for resource allocation, accounting, and monitoring. The SAB should include representation from the general scientific community, individuals with technical expertise in cloud systems, Kbase staff, and IT administrators. All SAB meetings should be attended by DOE Program Officers and possibly other funding agency representatives. All survey results, performance metrics, usability studies, and newly developed procedures will be reviewed by the SAB and DOE Program Officers. The SAB will also review long-term strategic objectives of the cloud resource and specify new objectives when necessary during the course of the project.

Outreach and communication. The Kbase cloud should also dedicate resources to education, training, and outreach. Staff should provide training materials, on-line presentations, and

documentation so users can fully utilize all aspects of the system. Training should be directed to users and developers with a broad range of experience levels. Topics on how to perform genome annotation, assembly, and expression analysis as well as general workflow development techniques should be addressed. Incorporation of the Kbase cloud into existing class curricula in university courses should also be considered. The Kbase should also establish multiple modes of communication such as an online wiki, error-reporting systems, electronic newsletters, and an email contact list. Kbase staff should regularly present on the Kbase cloud at workshops and scientific conferences. Other forms of outreach such as posters, promotional materials, and advertising should be used.

Glossary*

Appliance (or Software appliance)

A software appliance is a software application that might be combined with just enough operating system (JeOS) for it to run optimally on industry standard hardware (typically a server) or in a virtual machine.

IAAS

Infrastructure as a Service (IaaS) is the delivery of computer infrastructure (typically a platform virtualization environment) as a service.

Omics

Informally refers to genomics, transcriptomics, proteomics, metabolomics, and other global molecular analyses that identify and measure the abundance and fluxes of key molecular species indicative of organism or community activity under defined environmental conditions at specific points in time.

SAAS

Software as a service (SaaS, typically pronounced 'sass') is a model of software deployment whereby a provider licenses an application to customers for use as a service on demand. SaaS software vendors may host the application on their own web servers or download the application to the consumer device, disabling it after use or after the on-demand contract expires.

In the Kbase context SAAS is often used to refer to a setup where the cloud service provider provides a fixed set of VM images that users can choose to run on the machine. The other option (open set of VM images) is often referred to as IAAS.

VM (virtual machine)

System virtual machines (sometimes called hardware virtual machines) allow the sharing of the underlying physical machine resources between different virtual machines, each running its own operating system. The software layer providing the virtualization is called a virtual machine monitor or hypervisor.

*sources include Wikipedia

Joint USDA-DOE Plant Genomics Knowledgebase Workshop Report

Setting the Stage for the Plant Knowledgebase Workshop: Bioinformatics Use in Advancing Plant Genomics, Genetics, and Breeding

Friday, January 8, 2010, 10:30 a.m. – 4:00 p.m.
Plant and Animal Genome XVIII
Town & Country Hotel
San Diego, California

Convened by the

U.S. Department of Energy (DOE)
Office of Science
Office of Biological and Environmental Research

U.S. Department of Agriculture (USDA)
National Institute of Food and Agriculture

Workshop Organizers: Catherine Ronning (DOE), Susan Gregurick (DOE), Ed Kaleikau (USDA), Gera Jochum (USDA), and Bob Cottingham (Oak Ridge National Laboratory)

Audience: 100 plant scientists, geneticists, breeders, and bioinformatics specialists

Speakers: Catherine Ronning (DOE Office of Biological and Environmental Research), Bob Cottingham (Oak Ridge National Laboratory), David Francis (Ohio State University), Steve Rounsley (University of Arizona), Eva Huala (The Arabidopsis Information Resource), Doreen Ware (Cold Spring Harbor Laboratory), and Dan Rokhsar (DOE Joint Genome Institute)

Introduction

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 (genomicscience.energy.gov/compbio/).

A DOE workshop held in May 2008 defined the vision for the Knowledgebase—an open cyberinfrastructure to integrate systems biology data, analytical software, and computational modeling tools that will drive two classes of work: (1) experimental design and (2) modeling and simulation. This community-driven Knowledgebase will need to be understandable and accessible to the entire research community and must have an intuitive design that facilitates sharing and contribution among all users. To provide computational capabilities that support DOE systems biology research and other application areas, the Knowledgebase will need to serve multiple roles, including a flexible, adaptable repository of data and results from high-throughput experiments; a collection of tools to derive new insights through data synthesis, analysis, and comparison; a framework to test scientific understanding; a heuristic capability to

improve the value and sophistication of further inquiry; and a foundation for prediction, design, manipulation, and, ultimately, engineering of biological systems.

For the USDA National Institute of Food and Agriculture, a grand challenge in plant genomics, genetics, and breeding is to identify gene combinations that lead to significant innovation in agriculture and production of raw materials for food, feed, fiber, and fuel. An interdisciplinary approach such as molecular plant breeding may be able to meet this challenge and revolutionize 21st century plant improvement. Molecular plant breeding is founded on the integration of advances in biotechnology, genomic research, and molecular marker applications with conventional plant breeding practices. This integration would require a combination of molecular markers and high-throughput genome sequencing efforts, new knowledge of genome structure and function, statistical approaches to estimate genetic effects, experience in both laboratory molecular methods and field-based breeding practices, and the ability to manage large datasets with diverse data types. This workshop also was intended to assist in developing strategies to expand bioinformatic tools to enable the breeder-centric, high-throughput data management and visualization tools and platforms necessary for integrating genome sequence information with other data types and to provide the breeder-centric views of map and trait data that best serve plant breeders' needs. Implementing such strategies will require (1) broadly training a new generation of plant researchers to fully master key areas such as bioinformatics and quantitative genetics and breeding; (2) establishing partnerships with universities, federal laboratories, industry, and international centers to take advantage of the best training opportunities; and (3) developing a new cohort of researchers able to translate and integrate basic research endeavors with applied plant improvement and value added outcomes for sustainable bioenergy production systems.

Workshop Description

The Plant Genomics Knowledgebase Workshop—held in conjunction with the Plant and Animal Genome XVIII conference in San Diego, California—brought together 100 plant scientists, geneticists, breeders, and bioinformatic specialists to discuss current issues facing plant breeders in light of ever-increasing amounts of genomic data. The workshop featured lectures by leaders in the plant breeding, genomics, and bioinformatics communities. These presentations set the stage for afternoon breakout discussions by addressing the data needs of more-applied breeding programs and describing resources emanating from more-fundamental plant genomics and bioinformatics research. This event is part of a series of DOE-supported workshops to engage the scientific community in discussing scientific objectives the Knowledgebase could serve and determining which endpoints could be achieved in the near, mid-, and long term.

The overarching goal of the workshop was to address the following question:

> How can we best design the Knowledgebase to have the flexibility to grow with and adapt to new data and information challenges in the future?

A key objective was to specifically identify the requirements for effectively developing data capabilities for systems biology as applied to plants, particularly the research and development of plant feedstocks for biofuels. The current state of plant informatics is represented by many

disparate databases primarily focusing on specific taxonomic groups or processes. To enable a systems biology approach to plant research, integrating all types of data (including molecular, morphological, and "-omics") for bioenergy-relevant plant species is important. Thus, the challenge will be to develop uniformity of data format and database architectures to effectively integrate diverse data types and enable user-friendly acquisition and analysis.

Charge Questions

All participants were asked to address two charge questions:

1. What types of experimental data are currently available, and of these, which format(s) are most useful and valuable? Can data from various sources and of various types be standardized into this "ideal" format and then be organized and integrated into one common, searchable application?

   For example, a researcher studying cell wall biosynthesis in grasses may benefit from work being performed in poplar. How can we best facilitate cross-species comparisons? How can we use these tools to leverage and apply knowledge gained from model species (e.g., *Arabidopsis* and rice) to crop plants?

2. What are the challenges for plant bioinformatics in a 2- to 3-year time frame? Given the development of an integrated, uniform system (Question 1), what types of analyses do you foresee, and what types of analysis tools will maximize the system's utility?

   How do we best organize, for example, pathways and processes, and how can we organize and distinguish common processes from those that are taxon-specific? How can these informatics resources best be used to enhance plant breeding (i.e., "genotype to phenotype")? Will these resources be effective in designing decision support tools for plant breeders in the field?

Summary of Workshop Recommendations

Three recommendations from the workshop are:

1. **Establish community–agreed upon data formats and storage protocols for environmental and experimental metadata and workflows.**

   This includes gene annotations; gene product functions; protein interactions; expression and methylation data; natural variation data; and phenotypic data such as geographical coordinates of a field, sampling dates, weather conditions, experimental designs, scoring methods, and images. Although some of these data types and metadata informatics have well-established formats and protocols, others do not and are not well linked to upstream genomic data. Standards development endorsed by the research community needs to be a collaborative and iterative effort between data generators and developers of cyberinfrastructure such as the Knowledgebase. Active, community-driven development of standards will require resource commitments in the form of coordination workshops; new tools to facilitate annotation and data deposition; curation; and compliance through journals, agencies, and peers.

2. **Develop the ability for comparative analyses of gene sequences, transcript and protein abundance, phenotypes, and the relationships among these components across multiple species.**

   Because the plant community comprises both systems biologists and plant breeders, the Knowledgebase must be adaptive to different user needs. Developing comparative analyses across species will require different levels of community support for different research needs. Moreover, coordination is needed among the various plant bioinformatic efforts sponsored by different agencies to avoid duplication of effort and to identify opportunities for collaboration.

3. **Establish long-term support for maintaining repositories of a variety of genomic and phenotypic data types.**

   This will be key to success of a knowledgebase that tries to integrate information from these resources.

Data and Analytical Challenges for Bioenergy Feedstocks

Although this workshop focused on data capabilities relevant to developing plant feedstocks for biofuels, many of the tools, approaches, and issues discussed are applicable to non-biofuel plant species, including well-studied model organisms such as *Arabidopsis*. Thus, Knowledgebase efforts can be leveraged to other plant bioinformatics systems and biological research areas and vice versa. Workshop participants identified several data and analytical issues for plant genomics, including the diversity of data types available, the challenges of dealing with phenotypic data, cross-species analyses, data integration, and standards for interoperability among data and information resources.

Available Data Types for Plant Genomics

The range and quality of available data depend on the extent to which a particular genome has been studied. For a well-studied model organism like *Arabidopsis*, a broad range of data types supported by a rich history of published research helps researchers move from gene sequence to molecular function, associated phenotype, relevant metabolic or regulatory pathways, and interaction partners. As the types of data being generated for different species of bioenergy feedstocks continue to grow, a top priority will be developing appropriate repositories for handling each data type.

What Kinds of Data are Available from Arabidopsis Research?

- **High-quality genome annotation.** The annotated genome forms the basis of all other "-omics" data. The *Arabidopsis* genome has been revised nine times since the initial sequence was completed in 2000, and its annotation continues to evolve. Current revisions to the annotation include adding splice variants and untranslated regions (i.e., 5' and 3' UTRs) as transcript data improves, correcting sequencing errors, and adding features that are more difficult to annotate such as noncoding RNAs and genes that encode small proteins. In the last 5 years, The Arabidopsis Information Resource (TAIR)

has added or updated about half of the genes in the current release, and large, new datasets continue to be generated. Revising genome annotation is a continuous process.

- **Experimental gene function data.** In addition to refining gene structures, TAIR curators have been adding gene function annotations based on experimental data from research articles. To date, 8,622 genes have been annotated with results from published experiments—a total that continues to increase rapidly. Most data used in this manual annotation process were not from high-throughput experiments but from those focusing on a single gene. Gene Ontology annotations describe biological process, molecular function, and cellular compartment. Plant ontology annotations describe the anatomical part and developmental stage associated with expression patterns. This is a rich dataset to consider transferring to other plant species.
- **Phenotypic data** for *Arabidopsis* have largely been qualitative. Currently, these data are in a free-text form, and efforts are needed to use a plant ontology to describe these phenotypes. *Arabidopsis* phenotypic data also include about 5,000 images in a form that is not yet readily transferable to other plants.
- **Protein interaction data** build on existing foundational datasets to generate networks of interactions.
- **Natural variation data** include more quantitative data than some other kinds of data.
- **Expression, methylation, pathways, and networks data** provide more of a genome-wide view of how this plant functions.

Next-Generation Sequencing Data. With the expanding use of next-generation sequencing technologies such as Illumina and 454, an important challenge will be dealing with the vast volume and variable quality of short read sequences generated by many different sources. Needed are resources for assembling and curating these massive amounts of data and tools for using the data to identify and develop single nucleotide polymorphism (SNP) markers, such as the current iPlant effort.

Environmental Metadata. One of the more difficult data challenges identified by workshop participants will be defining appropriate data formats and storage protocols for environmental and experimental metadata. Such data include geographical field coordinates, sampling dates, weather conditions, experimental design, scoring methods, and images. Metadata collection systems will need to be standardized and automated (e.g., using bar codes).

Phenotypic Data Challenges

Enabling large-scale generation of useful phenotypic data and ensuring easy access to it are some of the most important challenges for the bioenergy feedstock research community. Many bioinformatic efforts for plant biology have emphasized non-phenotypic data (e.g., DNA sequence, SNP markers, gene expression, and epigenetics). Phenotypic data—an extremely broad category of data—are subject to considerable noise, have few or no uniform standards, and are highly dependent on genetic context (e.g., particular individuals that have a specific genotype) and environmental context (e.g., timescales, locations, and precipitation). Some

critical challenges identified by workshop participants include developing standards and more efficient methods for generating and managing phenotypic data, improving the ability to link specific genes to phenotypes on a quantitative scale, and establishing central repositories for storing phenotypic data and genetic material. One organization that will address these issues is the International Plant Phenomics Initiative (www.plantphenomics.com), which is being organized by European, Canadian, and Australian researchers to promote international collaboration for plant phenomics.

Limited Availability of Phenotypic Data. The availability of phenotypic data is key to identifying quantitative trait loci and genes associated with important bioenergy-related traits. However, phenotypic data is currently limited. At present, for example, 33% of the protein-coding genes in the *Arabidopsis* genome have experimental annotations, and only 9% have phenotypic descriptions (including "no visible phenotype"). Understanding of genes in biofuel species is even less developed. The lack of robust phenotypic data and functional annotation will result in the continued extensive use of transitive annotation based on sequence similarity from generic databases such as Pfam and UniProt. This is a primitive approach to improving our understanding of plant biology with respect to biofuels.

The amount of meaningful phenotypic data available in public databases is very small compared to the amount of genomic data available. Moreover, the limited resources handling phenotypic data do not address all phenotypes and often do not include lines used for breeding. They thus are not providing breeders with needed information. Participants suggested the need to develop a system of quality scores that could provide a measure of confidence for the heritability and/or measurement of a particular phenotype.

For many applied breeding objectives, a greater focus is needed on generating more phenotypic information in more populations of a given species and, importantly, generating data in actual elite breeding populations. Collecting phenotypic data for complex traits in plants is time consuming. Many potential bioenergy crops are perennial, so successive-year data are needed for individual plants or accessions—information difficult and expensive to obtain. In addition, measuring environmental effects on phenotypes, which requires quantitative data, is as important as defining genotype.

More Objective and Quantitative Phenotypic Data. Descriptors supported by the Union for the Protection of New Varieties of Plants (UPOV) or USDA's National Plant Germplasm System (NPGS) Germplasm Resources Information Network (GRIN) are used by breeders to classify traits into defined categories. For example, in GRIN, a trait such as color is assigned a numerical color category like "1" for green or "2" for yellow. Although these descriptor systems attempt to make all trait data more uniform, they fail to account for inherent variation within an accession (e.g., how "green" is it?). They also are disassociated from contemporary systems of measurement and disconnected from data scales used by expert practitioners.

Trait data should be quantitative and objective whenever possible. For example, there are very objective systems for measuring color, such as the RGB system for computers. High-throughput systems are needed that can extract quantitative phenotypic data from images. The advantages of such objective measures for phenotype are clear: the ability to interconvert systems of

measurement and the ability to easily obtain estimates of variation (and therefore estimates of heritability). Tools that enable the mapping of one system to another (e.g., a descriptor to an ontology and a scale to quantitative data) also are needed. The Australian Plant Phenomics Facility (www.plantphenomics.org.au/) is developing high-throughput phenotyping platforms for reproducibly capturing quantitative phenotypic data in parallel with environmental conditions.

Organizational Systems for Phenotypic Data. One organizational system pioneered through GRAMENE and other "-omics"-related projects (www.plantontology.org) uses hierarchical ontology for trait data, with vocabularies derived from published sources and terms appropriately defined. Input from user communities outside the basic researcher vary within trait ontology efforts. Within the international Solanaceae sequencing effort and the Solanaceae Genome Network, interaction with applied research communities is growing, and the system is proving flexible enough to account for diverse traits. Efforts are under way to ensure that ontologies are consistent with existing descriptors and have quantitative definitions. However, use of these ontologies by the community is lagging, an issue that needs to be addressed. Other initiatives include the development of Phenom-Networks (phnserver.phenome-networks.com/icis/), a web-based system to import raw data and facilitate analysis across experiments. Phenom-Networks draws its standards from the International Crop Information System (ICIS), a framework for integrated management of crop-improvement data for both individual crops and farming systems. The ICIS framework is being developed by the Consultative Group on International Agriculture (CGIAR) and has established guidelines for germplasm and data management (www.icis.cgiar.org/icis/index.php/ICIS_Concepts).

Support for Germplasm Stock Centers. The availability of germplasm (plant genetic material) linked to genetic information presented in bioinformatic resources will strongly influence both the value and audience of these resources. Germplasm housed within the National Plant Germplasm System (NPGS) is of historical interest but often does not meet the needs of breeding programs today. In contrast, immortal mapping populations (e.g., recombinant inbred lines and segmental substitution populations) may be too limited for broad inferences or may be based on accessions that are more interesting to basic scientists than those actively engaged in crop improvement. Databases designed to foster crop improvement will need to accommodate mapping populations, breeding populations and pedigrees, and germplasm accessions as defined by the user community. Permanent, long-term support to maintain a germplasm stock center for bioenergy-related species is critical.

Cross-Species Analyses

An important goal is developing databases that permit comparative analyses of gene sequences, transcript and protein abundance, phenotypes, and the relationship among these components across multiple species so that the value of genomic information can be expanded. However, the challenge of developing databases designed to be useful across species begins prior to data collection or formatting. It is critical that gene orthologs, experimental conditions, and genotypes be considered before any meaningful comparison can be achieved between any two genomic datasets. Also highly valuable would be resources for connecting gene or protein

expression data and other information available in the database for one species to the most likely ortholog in other species.

A few critical data integration requirements must be considered when developing the standards and tools needed to connect data across species. These requirements include defining common terms for gene function among different species, having high-quality genome annotations with accurate depictions of gene structures, and obtaining standardized ortholog sets for navigating between genomes. In addition to comparing across species, analytical tools are needed that permit meta-analysis across experimental studies.

Defining "Data Integration"

Workshop speaker Eva Huala noted in her presentation that although "data integration" is a widely used expression, it can have different meanings depending on audience and context. For example, "data integration" can mean:

- Integration of data from many experiments of a similar type in a single species (e.g., many different microarray experiments on *Arabidopsis*).
- Integration of data from experiments of different types in a single species (e.g., gene expression, protein expression, metabolic pathways to generate a network diagram or create a summary of all data for one gene).
- Integration of data from two or more species.
- Use of an integrated dataset to extract new knowledge.

Each type of "data integration" involves different sets of problems and bottlenecks. Determining whether or not data have been integrated appropriately entails much more than simply combining data; it also involves determining whether or not useful information can be extracted from the combined data.

Standards for Interoperability

There is a perception that funding practices and cultural pressures for attaining professional recognition within research communities often encourage the development of more new tools and bioinformatic resources rather than support maintenance and improvement of existing resources. With this push to build isolated, project-specific bioinformatic resources, there is little incentive to set the standards needed to promote interoperability among these resources. User metrics, such as web statistics and literature citations, are useful for evaluating the impacts and quality of tools, databases, or datasets.

When summarizing recommendations from the Workshop on Plant Bioinformatics and Databases sponsored by the European Commission-United States (EC-US) Task Force on Plant Biotechnology Research, Doreen Ware noted several efforts for which standards development could help create a unified platform for plant genome biology:

- **Assessments of genomic tools and datasets.** Establishing periodic assessments of important genomic tools and datasets, similar to CASP (Critical Assessment of Techniques for Protein Structure Prediction), will be important for monitoring the

quality of datasets and selecting the best tools for data analysis and integration. This effort will be essential for ensuring best practice and quality for reference data.

- **Genome sequence assemblies.** There is a life cycle associated with sequence datasets whereby additional improvement in the annotation for a reference genome sequence is needed even after the reference genome has been completed. With recent developments in next-generation sequencing, a standardized system should be established for evaluating the quality of sequence assemblies. Mechanisms are needed to describe the range of genome sequence models and assemblies that can now be produced, and researchers need to understand the status and quality of the genome annotations with which they are working.
- **Plant-specific ontologies.** Multiple plant-specific databases have ontologies, but there are no consistent standards among them. Coordinated efforts are needed with respect to controlled vocabularies for data collection and submission across databases, such as those used by the Plant Ontology Consortium (PO; www.plantontology.org), as well as leadership-driven efforts to generate phenotype ontologies. For phenotypes relevant to plant breeding activities, a system is needed for linking the terms used in genomic functional annotations to the phenotype terms used by breeders.
- **Curation.** For community-based curation and the curation of legacy data, there currently are no agreed-upon standards.

Knowledgebase Usability and Data Availability Issues

Long-term Sustainability of Data and Databases

Workshop participants were concerned that expiration of funding for existing databases could be problematic for sustaining the availability of important data types. This issue applies to both small boutique databases and larger community databases. Transfer of data from small, project-based databases into larger, more permanent data repositories can be difficult because of differences in schema design and scope. An additional challenge for small project-scale databases is frequent periods of unavailability due to server or network problems. Although a standard database schema (Chado) exists, it is not ideal for all purposes and has performance issues for high data volumes and usage levels. Participants also noted that getting funding for new databases currently is easier than securing continued funding for existing databases, compounding the problem of data longevity. Some participants believe the creation of new resources should continue to be the funding priority, since development of something new ensures that it will be tailored to the needs of the project. Others think that funding support needs to shift toward promoting reuse of existing resources and tools to encourage emergence of standards. Promoting reuse would require that money be made available for adapting existing tools to fit new projects, as there is always some work to be done before an existing tool or standard can be used.

Cultural Differences within the Potential User Community

The diversity of the potential Knowledgebase user community suggests that a one-size-fits-all solution may be difficult to achieve. A user's scientific culture influences how he or she views

data and asks questions about the data, so different users within the plant-science community have different needs and expectations from a knowledgebase. A systems biologist, for example, needs tools to discover how a plant works. A plant breeder, however, is simply interested in predicting the phenotype that results when a particular genotype is grown in a certain environment—without really needing to know how and why the observed phenotype is produced. In this case, black box methods for predicting phenotypes may suffice. A knowledgebase therefore needs to be adaptable to the different needs of diverse users. Although many existing bioinformatics resources have focused on engaging molecular biologists, genome scientists, computational biologists, and bioinformatics specialists, more effort is needed to bridge the gap and explore the information needs of users in more-applied fields such as plant breeders and crop scientists.

User-Dependent Data Formats

Users whose daily work is focused on plant breeding or laboratory experiments want to access bulk data in relatively simple formats (e.g., CSV flat files or GFF) for further manipulation on their own computers. Some interest was expressed in portability of data or databases so that work could be performed offline (e.g., while traveling or in remote areas where internet access is slow or unavailable). Other users with a more computational focus preferred more complex data formats such as XML. Nexus format for phylogenetic data also was suggested as a good standard. Participants pointed out that certain data types (e.g., sequence and microarray data) already have well-defined standards. In general, many scientists do not want to spend time addressing format issues; they want data presented to them in an intuitive way that does not require them to become programming experts.

Education, Training, and Communication

In the life sciences, adopting informatic resources requires a user community that is educated in bioinformatics concepts, methods, and tools and is equipped with skills in computational and quantitative analytical approaches from the fields of computer sciences, statistics, and mathematics. A key problem is a lack of people with sufficient training to fully exploit the genomic information and resources available. Training the current and next generations of biologists in computational and statistical methods is a major challenge.

In the physical sciences, the computational skills required to manipulate large datasets are considered indispensable and are taught to every undergraduate and graduate student in these disciplines. Similar training in computational approaches to biology is needed at all levels, especially the undergraduate. Workshop participants specifically proposed pre- and postdoctoral cross-training fellowships in quantitative genetics, bioinformatics, and computational biology of biofuel species. For maximum impact, these fellowships should not be tied to standard research grants, where typical 3-year cycles would impede recruitment of fellows, as the hire needs to be coordinated with the duration of the grant.

Existing databases can play an important role through tools that assist self-learning (e.g., online tutorials). Although there is a need to provide tools simple and intuitive enough for those without computational training, these resources should be designed to gradually enhance

understanding of underlying concepts and progressively lead the user to use the tools in more sophisticated ways. An example is a query tool that provides canned statements (in Structured Query Language or other appropriate formats) that can be altered easily by users to fit their particular needs. Eventually a user should be able to write new queries based on the knowledge gained from using and modifying the examples.

Plant Bioinformatic Efforts Relevant to Knowledgebase Development

Two ongoing bioinformatics efforts for plant biology were featured in the presentations at this workshop: the iPlant collaborative funded by the National Science Foundation (NSF) and presented by Steve Rounsley (University of Arizona) and the DOE Joint Genome Institute's Phytozome, presented by Dan Rokhsar (JGI).

The iPlant Collaborative: Cyberinfrastructure for the Plant Sciences

The NSF-supported iPlant Collaborative is an effort to develop a cyberinfrastructure that is nimble enough to address an evolving array of plant science grand challenges. According to NSF, the cyberinfrastructure is a combination of High Performance Computing (HPC), data, data analysis capabilities, and virtual organizations that also can serve as a resource for training and workforce development. The collaborative establishing iPlant includes more than 25 institutions and 45 additional researchers and continues to grow. Once the research community identifies the major problems in plant sciences, iPlant's mission is to provide the cyberinfrastructure that brings together the information needed for researchers to address these grand challenges.

iPlant's community-driven process identified two grand challenge projects that will be the focus over the next 2 years:

1. **Plant Tree of Life (iPToL).** The iPToL goal is to "build the cyberinfrastructure needed to scale up phylogenetic methods by 100-fold or more, to enable the dissemination of data associated with such large trees, and to implement scalable 'post-tree' analysis tools to foster integration of the plant tree of life with the rest of the botanical science." The largest phylogenetic tree that currently can be built is about 100-fold smaller than the number of green plants that exist. For this grand challenge, iPlant aims to design the computational approach that can be used to build a tree with 500,000 taxa in it. Using algorithms available today, the largest trees that can be built contain about 55 taxa and take about 3,000 CPU hours to construct. Some of the significant computational bottlenecks that iPToL will address will require redesigning algorithms. In addition to providing needed cyberinfrastructure, this project involves building, visualizing, and extracting data from the trees.
2. **Genotype to Phenotype (iPG2P).** The goal of the Genotype to Phenotype grand challenge is to elucidate "the relationship between plant genotypes and the resultant phenotypes in complex (e.g., non-constant) environments, one of the foremost challenges in plant biology." Although solving this grand challenge is not possible in a 2-year time frame, the project aims to help overcome the current computational and data management bottlenecks preventing researchers from even attempting to address this challenge today. Much of this effort concerns handling the different data generated

from genomics experiments (e.g., sequence, expression, metabolic, whole-plant, environmental), integrating these data, bringing in the modeling and statistical inference tools to analyze the data, visualizing the results, and providing the interfaces that researchers can use to work with their own results.

Phytozome: A DOE JGI Resource for Green Plant Comparative Genomics

The DOE Joint Genome Institute (JGI) has developed a central hub (www.phytozome.net) to provide all researchers with an interface to interact with plant genomic data in a unified way. About 20 plant genomes are included in version 5 of Phytozome, and a year from now JGI is expected to have 50 genomes of similar quality. All the genomes in Phytozome are reasonably high quality drafts, with enough data available to provide an approximation of the gene set. The genomes at Phytozome range from *Arabidopsis*, which is a highly developed, well-annotated genome, to cassava, which is a 454 draft genome that has just recently become available.

Genomes can serve as an organizing principle for much of the information emanating from modern biological studies. Looking across a phylogenetic tree of angiosperms, the timescale for their radiation is comparable to diversification of mammals (~150 million years), so the extent of diversity seen in angiosperms parallels what is seen among mammals (ranging from bats to elephants to humans). Thus, the work that has been done to compare mammal and other animal genomes indicates where comparisons of plant genomes could be in a few years. Genomes are a central axis for moving from organism to organism and seeing how different species have evolved, and certain comparisons between two different species can be useful in identifying particular kinds of candidate functional elements.

Principles Guiding Future Development of Phytozome

- Adopt open-source, community standards where possible, pulling from advanced comparative genomics already under way in vertebrates.
- Provide standardized datasets to the community. Although several versions of annotation for a genome may exist, the research community needs to agree that one version serves as the reference set at any given time.
- Take advantage of the handful of reference genomes (e.g., *Arabidopsis* and maize) that have benefited from a richer history of past research to help develop resources for the numerous new genomes that will be generated from Illumina and 454 sequencing.
- Continue to develop genome annotation assistance and browsers [e.g., JGI plant pipeline and GMOD (Generic Model Organism Database project)] using open-source community standards so that any researcher can locally set up a customized GBrowse for a particular species.
- Improve an array of features by building on existing resources:
    - "Phylogenomic" gene families (calibrated molecular divergence, synteny, molecular phylogenetic methods).
    - Comparative genomics taking advantage of VISTA and comparative tools for animal genomes.

  - Genomic diversity that builds on resources developed for human HapMap.
  - Complex queries. A guiding principle is to be able to download data in a standardized format that researchers can use in a customized way.
  - Customized analysis. GALAXY and other tool kits are built to hold data in a standardized format. Once a tool is brought into GALAXY, anyone can use it on any genome.
  - Links to TAIR, DOE Bioenergy Research Center knowledgebase efforts, iPlant, and other resources.
- Support workshops to systematically annotate the gene complement across plants.

Appendix 1: Agenda

USDA National Institute of Food and Agriculture Plant Genome, Genetics, and Breeding Project Directors' Meeting

and

Joint USDA-DOE Plant Knowledgebase Workshop

Town and Country Resort and Convention Center
San Diego, California
Friday, January 8, 2010

7:30 a.m.	Light refreshments available
7:45 – 10:00 a.m.	**Morning Session I: Plant Genome, Genetics, and Breeding** *(Pacific Salon 3)*
7:45 a.m.	Ed Kaleikau, USDA NIFA "AFRI Plant Genome, Genetics and Breeding Program"
8:00 a.m.	Phil McClean, North Dakota State University "BeanCAP – A NIFA Coordinated Agricultural Project"
8:20 a.m.	Scott Jackson, Purdue University "Genome Sequence for Common Bean"
8:30 a.m.	Gary Muehlbauer, University of Minnesota "Barley Coordinated Agricultural Project: Leveraging Genomics, Genetics, and Breeding for Gene Discovery and Barley Improvement"
8:50 a.m.	Tim Close, University of California, Riverside "Advancing the Barley Genome"
9:00 a.m.	Jeff Bennetzen, University of Georgia "Development of Genomic and Genetic Tools for Foxtail Millet: Use of these Tools in the Improvement of Biomass Production for Bioenergy Crops"
9:20 a.m.	John Vogel, USDA ARS "*Brachypodium distachyon*: A New Model for the Grasses"
9:40 a.m.	Peter Bretting, USDA ARS "GRIN-Global: An International Project to Develop a Global Plant Genebank Information Management System"
10:00 – 10:30 a.m.	Break
10:30 – 12:30 p.m.	**Morning Session II: Setting the Stage for the Plant Knowledgebase Workshop: Bioinformatics Use in Advancing Plant Genomics, Genetics, and Breeding** *(Pacific Salon 3)*

10:30 a.m.	Cathy Ronning, DOE BER "Introduction to the Workshop"
10:35 a.m.	Bob Cottingham, Oak Ridge National Laboratory "DOE Systems Biology Knowledgebase"
10:50 a.m.	David Francis, Ohio State University "A Plant Breeding Perspective"
11:10 a.m.	Steve Rounsley, University of Arizona "The iPlant Collaborative"
11:30 a.m.	Eva Huala, TAIR "Leveraging *Arabidopsis* Data for Research on Other Plant Species"
11:50 a.m.	Doreen Ware, Cold Spring Harbor Laboratory "US-EC Plant Bioinformatics"
12:00 p.m.	Dan Rokhsar, Joint Genome Institute "Genomes as an 'Organizing Principle' for the Knowledgebase"
12:20 p.m.	Instructions and Move to Breakout Rooms
12:30 – 2:30 p.m.	**Working Lunch: Plant Knowledgebase Breakout Sessions and Discussion** 5 Groups; Facilitators: Rex Bernardo, Steve Knapp, Robin Buell, Lukas Mueller, and Todd Mockler *(Pacific Salons 2, 4, 5, 6, 7)*
2:30 – 2:45 p.m.	Coffee Break
2:45 – 4:00 p.m.	Report out (15 minutes for each group) *(Pacific Salon 3)*
4:00 – 4:30 p.m.	Facilitators gather to summarize and wrap up *(Pacific Salon 3)*
4:00 – 6:00 p.m.	**Poster Session** *(Golden Ballroom)*

Appendix 2: Participants and Observers

Participants

Eduard Akhunov (Kansas State U.)
Steve Baenziger (U. Nebraska)
Ali Barakat (Pennsylvania State U.)
William Barbazuk (U. Florida)
Eric Beers (Virginia Tech U.)
Jeffrey Bennetzen (U. Georgia)
Rex Bernardo (U. Minnesota)
William Berzonsky (S. Dakota State U.)
Jim Bradeen (U. Minnesota)
Charles Brummer (U. Georgia)
Marcia Buanafina (Pennsylvania State U.)
Robin Buell (Michigan State U.)
John Burke (U. Georgia)
Victor Busov (Michigan Technological U.)
Patrick Byrne (Colorado State U.)
John Carlson (Pennsylvania State U.)
Tim Close (U. California - Riverside)
Luca Comai (U. California - Davis)
Carlos Crisosto (U. California - Kearney)
Richard Cronn (USDA FS)
Thomas Davis (U. New Hampshire)
Katrien Devos (U. Georgia)
Amit Dhingra (Washington State U.)
David Douches (Michigan State U.)
Andrew Doust (Oklahoma State U.)
Jorge Dubcovsky (U. California - Davis)
Ismail Dweikat (U. Nebraska)
David Francis (Ohio State U.)
Bikram Gill (Kansas State U.)
Jim Giovannoni (Cornell U.)
Jose Gonzalez (S. Dakota State U.)
Pam Green (U. Delaware)
Maria Harrison (Cornell U.)
Patrick Hayes (Oregon State U.)
Sam Hazen (U. Massachusetts)
Eva Huala (TAIR)
Amy Iezzoni (Michigan State U.)
Eric Jackson (USDA ARS)
Scott Jackson (Purdue U.)
James Kelly (Michigan State U.)
Matias Kirst (U. Florida)
Steve Knapp (U. Georgia)
Jan Leach (Colorado State U.)
Thomas Lubberstedt (Iowa State U.)
Laura Marek (Iowa State U.)
Michael Mazourek (Cornell U.)
Phil McClean (North Dakota State U.)
Susan McCouch (Cornell U.)
Richard Michelmore (U. California - Davis)
Amit Mitra (U. Nebraska)
Todd Mockler (Oregon State U.)
Gary Muehlbauer (U. Minnesota)
Lukas Mueller (Cornell U.)
Seth Murray (Texas A & M U.)
David Neale (U. California - Davis)
Joseph Onyilagha (U. Arkansas - Pine Bluff)
Jiwan Palta (U. Wisconsin)
Cameron Peace (Washington State U.)
Zhaohua Peng (Mississippi State U.)
Andy Pereira (Virginia Tech U.)
Dan Rokshar (JGI)
Pam Ronald (U. California - Davis)
Jeffrey Ross-Ibarra (U. California - Davis)
Steve Rounsley (U. Arizona)
John Warner Scott (U. Florida)
Kevin Smith (U. Minnesota)
Carol Soderlund (U. Arizona)
David Spooner (U. Wisconsin)
Dina St. Clair (U. California - Davis)
Steve Strauss (Oregon State U.)
Christian Tobias (USDA-ARS)
Jerry Tuskan (ORNL)
Allen Van Deynze (U. California - Davis)
Richard Veilleux (Virginia Tech U.)
Wilfred Vermerris (U. Florida)
John Vogel (USDA-ARS, Albany CA)
Dong Wang (U. Nebraska)
Shizhong Xu (U. California - Riverside)
Janice Zale (U. Tennessee)
Hongyan Zhu (U. Kentucky)

Observers

Peter Bretting (USDA)
Randy Johnson (USFS)
Ed Kaleikau (USDA)
Shing Kwok (USDA)
Liang-Shiou Lin (USDA)
Gail McLean (DOE)
Jack Okamura (USDA)
Jane Silverthorne (NSF)
Sharlene Weatherwax (DOE)

DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop

This workshop was part of the U.S. Department of Energy Office of Science's 2010 Genomic Science Contractor-Grantee and Knowledgebase Workshop in Crystal City, Virginia, February 7–10, 2010.

Organized By: Susan Gregurick (U.S. Department of Energy)
Robert Cottingham (Oak Ridge National Laboratory)

Co-Chairs: Adam Arkin (Lawrence Berkeley National Laboratory; University of California, Berkeley)
Robert Kelly (North Carolina State University)

Report Contents

Section I: Knowledgebase Concept and Workshop

Section II: Workflows—Knowledgebase Use Cases

Section III: Strawman Knowledgebase Architecture

Section IV: Workshop Summary and Conclusions

Section I: Knowledgebase Concept and Workshop

The Department of Energy (DOE) Genomic Science program within the Office of Biological and Environmental Research (BER) supports science that seeks to achieve a predictive understanding of biological systems. By revealing the genetic blueprint and fundamental principles that control plant and microbial systems relevant to DOE missions, the Genomic Science program (genomicscience.energy.gov) is providing the foundational knowledge that underlies biological approaches to producing biofuels, sequestering carbon in terrestrial ecosystems, and cleaning up contaminated environments.

Knowledgebase Vision and Background

The emergence of systems biology as a research paradigm and approach for DOE missions has resulted in dramatic increases in data flow from new generations of experimental technologies in areas such as genomics and imaging. While some resource centers are generating large datasets with workflows designed to answer specific scientific questions, there is also a great increase in data production, generally from individual laboratories. New scientific questions arise and can be answered by combining and analyzing such data across laboratories and projects. Great value has derived from the ability to combine sequence and structure data across producers, and in some research communities, such as the yeast field, general access to functional genomic data has greatly accelerated discovery and technology development. Over the last decade, BER—through its Genomic Science program—has sought to solve bioenergy, environmental remediation, and carbon sequestration challenges that demand understanding biological activities exhibited by complex populations and the individuals within them. Since we seek to understand the molecular basis of these dynamics and activities on scales from individual genomes through cellular networks to community function and evolution, these projects are generating multiscale information that could be organized more effectively to aid

the science of individual projects and to synergize data across projects with related missions. Perhaps even more important, the data from multiple, possibly unrelated programs could be flexibly reorganized and analyzed to aid new scientific discoveries and provide insight to researchers in environmental microbiology and biotechnology generally.

Enabling the community to serve, query, combine, and analyze these diverse data types is therefore imperative, as is building a blueprint and system to enable the design, implementation, and use of new analytical tools and frameworks for working with such data. To manage and effectively use this ever-increasing volume and diversity of data, the Genomic Science program is developing the DOE Systems Biology Knowledgebase—an open, community-driven cyberinfrastructure for sharing and integrating data, analytical software, and computational modeling tools. Historically, most bioinformatics efforts have been developed in isolation by people working on individual projects, resulting in isolated data and methods. An integrated, community-oriented informatics resource such as the Knowledgebase would provide a broader and more powerful tool for conducting systems biology research relevant to BER's complex, multidisciplinary challenges in energy and environment. It also would be easily and widely applicable to all systems biology research.

In general, a knowledgebase is an organized collection of data, organizational methods, standards, analysis tools, and interfaces representing a body of knowledge. For the DOE Systems Biology Knowledgebase, these interoperable components would be contributed and integrated into the system over time, resulting in an increasingly advanced and comprehensive resource. Other elements of the Knowledgebase vision are defined in a March 2009 report (genomicscience.energy.gov/compbio/) based on a DOE workshop that brought together researchers with many different areas of expertise, ranging from environmental science to bioenergy. The report highlights several roles the Knowledgebase will need to serve, including:

- An adaptable repository of data and results from high-throughput experiments;
- A collection of tools to derive new insights through data synthesis, analysis, and comparison;
- A framework to test scientific understanding;
- A heuristic capability to improve the value and sophistication of further inquiry; and
- A foundation for prediction, design, manipulation, and, ultimately, engineering of biological systems.

Beyond these perspectives from the last report, the Knowledgebase is now envisioned as a robust, flexible, and well-documented open architecture. This architecture would allow for both organized and distributed community development, facilitate the sharing of data and tools for data transfer, integration, query, analysis, and visualization, and be committed to interoperating with community resources and standards.

The Knowledgebase would differ from current informatics efforts by integrating data and information across projects and laboratories—tracking diverse, multiscale biological data from the genome through molecular networks, to cellular populations and communities, to environmental function, and combining data centralization with distributed data. Integration

implies that the Knowledgebase be a community effort rather than a monolithic project overseen and contributed to by only a few people. The Knowledgebase also will need to be more standardized than today's informatics resources. Although standardized components may not always be "cutting edge," they will be more interoperable, enabling comparisons among different laboratories and thus yielding important new insights. Standardization will involve not only data but also experimental protocols.

Another fundamental feature is that Knowledgebase development will have a more mature software engineering approach. In the past, biologists not necessarily trained in state-of-the-art computational techniques reasonably applied the computational tools of the day to their research. However, the dramatic increase in the amount of sequencing and other data being generated requires the support of a more robust computational infrastructure, with analyses that no longer are carried out in an *ad hoc* manner. Many current development efforts are based on computational technologies created 10 to 15 years ago. More modern analytical technologies are needed. To be useful, these new techniques must be developed by the entire research community rather than by informatics specialists working in isolation.

To establish the Knowledgebase as a community effort, several basic principles need to be considered. One is *open access*—the concept that data and methods contributed to the system will be available for anyone to use. Another is *open source* or *open contribution*, meaning that source code is managed in an open environment and is freely available to access, modify, and redistribute under the same terms. Perhaps the most important concept is *open development*, which would allow anyone to contribute to Knowledgebase development under organizational guidelines. Analogous to submitting a publication, this would involve a review process by an authoritative group that would determine if a particular contribution meets established criteria. In such an environment, different groups would work together on a common piece of software to meet common needs, the review process would facilitate integration into the Knowledgebase and quality control, and the product would be better than what an individual alone could create.

Several existing systems and applications can serve as reference models for thinking about Knowledgebase development. Exemplifying the concept of an open-source development is the computer operating system Linux, which is being built by a community of software developers working collaboratively to create a sophisticated and fairly successful system. Other familiar examples include iPhone or Google apps that enable users to pick and choose the kinds of features and capabilities they want and integrate them into a phone or other device. We are familiar with user interfaces that show layering of data from Google maps and Google Earth annotations (e.g., locations of landmarks and restaurants). Experimental design and research in the future will be conducted in the context of a user model similar to these successful systems. As research users gain new insights in systems biology from experiments and analyses, their interaction with the Knowledgebase populates new detail in the biological systems, forming the basis for new referential insight.

Wikipedia development also is open source and open development, allowing individuals or groups to contribute content. It has an editorial model, and, over time, the quality of its content evolves and improves. For the Knowledgebase, such an open-development environment

conceivably would enable noncomputing experts to play a role in the project's development and evolution.

Although these historical examples are approximations of the Knowledgebase vision, they provide a notion of possibilities in their commonly understood characteristics of flexible community development, data layering, editorial control, and peer review integration. The take-away lesson is that we see the initial Knowledgebase development like an operating system kernel that provides a platform on which open contribution of new applications can occur while the Knowledgebase simultaneously is managed to provide core functions like protection of legacy data and development of the underlying access and sharing model and architectural methods.

Workshop Description, Goals, Inputs, and Outputs

Although the 2009 Knowledgebase report describes a vision and long-term objectives for the Knowledgebase, it does not provide details about a plan to implement the system. To that end, DOE has launched an R&D project to establish the requirements for the Knowledgebase and to outline a plan for implementing them. As part of this project, DOE is sponsoring a series of community workshops. The first—held in conjunction with the November 2009 Supercomputing conference in Portland, Oregon—explored the potential for applying the cloud computing approach to systems biology research. The second workshop—held prior to the January 2010 Plant and Animal Genome meeting—addressed the Knowledgebase requirements necessary for developing data capabilities for plants. The output for these and subsequent workshops is now or will soon be posted online at www.systemsbiologyknowledgebase.org/workshops. As the third event in this series, the DOE Genomic Science Microbial Systems Biology Knowledgebase workshop was held Feb. 9–10, 2010, during the DOE Genomic Science Contractor-Grantee meeting in Crystal City, Virginia.

The goals of this workshop were to outline the near-, mid-, and long-term trajectory of microbial sciences for energy and environment and to map the associated workflows and data integration methods that can inform Knowledgebase specifications and requirements. Participants were asked to provide responses to six charge questions:

1. For systems biology of interest to genomic sciences, what are the scientific objectives that a knowledgebase could address in both a 5-year and longer time frame?
2. What are the key workflows that could be developed to accomplish these goals? Provide comprehensive usage examples that lead to scientific objectives.
3. What types of data are required to accomplish these objectives?
4. What bottlenecks to data integration and data usability need to be addressed to accomplish these goals?
5. What bottlenecks in bioinformatic and computational algorithms need to be addressed to accomplish these goals?
6. What would success look like? What would the benefit be?

The workshop featured presentations discussing the current, near-, and long-term prospects for microbial systems biology research in the context of the Knowledgebase. Formal presentations were given by Robert Cottingham (Oak Ridge National Laboratory) describing Knowledgebase background and objectives, by Robert Kelly (North Carolina State University) on the "Near-Term Prospects for Functional Microbial Genomics: Moving Beyond the Monoculture Paradigm," and by Adam Arkin (University of California, Berkeley, and Lawrence Berkeley National Laboratory) on "From Pathways to Populations and Back Again: Long-Term Prospects for the Microbial Systems Biology Knowledgebase."

Kelly indicated the rapidity with which new genome sequence information appears in public databases is presenting a growing challenge for the data storage, analysis, and utilization necessary to foster scientific and technological advances. The systems biology framework has arisen in response to this challenge, but new computing strategies are needed to take advantage of this new context for examining microbial biology.

Kelly also pointed out that most of what is now known about microbial biology was learned from the study of pure laboratory cultures. The "monoculture" paradigm has been quite productive and will continue to be at the heart of microbiology. However, monocultures are not representative of how microbial systems exist in nature. To this end, metagenomics has provided a means for examining microbial complexity, but complementary functional information is still needed to understand the "metaphenotype."

Illustrating the need for microbial community studies is the hypothesis that a significant portion of every microbial genome encodes elements designed to regulate and mediate intercellular interactions. These elements may not be responsive in laboratory monocultures and may be triggered only by certain environmental and ecological stimuli. Do these genomic elements exist? What are the studies needed to make this determination? If these genomic elements exist, how can they be identified, characterized, and manipulated? If multispecies systems are to be examined via systems biology, what are the consequences in terms of experimental design and analysis? What is the best way to construct a systems biology knowledgebase for multispecies (multiphenotype) investigations?

Over the next several years, efforts are needed to link the complexity reflected in metagenomes to what is already known from monoculture studies. Kelly indicated this learning curve will necessarily start with relatively simple systems because even co-cultures can exhibit phenotypes not easily predicted from pure culture information. Extending functional microbial genomics beyond monocultures was discussed with a view toward the integration of experimental design, experimental methods, and data analysis strategies. Kelly used hyperthermophile communities to illustrate some challenges that arise when moving beyond monocultures.

In his presentation, Arkin indicated the grand challenge to predict phenotype from genotype is particularly difficult in the microbial world. At its core, this challenge seeks to understand the principles of biological architecture and function sufficient for predicting behavior and, of course, for changing it. A systems biology knowledgebase should grow into an indispensible tool for molecular, environmental, evolutionary, medical, and epidemiological microbiologists

and for biotechnologists to understand and engineer their systems. However, there are challenges in accomplishing this that are found in few other systems.

Microbes rarely work alone but operate in complex communities that form spatial and temporal webs of mutual support, parasitism, and predation. Perhaps unique to microbes and their communities are the astonishingly rapid mechanisms for evolution and the deeply intertwined ecology of mobile genetic elements that aid in the preservation, diversification, and dissemination of function and may be central drivers themselves of the architecture of microbial networks.

The Knowledgebase, in the long term, will be faced with capturing and interrelating data about all these processes at scales from molecules to meters. Sequencing technologies reveal information on the identities of microbial players in these communities and can hone in on some aspects of gene expression. Structural techniques can provide key information on molecular identity and sometimes function. New imaging technologies can give us information on the arrangements and interactions among molecules, cells, and their environment. However, the complexity of the data increases greatly when moving beyond the sequence of single genomes and crystal structures of single proteins. The data also become far more conditional on unmeasured conditions and interactions and less precise and accurate metrologically, all of which present challenges for organizing and navigating this information. Arkin presented an example process outlining how such information could be assembled, navigated, and used in a knowledgebase. At each level, the challenges and acuteness of need for the community were described.

In ensuing discussions at the workshop, emphasis was placed on establishing agreed-upon scientific objectives that will result in a successful, community-driven Knowledgebase. To build a system that helps achieve important scientific goals, informatics experts need input from and frequent dialogue with the research community on what these goals are, including how the research technologies, data types and quantities, and goals change over time (see Fig. 1.1. Knowledgebase R&D Project).

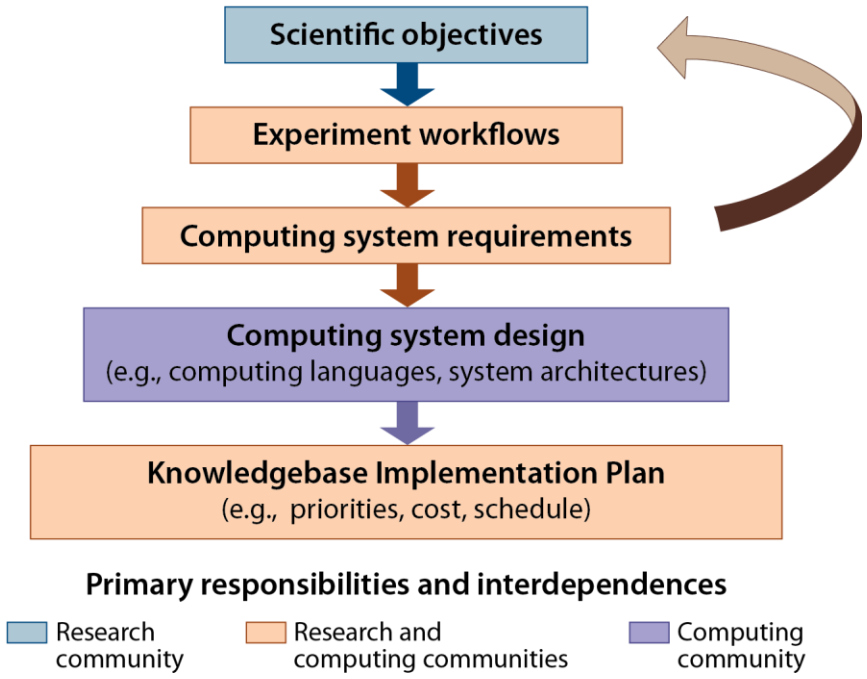


Fig. 1.1. Knowledgebase R&D Project: Scientific Objectives, Intense Collaboration Critical to Successful Knowledgebase Implementation Plan. The final product of this project, the Knowledgebase Implementation Plan, is being developed to incorporate the components and functionality necessary for the systems biology research community to meet its defined scientific objectives. To do this, the research and computing communities must work closely together to define—realistically and at a significant level of detail—the scientific objectives and experiment workflows (protocols) necessary for defining computing system requirements and design and for completing the implementation plan for a robust, durable Knowledgebase.

Workshops, such as this one, provide opportunities to discuss and identify appropriate and community-generated scientific objectives. Any and all input was welcomed, and participants were encouraged to contribute to the final R&D report at www.systemsbiologyknowledgebase.org. To be effective, scientific objectives must be credible, impactful, and achievable in a few years. Participants were asked to discuss objectives based on current research activities and consider candidates and priorities to recommend.

Several examples of potential scientific objectives related to microbes were presented to stimulate discussion. The first was improved prediction of gene regulatory networks based on integrating genomic sequences from phylogenetically related organisms with high-resolution expression (RNA-Seq) data from multiple biological states. Suppose the goal was to predict gene regulation in a particular situation. What are the Knowledgebase capabilities necessary for predicting gene regulation in a subsystem? One need would be the ability to upload raw RNA-Seq sequence data or provide access it. Another need would be tools to process raw sequence into standard formats. A third involves data visualization capabilities.

The limit in the future might be how many biological samples are available to be assayed by RNA-Seq and not the availability or cost of the technique. As cost rapidly declines, it is conceivable that thousands of states could be measured. From plots of expression profiles, genes that are statistically represented in a particular biological state can be readily visualized.

Which genes are in probable states or pathways that are of particular research interest? This should all be readily available to any researcher. Even small regulatory sequences are visible. Imagine doing RNA-Seq analysis on a set of phylogenetically related organisms and comparing them based on genomic structure. This new data will reinforce past experiments in these same organisms. Based on the alignments, we can find promoters and make predictions about gene regulatory binding sites. This illustrates the type of understanding achievable with a knowledgebase characterized by good algorithms and data integration technologies that have been built up over time. When using Google maps to find the nearest Starbucks location, users rely on a series of technologies that have been developed in such a way. A set of standards allows this information to be mapped together, enabling the system to generate the appropriate directions. The data integration underlying Google maps is analogous to many of the current challenges associated with integrating biological data.

A second example of a scientific objective would be integrating phenotypic response with specific genotypes or pathways so that regulatory or genetic changes could be predictably associated with microbial behavior and response. The idea of relating phenotypes to genotypes and putting that information in context is of wide interest. What are the sources of data? How do we transform them? What are the analytical steps, and what tools are currently available?

As with any scientific objective considered for the Knowledgebase, these two examples would be evaluated to determine if they are credible, impactful, and achievable. If a particular objective meets these three criteria, then community input would help set priorities for the development and implementation timeline of the Knowledgebase.

Section II: Workflows—Knowledgebase Use Cases

Workflows as a Bridge from Bench to Computer

The focus of this workshop, particularly on the second day, was on creating workflows. In research, a scientific objective is satisfied by creating hypotheses and doing one or more experiments depending on the scope of the objective. For every experiment, there are rationales, protocols to be executed, a number of data inputs (data sources) and outputs (results), and analysis tools. Workflows describe this information. Detailed workflows are bridges between the research and computing communities and thus are key to translating research into computing requirements that will most effectively advance the science.

Workflows provide important details for Knowledgebase design, both in terms of the underlying data as well as the experimental or analytical objective. Knowledgebase architecture will have layers such as data repositories, workflow management, and output visualization, all of which relate to workflows developed by the scientific community participating in this Knowledgebase development process. Workflows are essentially communication mechanisms that exchange ideas and information between the researchers and those who actually build the computing system. Included in this report are six workflows drafted to satisfy research objectives important in DOE systems biology. These workflows encompass diverse problem-solving methodologies representative of the broad scientific community and are works in progress—presented here to stimulate discussions between the research and bioinformatics

communities so that robust computing system requirements and an implementation plan can be developed.

Developing an executable Knowledgebase Implementation Plan must be a community effort—involving both the experimental and computing research communities—where we integrate across projects and research laboratories. Fully developed, robust workflows will foster this integration and lead to a more standardized approach. To handle a new level of biological complexity, we need to embrace more strategic software engineering approaches; we can no longer afford to create isolated and ad hoc systems.

As the key products of this workshop, workflows are critical inputs to the participants of the final workshop (June 1–2, 2010, Crystal City, Virginia). Prior to and during the final workshop, representatives from the computing and biological research communities will work closely together to refine the scientific objectives and workflows and to translate the workflows into computing system requirements. These requirements will form the basis of the Knowledgebase design—a prerequisite to the Knowledgebase Implementation Plan, which is the final product of the DOE Systems Biology Knowledgebase Research and Development Project.

The workflows described in this section are critical to the success of systems biology research and reflect the data inputs, outputs, and experiments being carried out in the DOE-sponsored research community. Over the next several months, assessments will be made to ensure that the highest priority workflows, as identified by community consensus, will be included in the Knowledgebase Implementation Plan. The workflows generated in this workshop are:

1. Metabolic Network Reconstruction (Ines Thiele)
2. Metabolic Flux Analysis via Isotope Labeling (Hector Garcia Martin)
3. Inference of Gene Regulatory Networks (Adam Arkin and Nitin Baliga)
4. Signaling (Aindrila Mukhopadhyay and Loren Hauser)
5. Structural Biology (Paul Adams)
6. Imaging Bioinformatics (Bahram Parvin)

To foster further interactions among the biology research communities, both experimental and computational, most of these have been included as originally submitted as a snapshot in time showing the current range of thought on what a workflow is and how the concept relates to various researchers and areas of research. [Note: An additional workflow on microbial community science is under development and will be available in May. This workflow is based on discussions from the Knowledgebase workshop held in conjunction with the DOE Joint Genome Institute's annual user meeting (March 23, 2010). Workflows associated with the microbial community scientific objectives will be discussed by the interdisciplinary participants at the June Knowledgebase workshop where Knowledgebase system requirements will be discussed and drafted.]

To facilitate workflow development, participants in this workshop were instructed to focus on describing several workflow components: data and sources (inputs), process steps (transformation rules or algorithms), and results or output. They also were asked to explain why

the workflow is important to the research endeavor and how it might be improved. As an example, consider the first workflow on Metabolic Network Reconstruction starting on p. 11. This example lists input data and even outlines how to obtain it. The process diagram associated with this workflow identifies each process step (see Figure 1. Detailed Workflow for Metabolic Network Reconstructions, p. 12). Many of these steps are common bioinformatic transformations that could readily be included in a future Knowledgebase. As the authors note, many steps are not precise and require some type of manual intervention such as curation. This identifies areas for improvement in either the underlying data or a need for better standards. Some of the process steps are experimental and produce specific results. Again, issues of data quality and accuracy can be important. Although not entirely automatable, this process is of wide utility and interest. This presents an excellent example of a workflow that the research community could prioritize to focus on in the Knowledgebase. By having a range of researchers focused on the bottlenecks, there would likely be improvements not only for metabolic reconstruction, but for other areas of research that depend on similar process steps.

Workflow 1: Metabolic Network Reconstruction

Summary

The metabolism workflow consists of two parts:

1. The metabolic network reconstruction protocol [1] and required data and
2. The protocol to obtain fluxomic data required by the metabolic network reconstruction protocol.

Genome-scale metabolic network reconstructions are biochemically, genetically, and genomically (BiGG) structured knowledgebases, the goal of which is to formally represent the metabolic activities of a specific organism. Genome-scale metabolic networks have been published for more than 30 organisms to date, though they are of varying quality and completeness. Reconstructions are useful because they can be mathematically converted into constraint-based models, allowing important predictive calculations like flux balance analysis to be performed. This comprehensive workflow details nearly 100 iterative steps in the following categories:

1. Draft reconstruction
2. Refinement of reconstruction
3. Conversion of reconstruction into computable formats
4. Network evaluation
5. Data assembly and dissemination

The output of this workflow is a highly curated, accurate, and comprehensive representation of biochemical transformation taking place in the organism of interest. It is not yet possible to automate all steps within the process without loss of accuracy or correctness.

We also attached the comprehensive standard operating procedure (SOP) for biochemical network reconstruction [1] to this workflow.

Input

Required organism-specific data

- Gene information (ID, coordinates, function)
- Protein information (function, location, complex formation)
- Enzymatic reaction (stoichiometry at cellular pH, substrate specificity, cofactor specificity, location, directionality)
- Biomass composition (fraction of macromolecule, molecular composition of macromolecules)
- Phenotyping data (growth medium composition, other growth conditions – e.g. temperature, pH, etc)
- Knock-out strain information (growth phenotypes, other characteristics)

- P/O ratio

How to obtain this information

Online resources:

- Genome database containing genome annotation (i.e., locus ID, gene coordinates, (putative) annotation) – e.g., GOLD, TIGR, SEED, etc.
- Biochemical reaction database for metabolic reactions – e.g., KEGG, BRENDA
- Transport database for transport reaction mechanisms – e.g., Transport DB
- Organism-specific databases – e.g., EcoCy, PyloriGene, GeneCards
- Protein location prediction – e.g., PSORT, PA-SUB
- Thermodynamic information (estimation of standard Gibbs free energy of formation ($\Delta_f G'°$) and of reaction ($\Delta_r G'°$)) – e.g., Web GCM
- CMR database (estimation of DNA, RNA and protein composition)

Tools:

- Blast (if not or insufficient genome annotation is available), other gene function annotation tools

Bibliome:

- Primary and review literature about organism, its metabolic characteristics and its components (proteins, genes)
- Biochemical textbooks
- Organism-specific books

Experiments:

- Measurement of biomass composition (lipids, amino acids, nucleotides, cofactors, etc.)
- Measurement of growth environments (e.g., biolog)
- Measurement of single and double knockout mutants
- Measurement of possible secretion products (and ratios) at different growth environments
- Omics data: Metabolomics, fluxomics, proteomics, transcriptomics
- Transcriptional regulatory information – which pathways are active under which conditions

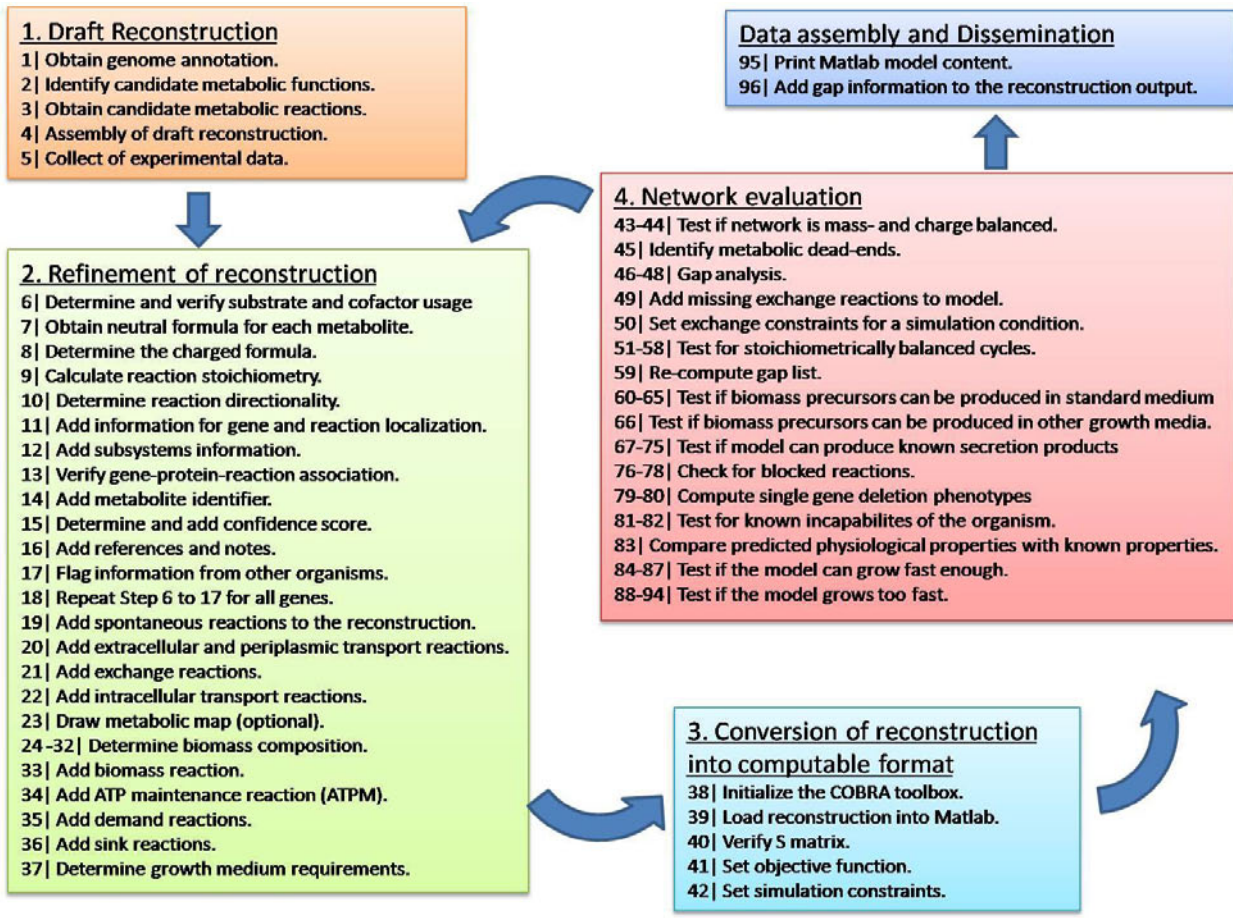


Figure 1. Detailed Workflow for Metabolic Network Reconstructions.

Workflow Process to Metabolic Network Reconstruction

The biochemical network reconstruction process is well established for metabolism and has been applied to many model organisms. The same approach can also be applied for other cellular functions, such as signaling [2, 3] and macromolecular synthesis [4]. The reconstruction process has been reviewed by numerous groups [5-8]. More recently, it has been formulated in the form of a standard operating procedure (SOP), or protocol, which explains the necessary stages and steps in great details [1]. Readers interested in reconstruction are advised to also refer to the SOP.

The metabolic reconstruction process can be grouped into 5 major stages (see Figure 1):

1. **Generation of a draft reconstruction based on genome annotation and biochemical databases.** Generally, the genome annotation is downloaded from a repository (e.g., NCBI) or the sequencing center (e.g., TIGR), and it should list at least a unique identifier, genome coordinates, and potential gene product function. Many of genome resources have also enzyme commission (EC) numbers for the genome encoded enzymes. These EC numbers along with key words can be used to compile a sublist of potential metabolic functions in the target organism. This list can be then used to obtain from

biochemical databases (e.g., KEGG [9]. BRENDA [10]) the metabolic reactions catalyzed out by the enzymes. This list represents the draft reconstruction. The characteristics of this draft reconstruction are that it is incomplete (missing or wrong annotations) and it has an organism-independent reaction list: KEGG, as well as partially BRENDA, list all possible metabolic transformation catalyzed by a particular enzyme. However, the enzyme of the target organism may be able to bind to a subset of the listed substrates, or only one of the listed coenzymes can participate in the reaction in the target organism. This substrate and coenzyme pluripotency is one of the main reasons why manual curation is necessary.

2. **Refinement and expansion of the draft reconstruction through manual curation and extensive use of biochemical literature specific for the target organism.** Starting from the draft reconstruction, every entry will be evaluated for the following criteria:

   a. Is the assigned function of the gene product correct? Use of biochemical literature, enzyme purification studies, a more detailed, phylogeny based annotation are helpful to answer this question.

   b. What is the substrate and coenzyme specificity of the target organism's enzyme? Use of biochemical data, enzyme assays and protein structure will be helpful for answering this question. Finding evidence for this issue can be difficult. The use of closed relative organisms can be helpful.

   c. Is the biochemical reaction(s) mass- and charge balanced? Therefore, the neutral formula of each metabolite in the reaction has to be obtained (e.g., from KEGG or PubChem [11]). The charged formula has to be determined for each metabolite for a set pH value (e.g., pH 7.2) by determining the protonation state of each functional group within the metabolite. Software tools are available to assist this step (see Thiele and Palsson for details [1]). Once the charged formula has been determined for each metabolite, the occurrences of each element (e.g., C, H, N, S, O, P), as well as the charge, on the left- and right-hand side has to be counted. Stoichiometric coefficients may need to be adjusted such that the same amount of each element appears on both sides of the reaction. In some cases, protons (H+) or water may be added to the reactions to obtain a mass- and charge balanced reaction.

   d. The reaction directionality needs to be determined using thermodynamic information (refer for details to Thiele and Palsson [1], Feist et al [12], and Fleming et al [13]).

   e. Localization of reaction needs to be determined, especially, if multiple compartments are considered (e.g., human metabolic network accounts for eight cellular compartments, while many bacterial reconstructions account for two or three compartments, which are extracellular space, periplasm, and cytosol). Information about reaction location may be obtained from the genome sequence if it encodes for a signal peptide (for protein export) or by targeted experiments (e.g., using GFP tagging and fluorescence microscopy).

f. Gene-protein-reaction (GPR) association needs to be determined: while the genome annotation indicates that the gene product has a particular function, one should investigate if further gene products are needed for function, as is the case for protein complexes, or if alternate gene products exist that can carry out the function, i.e., isozymes. The reconstruction contains these GPR associations in form of Boolean rules: for example, a protein complex is encoded as 'gene_A & gene_B', while isozymes are encoded as 'gene_A or gene_B'. Any combination of these rules is possible. Beside genome annotation, biochemical data, protein purification, and/or structural genomics can provide information regarding the GPR association.

g. Confidence score, references, and notes: The steps listed above collect valuable information for a particular enzyme or function in the target organism. This information should be associated with the network reaction (e.g., in special columns in the spreadsheet). This information is thought to increase the traceability of reaction/gene evidence as well as highlight/summarize the amount of knowledge currently available. Often, a confidence scoring system is employed, which allows easy identification of high-confidence/low confidence reactions in the network. This is of particular value during the network debugging and evaluation stage (see below). The highest confidence score (4) is given to reactions that have biochemical evidence (e.g., protein purification, protein assays, protein structure information). A score of 3 is given if genetic data is available (e.g., knock-out mutant characterization, knock-in experiments, over-expression of a protein). A score of 2 is given if either physiological data (e.g., secretion products, growth capability on substrate) or (high confidence) sequence annotation is available. A low confidence score of 1 is given if reactions are included for modeling purposes without any of the aforementioned evidence. In some cases, a confidence score of zero is also employed, which highlights reactions that have not yet been evaluated for supporting evidence.

h. Finally, different information should be collected in this stage of the reconstruction process to facilitate the following stages. This information includes the biomass precursors, necessary to produce a new cell (target organism) which is ideally derived from experimental data (see Thiele and Palsson for a detailed description on how to compile this information). Furthermore, information about enzyme reaction rates (v_{max}) should be collected, as many biochemical publications contain this information. Information about growth media should be also collected.

3. **Conversion of the manual curated metabolic reconstruction into a mathematical model.** The reconstruction process is an iterative process as shown in Figure 1, where the initial reconstruction is converted into a mathematical format, the so called stoichiometric (S) matrix. This model conversion also includes the addition of balances and bounds. Balances in biochemical networks can be, for example, mass- and energy conservation. For instance, the majority of modeling applications of metabolic models assume the system to be in quasi steady state. This assumption implies that the sum of producing reactions for a particular metabolite is equal to the sum of consuming reactions. Bounds on metabolic reactions can include maximal reaction rates based on

the catalyzing enzyme's properties, thermodynamic information (e.g., reaction directionalities), etc. Often, the mathematical models are stored and computed in Matlab (Mathwork, Inc). Commonly, metabolic reconstructions and models are stored in the systems biology markup-language (SBML) format [14], which is platform independent and can be loaded in numerous systems biology applications.

4. **Network debugging and evaluation to ensure that the metabolic model has similar phenotypic properties as the target organism.** Once the metabolic reconstruction is converted into a mathematical format and balances and bounds are applied, a comprehensive investigation of the model's properties begins. Most reconstructions contain initially numerous dead-end metabolites (i.e., metabolites that are only produced or consumed in the network). Due to the balance constraints, reactions which contain such dead-end metabolites cannot carry any reaction flux in any simulation conditions. A detailed evaluation of these dead-end metabolites is necessary to identify whether these metabolites can be connected to the remaining network by adding one or more reactions to the reconstruction. However, one has to be careful, as arbitrary filling of the so-called gaps will alter significantly the model's properties. All added reactions should have experimental, genome and/or physiological data as supporting evidence. Some dead-end metabolites may remain in the network, as current knowledge does not support any filling of gaps they are causing. In addition to these 'knowledge gaps' the reconstruction can contain 'scope gaps.' In the case of scope gaps, reactions are known, which could connect the dead-end metabolite, but they are either non-metabolic or not within a previously defined scope of the reconstruction (e.g., tRNA charging with amino acids).

   Once all dead-end metabolites have been characterized and partially connected to the network by repeating part of the second and third stage, the model's capability to produce biomass precursor is evaluated. This process will lead to further identification of network gaps, which need to be filled. This step can be quite time-consuming, and detailed evaluation of dead-end metabolites in the earlier step will directly pay off. When the model can produce all biomass precursors, one can compile them into one reaction (the biomass reaction) by considering their fractional contributions to cell composition. This stage also includes further (i) quality tests, such as the model's capability to grow on known carbon, nitrogen, phosphor and sulfur sources; (ii) the capability to reproduce accurately measured growth rates and to secrete known by-products. The list of tests depends on the properties of the target organism as well as the availability of experimental data (e.g., phenotyping data, knock-out mutant growth phenotype data, etc.). Note that this stage is iterative, in which network reactions will be added (by repeating partially, or in full, stage 2 and 3) or in some cases reactions will be removed from the metabolic reconstruction. This stage is deemed to be finished if the model reproduces accurately the target organism's phenotypic characteristics and/or experimental data is exhausted.

5. **Prospective use of the reconstruction and the metabolic models. This stage is certainly the most exciting part of the reconstruction process.** Numerous applications have been developed over last decade or so, including biological discovery [15], metabolic

engineering [16-17], prediction of outcome of adaptive evolution [18], network topology [19], and assessment of phenotypic behavior [20-22]. Some of these applications have been summarized in a recent review [23-24].

Output

The output of this workflow is a highly curated, accurate and comprehensive representation of biochemical transformation taking place in the organism of interest (Figure 2). Note that to date, it is not possible to automate all steps within the 5 stages without loss of accuracy or correctness.

A

Col 1	*Col 2*	*Col 3*	*Col 4*	*Col 5*	*Col 6*	*Col 7*
Rxn name	**Rxn description**	**Formula**	**GPR**	**Genes**	**Subsystem**	**Reversible**
PFK	phosphofructokinase	atp[c] + f6p[c] -> adp[c] + fdp[c] + h[c]	(b3916 or b1723)	b1723 b3916	Glycolysis/Gluconeogenesis	0

Col 8	*Col 9*	*Col 10*	*Col 11*	*Col 12*	*Col 13*	*Col 14*	*Col 15*
LB	**UB**	**Objective**	**CS**	**E.C. number**	**rxnKeggID**	**Notes**	**References**
0	1000	0	4	2.7.1.11	R00756	E. coli has to genes for PFK (pfkA and pfkB) where pfkA is the major form.	PMID: 6310120;149128;6310120

B

Col 1	*Col 2*	*Col 3*	*Col 4*	*Col 5*	*Col 6*	*Col 7*	*Col 8*	*Col 9*	*Col 10*
Metabolite name	**Description**	**Formula**	**Charge**	**Compartment**	**KeggID**	**PubChemID**	**CheBI ID**	**Smiles**	**InChi**
glc-D[c]	D-Glucose	C6H12O6	0	Cytoplasm	C00031	3333	17634	OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O	1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6?/m1/s1

Figure 13:
Figure 2. Data contained in the metabolic network reconstruction after completion of the presented workflow.

References

1. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction.** *Nature protocols* 2010, **5**(1):93-121.
2. Papin JA, Palsson BO: **The JAK-STAT Signaling Network in the Human B-Cell: An Extreme Signaling Pathway Analysis**. *Biophysical journal* 2004, **87**(1):37-46.
3. Li F, Thiele I, Jamshidi N, Palsson BO: **Identification of potential pathway mediation targets in Toll-like receptor signaling**. *PLoS Comput Biol* 2009, **5**(2):e1000292.

4. Thiele I, Jamshidi N, Fleming RM, Palsson BO: **Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization**. *PLoS Comput Biol* 2009, **5**(3):e1000312.
5. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO: **Reconstruction of biochemical networks in microorganisms**. *Nature reviews* 2009, **7**(2):129-143.
6. Reed JL, Famili I, Thiele I, Palsson BO: **Towards multidimensional genome annotation**. *Nature reviews* 2006, **7**(2):130-141.
7. Notebaart RA, van Enckevort FH, Francke C, Siezen RJ, Teusink B: **Accelerating the reconstruction of genome-scale metabolic networks**. *BMC Bioinformatics* 2006, **7**(1):296.
8. Durot M, Bourguignon PY, Schachter V: **Genome-scale models of bacterial metabolism: reconstruction and applications**. *FEMS microbiology reviews* 2009, **33**(1):164-190.
9. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic Acids Res* 2006, **34**(Database issue):D354-357.
10. Barthelmes J, Ebeling C, Chang A, Schomburg I, Schomburg D: **BRENDA, AMENDA and FRENDA: the enzyme information system in 2007**. *Nucleic Acids Res* 2007, **35**(Database issue):D511-514.
11. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2008, **36**(Database issue):D13-21.
12. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Molecular systems biology* 2007, **3**:121.
13. Fleming RM, Thiele I, Nasheuer HP: **Quantitative assignment of reaction directionality in constraint-based models of metabolism: Application to Escherichia coli**. *Biophys Chem* 2009.
14. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A *et al*: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics (Oxford, England)* 2003, **19**(4):524-531.
15. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO: **Systems approach to refining genome annotation.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(46):17480-17484.
16. Lee SY, Kim JM, Song H, Lee JW, Kim TY, Jang YS: **From genome sequence to integrated bioprocess for succinic acid production by Mannheimia succiniciproducens**. *Appl Microbiol Biotechnol* 2008, **79**(1):11-22.

17. Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO: ***In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid**. *Biotechnology and bioengineering* 2005, **91**(5):643-648.

18. Ibarra RU, Edwards JS, Palsson BO: ***Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth**. *Nature* 2002, **420**(6912):186-189.

19. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL: **Global organization of metabolic fluxes in the bacterium *Escherichia coli***. *Nature* 2004, **427**(6977):839-843.

20. Thiele I, Price ND, Vo TD, Palsson BO: **Candidate metabolic network states in human mitochondria: Impact of diabetes, ischemia, and diet**. *J Biol Chem* 2005, **280**(12):11683-11695.

21. Reed JL, Palsson BO: **Genome-Scale In Silico Models of E. coli Have Multiple Equivalent Phenotypic States: Assessment of Correlated Reaction Subsets That Comprise Network States**. *Genome Res* 2004, **14**(9):1797-1805.

22. Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ, Smid EJ: **Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model**. *J Biol Chem* 2006, **281**(52):40041-40048.

23. Feist AM, Palsson BO: **The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli***. *Nat Biotech* 2008, **26**(6):659-667.

24. Oberhardt MA, Palsson BO, Papin JA: **Applications of genome-scale metabolic reconstructions**. *Molecular systems biology* 2009, **5**:320.

Workflow 2: Metabolic Flux Analysis via Isotope Labeling

Summary: Metabolic fluxes are a key determinant of cellular physiology, representing the final functional output of the interaction of all the molecular machinery (genes, proteins, metabolites) studied by the other "omics" fields. This workflow (a schematic of which is presented in Figure 1) describes the input data required for measuring fluxes using an isotope labeled feed, along with the expected output and the processes needed to obtain it. The main input data are metabolite labeling patterns after a carbon labeling experiment, a metabolic reconstruction, and measured extracellular and biomass fluxes. The desired output is the rate (i.e., number of molecules through the reaction) for each of the reactions considered in the model, along with confidence intervals. Here, we will focus on the most common and well-established form of flux analysis through isotope labeling: ^{13}C Metabolic Flux Analysis (^{13}C MFA) from proteogenic amino acids in the exponential phase. Nonetheless, the modular nature of the workflow presented here will allow for other varieties of ^{13}C MFA in development to be easily incorporated.

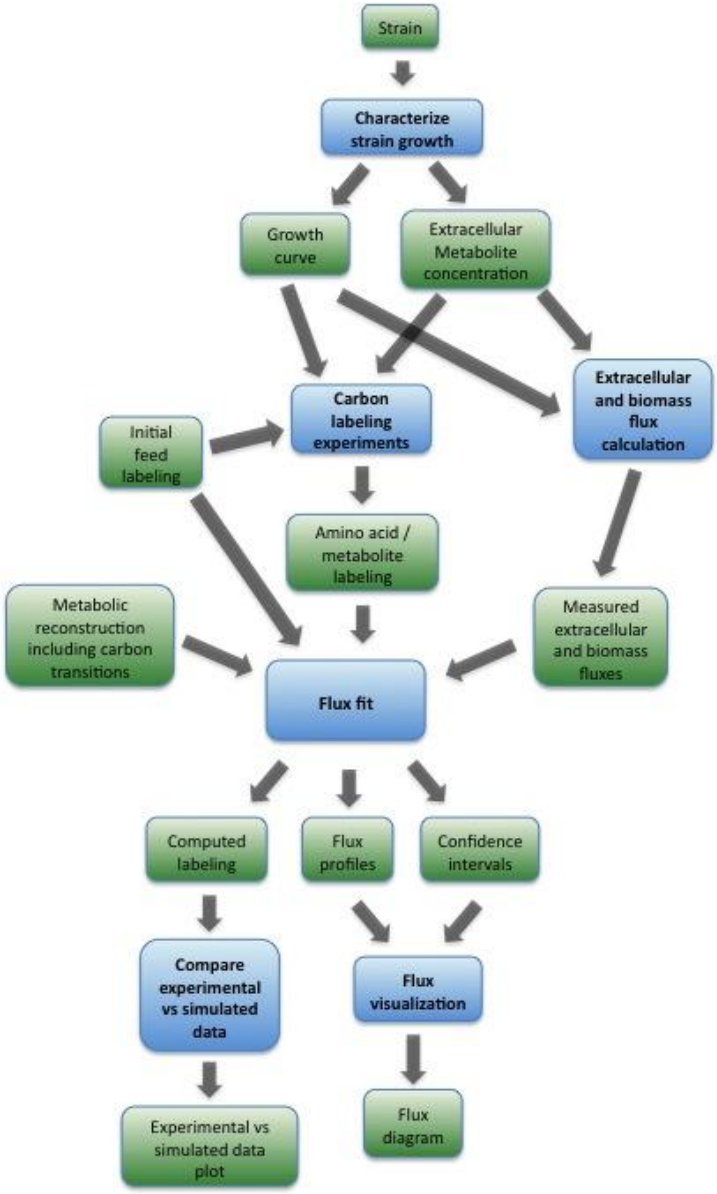


Figure 1. Workflow for ^{13}C metabolic flux analysis. Blue blocks indicate processes (e.g., experiments or algorithms), and the green blocks indicate datasets or physical objects. See text on next page for callout to this figure.

The following workflow for metabolic flux analysis though isotope labeling will focus on its most common and established form: 13C Metabolic Flux Analysis (13C MFA) from proteogenic amino acids in the exponential phase. This is not to say that it is the most important, but rather the most mature and where agreement on a common workflow is most likely. That having been said, the modular nature of the workflow presented here allows for other varieties to be easily incorporated, some of which are still in development. For example, if intracellular metabolite labeling were to be used instead of amino acid labeling, this data (and the necessary metabolite concentrations) could be easily inserted at the same point in the diagram as amino acid labeling

(along with a connection to the metabolomics workflow). Other such variations such as flux analysis in a non-steady state [1], labeling of atoms other than carbon [2], or usage of NMR data [3] can be added in a similar fashion.

The workflow described here is an important tool for us as researchers for several reasons: 1) it explains the process to new members of the group as well as collaborators, 2) it helps define the standards for stored data in order to replicate and compare results in the future 3) it defines the steps used to track project completion and to help plan and develop high-throughput experiments.

The first step we include in the workflow is the characterization of the strain growth, a process not exclusive to 13C MFA. This characterization produces two sets of data that will be useful for planning the isotope labeled experiment: the growth curve and the concentration of extracellular metabolites. The growth curve provides the mid-log point used for sampling, and the extracellular metabolite concentration provides a rough idea of which metabolic pathways are important in addition to measured transport fluxes for later use. An example of a possible data input of extracellular metabolite concentration is shown in Figure 2. Useful metadata involves strain details, including plasmid and genetic modifications, along with materials and methods for OD and extracellular metabolite measurements.

Compound Summary

Sequence start: 2/20/2009 11:22:46 AM
Operator:

Method file name: C:\CHEM32\1\DATA\FEB-20-2009 2009-02-20 11-22-22\YEAST_FERMENTATION_JENNIFER.M

Sample Name	Sample Amt [g/L]	Multip.* Dilution	FileName .D	RetTime [min]	Amount [g/L]	Compound
Blank	0.00000	1.0000	001-0801	9.222	-	-
				9.325	-	-
				9.876	-	-
				12.810	-	-
				13.083	-	-
				13.769	-	-
				14.072	-	-
				14.345	-	-
				15.418	-	-
				15.693	-	-
				22.392	-	-
1-1	0.00000	1.0000	011-0901	9.217	0.032	Pyruvate
				9.230	16.379	Glucose
				9.876	-	-
				12.809	5.33320e-1	Lactate
				13.085	5.23185e-1	Lactate
				13.777	9.03251e-3	Glycerol
				14.087	6.99238e-2	Formate
				14.344	6.28119e-2	Formate
				15.423	3.43505e-1	Acetate
				15.701	3.34890e-1	Acetate
				22.410	1.87606e-1	EtOH
1-2	0.00000	1.0000	012-1001	9.215	0.027	Pyruvate
				9.229	16.774	Glucose
				9.876	-	-
				12.807	4.39012e-1	Lactate
				13.083	4.06618e-1	Lactate
				13.769	-	-
				14.087	5.33141e-2	Formate
				14.343	3.99210e-2	Formate
				15.420	2.81520e-1	Acetate
				15.699	2.65478e-1	Acetate
				22.408	1.16859e-1	EtOH
1-3	0.00000	1.0000	013-1101	9.215	0.030	Pyruvate
				9.225	16.606	Glucose
				9.876	-	-
				12.807	4.11458e-1	Lactate
				13.079	3.89868e-1	Lactate
				13.769	-	-
				14.087	5.74127e-2	Formate
				14.342	4.45182e-2	Formate
				15.419	3.00310e-1	Acetate
				15.694	2.99122e-1	Acetate
				22.401	1.36531e-1	EtOH
2-1	0.00000	1.0000	014-1201	9.215	0.028	Pyruvate
				9.226	16.606	Glucose

Figure 2. Example of output of HP-LC analysis used as input of extracellular metabolite concentration. A standard format for this information would be useful.

The main experimental process in the workflow is the performance of the labeling experiment, the workflow for which has been described by Zamboni et al [4] (see Figure 3). The necessary input for planning and performing the experiment includes the growth curve and extracellular metabolite concentrations, which has been discussed above, and the feed labeling, which affects the range of fluxes that can reliably be determined [5] [6]. The output includes the main piece of data needed to constrain the metabolic fluxes: the amino acid labeling pattern. The labeling information should include as metadata details of the experiment including sampling points, initial feed labeling and materials, and methods for labeling measurement. Examples of amino acid labeling data in terms of the derivatized fragments [7] or amino acid backbone labeling can be seen in Figs. 4 and 5 [8] [9].

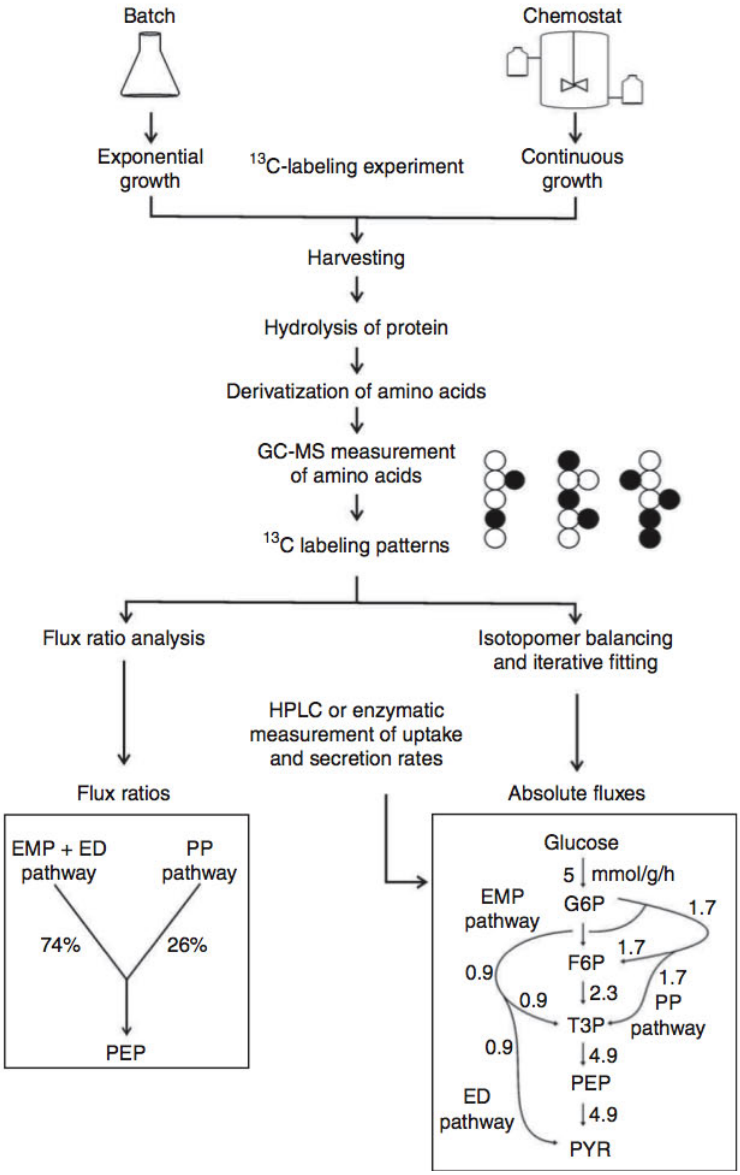


Figure 3. Workflow for carbon labeling experiment as per Zamboni et al [3] showing the two types of methods to obtain flux profiles: through flux ratio analysis or isotopomer balancing and iterative fitting.

SAMPLE:	WT+MET	WT+MET	WT+MET	MET28	MET28	CBF1	CBF1
REPLICATE	A	B	C	A	B	A	B
TBDMS FRAGMENT							
Ala232 (M0)	0.5666	0.5638	0.5633	0.5757	0.5755	0.5632	0.5699
Ala233 (M1)	0.3151	0.3174	0.3170	0.3102	0.3135	0.3173	0.3133
Ala234 (M2)	0.0907	0.0912	0.0916	0.0876	0.0857	0.0912	0.0898
Ala235 (M3)	0.0238	0.0235	0.0240	0.0227	0.0214	0.0242	0.0234
Ala236 (M4)	0.0034	0.0035	0.0036	0.0035	0.0037	0.0035	0.0032
Ala237 (M5)	0.0004	0.0005	0.0005	0.0003	0.0002	0.0005	0.0004
Ala260 (M0)	0.5603	0.5570	0.5542	0.5678	0.5626	0.5582	0.5595
Ala261 (M1)	0.3171	0.3192	0.3197	0.3115	0.3155	0.3174	0.3161
Ala262 (M2)	0.0930	0.0934	0.0944	0.0913	0.0898	0.0941	0.0939
Ala263 (M3)	0.0250	0.0255	0.0269	0.0247	0.0269	0.0258	0.0258
Ala264 (M4)	0.0040	0.0040	0.0040	0.0041	0.0045	0.0040	0.0041
Ala265 (M5)	0.0007	0.0008	0.0007	0.0007	0.0006	0.0005	0.0006
Ala266 (M6)	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Gly218 (M0)	0.7614	0.7592	0.7579	0.7612	0.7621	0.7585	0.7605
Gly219 (M1)	0.1603	0.1605	0.1613	0.1589	0.1575	0.1606	0.1596
Gly220 (M2)	0.0664	0.0685	0.0682	0.0664	0.0643	0.0687	0.0671
Gly221 (M3)	0.0098	0.0100	0.0104	0.0109	0.0128	0.0102	0.0107
Gly222 (M4)	0.0019	0.0017	0.0019	0.0021	0.0025	0.0019	0.0020
Gly223 (M5)	0.0002	0.0001	0.0003	0.0005	0.0008	0.0001	0.0002
Gly246 (M0)	0.7479	0.7491	0.7479	0.7529	0.7581	0.7511	0.7529
Gly247 (M1)	0.1694	0.1683	0.1696	0.1673	0.1640	0.1667	0.1664
Gly248 (M2)	0.0706	0.0710	0.0708	0.0687	0.0667	0.0703	0.0693
Gly249 (M3)	0.0103	0.0099	0.0100	0.0093	0.0094	0.0103	0.0096
Gly250 (M4)	0.0018	0.0016	0.0018	0.0018	0.0019	0.0017	0.0017
Val260 (M0)	0.4119	0.4087	0.4130	0.4269	0.4181	0.5802	0.6023
Val261 (M1)	0.3806	0.3828	0.3795	0.3737	0.3797	0.2760	0.2637
Val262 (M2)	0.1532	0.1531	0.1525	0.1482	0.1475	0.1105	0.1039
Val263 (M3)	0.0451	0.0447	0.0436	0.0430	0.0447	0.0273	0.0248
Val264 (M4)	0.0086	0.0094	0.0098	0.0082	0.0099	0.0056	0.0046
Val265 (M5)	0.0006	0.0013	0.0015	0.0000	0.0000	0.0003	0.0006
Val288 (M0)	0.4098	0.4054	0.4083	0.4247	0.4139	0.5779	0.5984
Val289 (M1)	0.3812	0.3839	0.3802	0.3729	0.3776	0.2771	0.2638
Val290 (M2)	0.1545	0.1552	0.1559	0.1497	0.1550	0.1120	0.1062
Val291 (M3)	0.0440	0.0448	0.0448	0.0428	0.0441	0.0272	0.0255
Val292 (M4)	0.0091	0.0093	0.0095	0.0087	0.0088	0.0052	0.0055
Val293 (M5)	0.0013	0.0014	0.0013	0.0012	0.0005	0.0007	0.0007
Leu274 (M0)	0.3110	0.3088	0.3100	0.3312	0.3190	0.7240	0.7293
Leu275 (M1)	0.3885	0.3898	0.3890	0.3845	0.3895	0.1899	0.1854
Leu276 (M2)	0.2111	0.2114	0.2100	0.2035	0.2063	0.0748	0.0723
Leu277 (M3)	0.0689	0.0688	0.0699	0.0624	0.0662	0.0098	0.0113
Leu278 (M4)	0.0170	0.0175	0.0174	0.0157	0.0161	0.0015	0.0017
Leu279 (M5)	0.0031	0.0032	0.0032	0.0025	0.0029	0.0000	0.0000
Leu280 (M6)	0.0005	0.0005	0.0004	0.0002	0.0000	0.0000	0.0000

Figure 4. Amino acid labeling for different derivatized fragments, taken from [7]. The name on the left column corresponds to the amino acid and the fragment type [6]. Each of the following columns corresponds to the fraction of molecules with 0,1,2... extra mass units incorporated due to isotopic variation (from carbon or other atoms).

Table 1. Relative Intensity and Error Associated with the Measurement of Each Amino Acid Isotopomer[a]

amino acid	rel intens (%)	error (ppm)	amino acid	rel intens (%)	error (ppm)	amino acid	rel intens (%)	error (ppm)	amino acid	rel intens (%)	error (ppm)
Gly			Val			Lys			Phe		
M0	1.30	0.21	M0	11.60	0.07	M0	0.90	0.15	M0	0.65	0.17
M1	4.20	0.22	M1	42.40	0.01	M1	4.20	0.09	M1	2.60	0.21
			M2	1.80	0.00	M2	8.00	0.10	M2	5.60	0.20
						M3	0.30	0.09	M3	3.70	0.21
Ala			Thr			Glu			Arg		
M0	4.70	0.00	M0	1.20	0.02	M0	8.30	0.16	M0	0.30	0.27
M1	19.50	0.01	M1	6.00	0.02	M1	23.30	0.08	M1	1.00	0.27
M2	0.40	0.00	M2	11.50	0.03	M2	1.00	0.09	M2	1.80	0.37
			M3	0.20	0.03				M3	0.07	0.21
Ser			Leu (and Ile)			Met			Tyr		
M0	2.30	0.03	M0	100	0.06	M0	0.15	0.17	M0	nd[b]	
M1	7.30	0.05	M1	71.70	0.01	M1	1.00	0.18	M1	0.10	0.25
M2	0.20	0.00	M2	4.20	0.01	M2	4.40	0.14	M2	0.20	0.42
			M3	0.10	0.04	M4	8.20	0.13	M3	0.15	0.15
Pro			Asp			His					
M0	19.00	0.03	M0	1.45	0.01	M0	2.40	0.12			
M1	55.20	0.02	M1	7.30	0.03	M1	3.80	0.12			
M2	2.50	0.02	M2	15.10	0.03	M2	0.90	0.09			
M3	0.06	0.40	M3	0.30	0.07	M4	0.30	0.08			

[a] M0, M1, M2, etc., refers to isotopomers with 0, 1, 2, etc., ^{13}C incorporated in the backbone of the amino acid. A RSD of ≤2% is associated with each relative inensity measurement. Errors refer to one single measurement. A variation of 10% is associated with it. [b] Not detected.

Figure 5. Amino acid labeling for carbon backbone. M0, M1, M2... indicate the fraction of molecules with 0,1,2.... labeled carbons incorporated [8].

Extracellular metabolite concentrations from the growth characterization experiment are used to derive the transport fluxes, (i.e. uptake and secretion rates). The calculation from extracellular metabolites is straightforward, and it involves calculating the change in metabolite concentration in the media. Another important set of known fluxes is the fluxes to biomass production, obtained from the change in OD and the cell composition.

Fitting the fluxes to the labeling data is the main computational process in the workflow. A variety of methods are available to do this [7] [10] [11]. Some involve determination of local flux ratios, and some are based on iterative fittings for the whole metabolic network under consideration (see Figure 3). Among the latter, the fit can either be performed in a search space involving fluxes and labeling, with the labeling pattern included as a constraint [12], or in a search space involving only fluxes, with the labeling determined for each flux profile. Labeling corresponding to each flux profile can be produced using several methods, including isotopomer mapping matrices [13], cumomers [14] or elementary metabolic units [15], to name a few. The search through the flux phase space can, as well, be carried over via a variety of techniques, including genetic algorithms, sequential quadratic programming and simulated annealing [7]. Software for flux calculations include 13CFLUX [16] [4] and FIATFLUX [17], none of which are available in open source format, and openFLUX [18], a recent application based on elementary metabolic units available in open source format. For the purpose of designing a workflow, what is important is not the differences among these methods but the fact that they all require the same input: 1) transport and biomass fluxes, 2) amino acid labeling, 3) initial feed labeling, and 4) the carbon transitions included in a metabolic reconstruction. Amino acid and initial labeling patterns, and measured fluxes have all been discussed above. The metabolic reconstruction has been discussed at length above; the only required condition is that it includes atomic transitions (see example in Figure 6 [19]). This metabolic model may be a coarse grained version of the models considered above. See, for example, Figs. 6 and 7, where

some reactions have been clumped together to ease calculations. Recently, a new tool for sharing, storing, and constructing these atomic transitions embedded in a metabolic reconstruction has become available [20]. Standard formats used in this program are the 13CFLUX format and SBML.

Supplementary Table S-2 Reaction lists:

```
% Lactate to pyruvate

1: ABC(1) -> ABC(2)

% Pyruvate to Acetyl-CoA

2: ABC(2) -> BC(3) + A(17)

% Oxaloacetate to citrate

3: AB(3) + abcd(11) -> dcbBAa(5)

% Citrate to isocitrate

4: ABCDEF(5) -> ABCDEF(6)

% Isocitrate to 2-oxoglutarate

5: ABCDEF(6) -> ABCDE(7) + F(17)

% 2-oxoglutarate to succinyl-CoA

6: ABCDE(7) -> BCDE(8) + A(17) :: ABCDE(7) -> EDCB(8) + A(17)

% Succinate to Malate

7: ABCD(8) -> ABCD(9)

% Fumarate -> malate

8: ABCD(9) -> ABCD(10)
```

Figure 6. Example of atomic transitions input needed form 13C MFA [18]. The first number indicates the reaction number as per Figure 7, and the numbers in parenthesis indicate the metabolite numbers. Carbon transitions are indicated as strings of letters: e.g., ABC -> AB + C indicates that the first two carbon in the reactant end up as the two carbons in the first product and the last carbon goes to the second product.

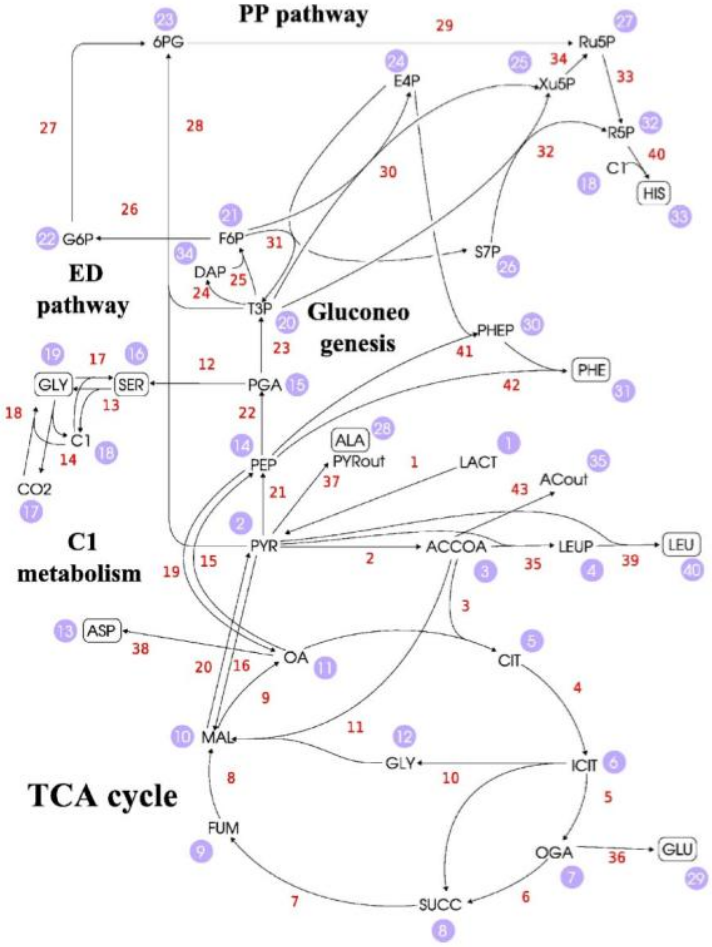


Figure 7. Reaction network for *Shewanella* central carbon metabolism [18]. Notice how some of the reactions (e.g., g6p to Ru5P in the penthose-phosphate pathway) have been clumped together to ease calculations).

The output should include the flux profile giving the best fit for the experimental data, a confidence interval, and the computed labeling patterns. The metabolic flux profile gives the best guess (compatible with the data) of the rate for each of the metabolic reactions considered in the metabolic model of the cell. This information is useful *per se* as a widely recognized highly relevant characteristic of the phenotype [21], and has numerous applications in (e.g.) metabolic engineering [11]. As with every experimental measurement, it is also desirable to assign confidence intervals to flux estimates, and a variety of algorithms are available for this purpose [7].

A simple list of Fluxes with their corresponding confidence intervals for each metabolic reaction can be very difficult to make productive use of, particularly for large models. Hence, visualization is an important part of the workflow and several possibilities are available [22] [23] [24] [25], although not all of them allow flux visualization for models with clumped reactions.

Finally, a useful visual check that the fit is appropriate is to compare computational predictions with experimental data, as shown in Figure 8.

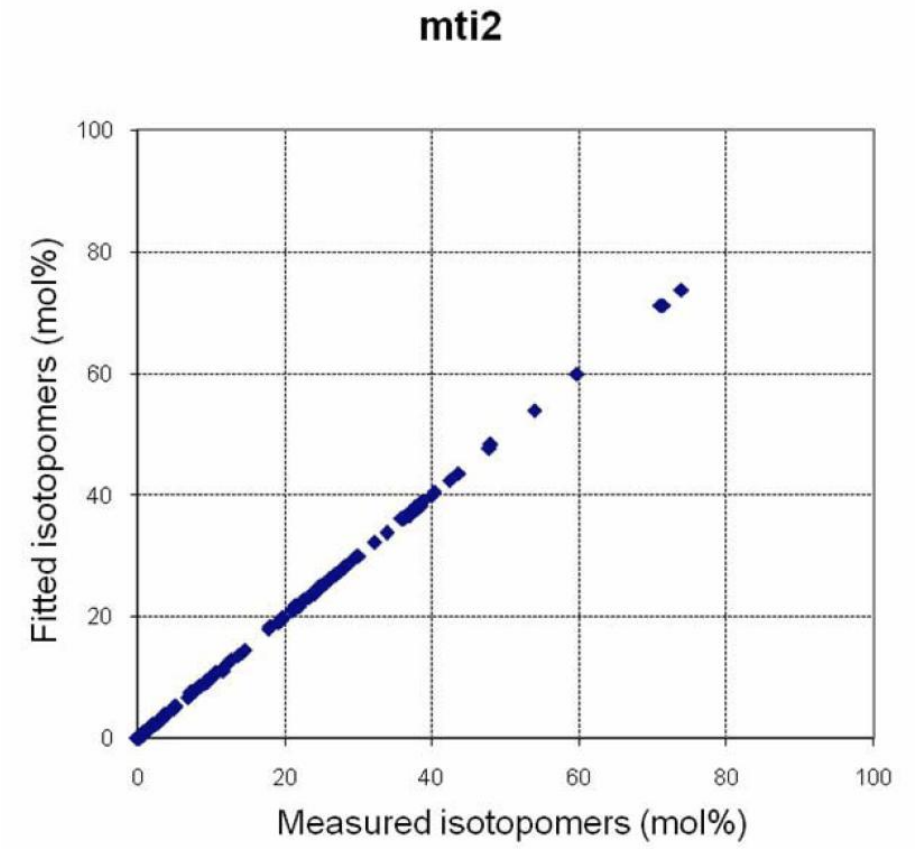


Figure 8. Comparison between computed and measured labeling data [7]. A good fit does not deviate from the diagonal.

References

[1] Katharina Nöh, Aljoscha Wahl, and Wolfgang Wiechert, Metab. Eng **8**, 554-577 (2006).

[2] Jie Yuan, William U Fowler, Elizabeth Kimball, Wenyun Lu, and Joshua D Rabinowitz, Nat Chem Biol **2**, 529-530 (2006).

[3] Ari Rantanen, Juho Rousu, Paula Jouhten, Nicola Zamboni, Hannu Maaheimo, and Esko Ukkonen, BMC Bioinformatics **9**, 266 (2008).

[4] Nicola Zamboni, Sarah-Maria Fendt, Martin Rühl, and Uwe Sauer, Nat Protoc **4**, 878-892 (2009).

[5] YoungJung Chang, Patrick F. Suthers, and Costas D. Maranas, Biotechnology and Bioengineering **100**, 1039-1049 (2008).

[6] M Möllney, W Wiechert, D Kownatzki, and A A de Graaf, Biotechnol. Bioeng **66**, 86-103 (1999).

[7] Yinjie J Tang, Hector Garcia Martin, Samuel Myers, Sarah Rodriguez, Edward E K Baidoo, and Jay D Keasling, Mass Spectrom Rev **28**, 362-375 (2009).

[8] Joel F. Moxley, Michael C. Jewett, Maciek R. Antoniewicz, Silas G. Villas-Boas, Hal Alper, Robert T. Wheeler, Lily Tong, Alan G. Hinnebusch, Trey Ideker, Jens Nielsen, and Gregory Stephanopoulos, Proceedings of the National Academy of Sciences **106**, 6477-6482 (2009).

[9] Francesco Pingitore, Yinjie Tang, Gary H Kruppa, and Jay D Keasling, Anal. Chem **79**, 2483-2490 (2007).

[10] Michael Dauner, Curr Opin Biotechnol (2010).

[11] Shintaro Iwatani, Yohei Yamada, and Yoshihiro Usuda, Biotechnol. Lett **30**, 791-799 (2008).

[12] Patrick F Suthers, Anthony P Burgard, Madhukar S Dasika, Farnaz Nowroozi, Stephen Van Dien, Jay D Keasling, and Costas D Maranas, Metab. Eng **9**, 387-405 (2007).

[13] Schmidt, Marx, de Graaf AA, Wiechert, Sahm, Nielsen, and Villadsen, Biotechnol. Bioeng **58**, 254-257 (1998).

[14] W Wiechert, M Möllney, N Isermann, M Wurzel, and A A de Graaf, Biotechnol. Bioeng **66**, 69-85 (1999).

[15] Maciek R. Antoniewicz, Joanne K. Kelleher, and Gregory Stephanopoulos, Metab Eng **9**, 68-86 (2007).

[16] Wolfgang Wiechert, Michael Möllney, Sören Petersen, and Albert A. de Graaf, Metabolic Engineering **3**, 265-283 (2001).

[17] Nicola Zamboni, Eliane Fischer, and Uwe Sauer, BMC Bioinformatics **6**, 209 (2005).

[18] Lake-Ee Quek, Christoph Wittmann, Lars K Nielsen, and Jens O Krömer, Microb Cell Fact **8**, 25-25 (n.d.).

[19] Yinjie J. Tang, Héctor García Martín, Paramvir S. Dehal, Adam Deutschbauer, Xavier Llora, Adam Meadows, Adam Arkin, and Jay. D. Keasling, Biotechnol. Bioeng. **102**, 1161-1169 (2009).

[20] Esa Pitkänen, Arto Akerlund, Ari Rantanen, Paula Jouhten, and Esko Ukkonen, J Integr Bioinform **5**, (2008).

[21] Uwe Sauer, Mol. Syst. Biol **2**, 62 (2006).

[22] Suzanne M Paley and Peter D Karp, Nucleic Acids Res **34**, 3771-3778 (2006).

[23] Nobuaki Kono, Kazuharu Arakawa, Ryu Ogawa, Nobuhiro Kido, Kazuki Oshita, Keita Ikegami, Satoshi Tamaki, and Masaru Tomita, PLoS ONE **4**, e7710 (2009).

[24] F Le Fèvre, S Smidtas, C Combe, M Durot, Florence d'Alché-Buc, and V Schachter, Bioinformatics **25**, 1987-1988 (2009).

[25] Eva Grafahrend-Belau, Christian Klukas, Björn H Junker, and Falk Schreiber, Bioinformatics **25**, 2755-2757 (2009).

Workflow 3: Inference of Gene Regulatory Networks

Summary

Gene regulatory networks (GRNs) are the "on-off" switches and rheostats of cells that operate at the gene level. They dynamically orchestrate the level of expression for each gene in the genome by controlling whether and how vigorously that gene will be transcribed into mRNA. Understanding how GRNs work is key to systems biology and its successful applications. An array of input data types exists. Knowledgebase users should be able to select an organism, upload, broadcast, or import expression data from public repositories, and submit a request for gene regulatory network inference. Meta-information on experiment design should be automatically parsed from public data, or the user should be prompted to upload this information. Users may want to start with a set of genes or a metabolic process and ask which factors are its regulators. Another use scenario is that researchers may want to know the gene targets of regulatory elements.

Genes will be grouped into putative regulatory modules whose transcription is correlated under specific conditions. For each module, the user selects a subset of known transcription factors and environmental factors that best predict the transcription levels. Additional inputs, such as motifs or protein interactions, may be statistically integrated in the clustering step or the network inference step, and shared regulatory motifs can be computed. Several algorithms have been devised for clustering and discovery of regulatory influences. Results can be exported as raw data or presented to the user in a searchable and browsable form. Subnetworks can be graphically displayed along with views of expression profiles and regulatory motifs and the gene content of individual clusters. Useful output will also include the ability to compute and present predictions (and confidence estimates on predictions) of effect of transcription factor deletions/overexpressions and/or environmental changes.

Inputs

1. Depending on the specific type of network inference analysis a user has in mind, a different combination of the following data might be necessary; but minimally, these seven types of information cover most of what is available today.
2. Measurements of transcription (with confidence values [if avail.]) in the form of an n x m matrix with n genes and m conditions (microarray or sequencing)
3. Measurements of fitness associated with systematic gene knockouts or over-expression (maybe these last two can be condensed in measures of genome-scale gene function with confidence in the form of an n x m matrix.... This could be generalized as phenotype and also have associated confidence depending on how it's measured.
4. Gene interaction network(s) = [nodes (genes), edges (interactions/type), confidence values or weights for edges]
5. Gene locations on genome—with RNA-Seq this is becoming extremely precise with direct measurement.
6. Genome sequence or individual upstream sequences (for motif detection)

7. A list of predictors (transcription factors, environmental factors including metabolites)
8. Machine-readable descriptions of conditions, specifically time series info with standardized measurements of environmental factors

User will specify an organism and import (or broadcast) the above data items. Many of these data types are stored in existing databases and can be loaded automatically through interoperability with these data sources. Many of these data types (items 2-4) can be obtained automatically given the organism. Item 1 may be obtained automatically from expression databases such as GEO or MicrobesOnline. Item 3 can be obtained from STRING, and some information is also accessible in MicrobesOnline. Items 4-6 can be obtained from NCBI or MicrobesOnline or other databases.

However, these data are not available for all organisms. One addendum to this work is that many of these types of measurements follow a standard experimental workflow. Once a genome of a cultivated organism gets sequenced, it might be useful to develop a minimal set of functional measurements to aid in this.

As these workflows are being developed and have increasingly precise data such as RNA-Seq and can have associated confidence measures that can be carried through the analyses, this is providing a basis for comparing the precision of results between methods and laboratories that would help to improve quality and would benefit existing systems such as GEO if applied consistently.

Apply clustering and network inference
Group the genes into putative regulatory modules whose transcription is correlated over a set of conditions. Select a subset of known transcription factors and environmental factors that best predict the transcription levels of each module. Additional inputs, such as motifs or protein interactions, may be statistically integrated in the clustering step or the network inference step, and shared regulatory motifs can be computed. Several algorithms have been devised for clustering and discovery of regulatory influences; some are available in R and MatLab.

Outputs

- Clusters of putatively coregulated genes or biclusters containing genes putatively coregulated under subsets of conditions
- Cis-regulatory motifs
- Regulatory network mapping: influences of predictors on genes within clusters/biclusters directly or through *and* and *or* operations. Confidence values for edges.

Results can be exported as raw data or presented to the user in a searchable and browsable form. Users may want to start with a set of genes or a metabolic process and ask which factors are its regulators. Or, users may want to take a given regulator and ask what are its targets. Subnetworks can be graphically displayed along with graphical views of expression profiles and regulatory motifs and the gene content of individual clusters. Useful output will also include the

ability to compute and present predictions (and confidence estimates on predictions) of effect of TF deletions/overexpressions and/or environmental changes.

Scope

A user should be able to select an organism, upload, broadcast, or import expression data from public repositories, and submit a request for network inference; metainformation on experiment design should be automatically parsed from public data, or a user should be prompted to upload this information. All other data types can be automatically parsed from public repositories—an advanced user should have privileges to change or override default settings by changing source of information, threshold of significance, etc. A user should be given options for choice of algorithms based on the amount and type of available data; the user should have access to published citations for the algorithms and basic information on workings of the algorithm in non-technical jargon-free language. It should be possible to store a session with the default or user-edited settings so the entire analysis can be recreated.

Data requirements and computational complexity

It seems that these might be important numbers both from the user's perspective and from the planning perspective, but can these be coherently calculated? We can give some perspectives to the user, for instance, to infer a network with causal influences time series data are a must; for better coverage of regulons one needs to probe responses to at least half a dozen or a dozen environmental perturbations with different dosages and over the time scale of the response; to incorporate mechanisms we need to have physical interactions (P-D, P-P), or information on TF-cis-regulatory motif relationships. However, in principle, one could learn a network based on correlations and cis-regulatory motifs with a relatively small dataset (30-50 experiments - see Gardner's CLR algorithm or Bar-Joseph's DREM). Such a network will give a very limited view of transcriptional control but could be deemed extremely valuable for an organism for which absolutely nothing was known previously. Given the diverse variations in use cases, while we could consider very simple to very sophisticated cases, I would argue that we should focus on use-cases of simple to mid-scale complexity. I say this because advanced users with sophisticated needs are likely to have the capability to do it themselves (without a knowledgebase).

We could have minimal requirements imposed on algorithm developers when they submit their work. This would include a listing of requirements (number of experiments, interaction data etc.). It might be instructive to have the following information as well; I am not sure if we can generalize this to other use cases.

- Estimates of number and diversity of experiments necessary for clustering
- Estimates of quality needed—issues of quality, compendium biases, etc.
- Estimates of computational complexity of biclustering/bayes nets/etc.

Notes

Are there other players we'd like to incorporate, like RNA regulatory elements, for instance? Would we want to get more out? Network motifs? How about Lee's fusion of kinetics and GRNs? Does that require additional input or generate additional output?

Certainly, moving beyond the inference of regulatory structure and gross dynamics would require different experiments. Inferring metabolism requires both different sorts of functional assays and genome-scale experiments; inferring signaling pathways has its own troubles (see Aindrila Mukhopadhyay's document); inferring complex regulation like that implemented in control of sporulation requires more detailed microscopic measurement and mechanistic modeling. However, here we have the opportunity for something that could almost become a standard after analysis of any sequenced genome.

Inference and Measurement

Is it possible to describe the situations where it is better to try to infer genetic regulatory network topology, rather than try to measure the regulatory interactions directly? There have been remarkable experimental strides made in determining the sequences to which transcription factors bind (e.g., Hesselberth et al, "Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting," Nature Methods 6, 283 - 289 (2009) doi:10.1038/nmeth.1313).

Workflow 4: Signaling

Summary

Due to the focus on microbes and microbial communities, the workflow pertains to signaling in bacteria. Bacterial signaling forms a subset of the Genetic Regulatory Network discussion (see Workflow 3) and therefore contains overlap in experimental design, analysis and workflow.

Microbial genomes present signaling systems to sense and respond to both external and internal stimuli[1,2]. Signals include numerous factors considered to be stresses, intracellular cues, and environmental changes. In bacteria, two component signal transduction systems, typically comprised of a sensor histidine kinase and a response regulator, provide the primary mechanism of signal sensing and response[3,4]. Signal transduction occurs via phosphotransfer or phophorelay and results in an activated response regulator. The best-studied response mechanisms include either the direct modulation of chemotaxis by the activated response regulator, or in a large number of studies, response regulator modulated differential expression of target genes. New classes of response regulators that modulate function via alternate mechanisms such as c-diGMP cyclase or phosphodiesterase domains have also been described[1,2]. Available sequenced genomes from environmental organisms encode numerous sensor and response regulator proteins containing domains of unknown function indicating that additional mechanisms for effector function have yet to be discovered. Environmental bacteria such as *Geobacter metallireducens*, *Desulfovibrio vulgaris,* and the cyanobactrium *Nostoc spp.* have upward of 60, to more than 150, sensor kinases[5]. The responses regulated by the corresponding two component systems are no doubt at the core of environmental process of key significance. These systems also provide the parts for developing valuable sensory modules to build sophisticated engineered systems (using synthetic biology methods).

Definition of a signal: With regard to the type of research being conducted by the Genomic Science groups, signals can vary widely. In environmentally relevant microbes, a signal could be a change in environmental cue (e.g., the lack/abundance of resources such as carbon source, electron acceptors, electron donors, amino acids, vitamins, etc.); stresses (e.g., salt, pH, heat, cold, metals, toxins, oxygen, a variety of small molecules); or variability in other organisms in the microenvironment. The responses to these signals, including the triggering of altered physiological states (e.g., biofilm formation, sporulation, virulence, swarming, etc.) are all initiated via signal sensing and corresponding response. In microbes that are being engineered for industrial uses (e.g., biofuel production), perturbation from toxins present in carbon feed, intracellular triggers due to imbalance in metabolic intermediates, and accumulation of final (often toxic) products serve as signals.

A vast body of knowledge exists for these systems from individually studied systems. Efforts to compile and integrate information on regulatory modules from such studies have only recently begun to emerge as described in Workflow 3. However, the impact of multiple stimuli on a given organism or comprehensive understanding of all signal sensing for a single organism is still rare. In the few cases where such studies have been undertaken, valuable and interesting phenomenon have been discovered[6,7]. The tremendous increase in sequenced organisms and

corresponding computational[2,5,8] and experimental tools[9,10] now makes it more possible to undertake such efforts.

Metadata: This documents what signals are being studied, under what defined conditions, on which organisms and by whom? What media and growth phases are being used? What methods are being used (sequencing, arrays, analytical)? What analysis tools, *in silico* prediction algorithms, and validation experiments are to be used?

Aspects of studying signaling

- *What are the genomes in question?* For a given organism, there may be more than one genome sequence if there are modified, engineered, evolved, adapted or shuffled versions.
- *Sensing and responding to signals:* Study of two-component, cAMP, and c-diGMP systems, transcriptional factors, global regulators, sigma factors; cell-wide studies (transcriptomics, proteomics, and metabolomic studies), and mapping ligand (signal) binding, phospho-transfer, and other post-translation modification.
- *Information gathered:* Ligand binding and transport, two component phosphorylation, other assays (chemotaxis, binding to cyclic-diGMP, DNA gel shifts), ChIP-chip arrays, ChIP-seq, microrrays, RNA-Seq, mapping post translational modifications (phosphoproteome, methylations etc), mapping protein interactions and localization.

Data types

This will form the core of the database for this topic

- Genome sequences
- Knockout and expression libraries: corresponding phenotypic data (e.g. from omniglogs or other such HT strategies)
- Transcript level data: microarray, RNA-Seq, absolute mRNA quants (e.g. nCOUNTER)
- DNA binding: ChIP-chip, ChIP-seq, microfluidics
- Mass Spec data: Protein levels, Post translational modifications, metabolites
- (data from different types of mass spectrometers)
- Ligand binding mapping: Semi HT
- Regulator-DNA binding: Semi HT
- Regulatory motifs and maps generated using computational methods

Resources

- Common sensory proteins include histidine kinases, methyl-accepting chemotaxis receptors, Ser/Thr/Tyr protein kinases, adenylate and diguanylate cyclases and c-di-GMP phosphodiesterases. A webpage maintained Galperin and coworkers contains a

fairly comprehensive repository of signal transduction systems from over 500 bacteria and archaea[11].

- Predictive tools and databases (e.g. Regtransbase [12], MicrobesOnline[13], and MiST[16]).
- Specific tools for predicting cognate partners of sensor histidine kinases and response regulators such as that developed by Burger and van Nimwegen[8].
- Methods developed for mapping two component phosphotransfer and relay[9], rewiring sensor kinases[14].
- Classification system developed for categorizing bacterial signaling proteins[1,2,15].

Illustration using one concrete problem

BESC is studying a number of Caldicellulosirupter species, which are non-sporulating, anaerobic, gram positive, thermophilic bacteria that can facilitate the direct conversion of cellulosic biomass (e.g. from switchgrass) to ethanol, H2 and other products.

A study of this organism will utilize the features afforded by other knowledgebases: Annotated genomes and their use in generating arrays, predictions for regulatory networks and motifs, predicted two component systems, transcriptional factors (including those that work with transporters), sigma factors, small RNA regulators.

A given experiment would entail growth of Caldicellulosirupter on ground plant material and monitor production of waste products including ethanol, H2, acetate etc. Genetic engineering could create the production of alternative end products in different proportions.

A range of factors (signals) would be examined in this context. Beneficial factors include C source, cell density, etc. Harmful factors include exposure to O2, cold shock, inhibitors from lignocellulosic biomass, acetate, non-optimal pH, salt, and the accumulation of other final products.

Current studies include: Log phase growth using cellobiose and switchgrass (pretreated), stationary phase growth in switchgrass (pretreated), log phase growth under ethanol stress with either cellobiose or switchgrass.

A systematic examination of any of the above factors could be conducted using the following:

1. Transcript level measurements: arrays, RNA-Seq, other targeted measurements.
2. ChIP-chip, ChIP-seq
3. Analytical assays:
    a. HT: Mass spec based analysis (protein levels, protein complexes, PTMs)
    b. MT: ligand-docking, transport,
    c. LT: Mapping HK-RR phosphotransfer, RR-DNA gel shifs
    d. LLT: Imaging for morphological changes or cellular localization of complexes.
4. Study of knockout or expression strain libraries (transposon, targeted, site specific). Corresponding phenotypic data and iterative (1), (2) and (3)

5. Collecting and integrating the above data types into previous or initial regulatory network prediction (see Workflow 3: Inference of Gene Regulatory Networks).

HT: High throughput; **MT**: Medium throughput; **LT**: Low throughput; **LLT**: Low Low throughput; **PMT**: Post translational modifications.

References

[1.] Galperin, M. Y., Diversity of structure and function of response regulator output domains, *Curr Opin Microbiol* 13 (2), 150-9, 2010.

[2.] Galperin, M. Y., Higdon, R., and Kolker, E., Interplay of heritage and habitat in the distribution of bacterial signal transduction systems, *Mol Biosyst* 6 (4), 721-8, 2010.

[3.] Stock, A. M., Robinson, V. L., and Goudreau, P. N., Two-component signal transduction, *Annu Rev Biochem* 69, 183-215, 2000.

[4.] Gao, R. and Stock, A. M., Biological Insights from Structures of Two-Component Proteins, *Annu Rev Microbiol*, 2009.

[5.] Alm, E., Huang, K., and Arkin, A., The evolution of two-component systems in bacteria reveals different strategies for niche adaptation, *PLoS Comput Biol* 2 (11), e143, 2006.

For histidine kinases in various genomes see www.microbesonline.org/cgi-bin/hpk/browse.cgi

[6.] Tagkopoulos, I., Liu, Y. C., and Tavazoie, S., Predictive behavior within microbial genetic networks, *Science* 320 (5881), 1313-7, 2008.

[7.] Skerker, J. M., Prasol, M. S., Perchuk, B. S., Biondi, E. G., and Laub, M. T., Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis, *PLoS Biol* 3 (10), e334, 2005.

[8.] Burger, L. and van Nimwegen, E., Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method, *Mol Syst Biol* 4, 165, 2008.

[9.] Laub, M. T., Biondi, E. G., and Skerker, J. M., Phosphotransfer profiling: systematic mapping of two-component signal transduction pathways and phosphorelays, *Methods Enzymol* 423, 531-48, 2007.

[10.] Zhou, L., Lei, X. H., Bochner, B. R., and Wanner, B. L., Phenotype microarray analysis of *Escherichia coli* K-12 mutants with deletions of all two-component systems, *J Bacteriol* 185 (16), 4956-72, 2003.

[11.] Galperin, M. Y., Higdon, R., and Kolker, E., www.ncbi.nlm.nih.gov/Complete_Genomes/SignalCensus.html, 2010.

[12.] Kazakov, A. E., Cipriano, M. J., Novichkov, P. S., Minovitsky, S., Vinogradov, D. V., Arkin, A., Mironov, A. A., Gelfand, M. S., and Dubchak, I., RegTransBase--a database of regulatory sequences and interactions in a wide range of prokaryotic genomes, *Nucleic Acids Res* 35 (Database issue), D407-12, 2007.

[13.] Dehal, P. S., Joachimiak, M. P., Price, M. N., Bates, J. T., Baumohl, J. K., Chivian, D., Friedland, G. D., Huang, K. H., Keller, K., Novichkov, P. S., Dubchak, I. L., Alm, E. J., and Arkin, A. P., MicrobesOnline: an integrated portal for comparative and functional genomics, *Nucleic Acids Res* 38 (Database issue), D396-400, 2009.

[14.] Salis, H., Tamsir, A., and Voigt, C., Engineering bacterial signals and sensors, *Contrib Microbiol* 16, 194-225, 2009.

[15.] Mascher, T., Helmann, J. D., and Unden, G., Stimulus Perception in Bacterial Signal-Transducing Histidine Kinases, *Microbiol. Mol. Biol. Rev.* 70 (4), 910-938, 2006.

[16]. Ulrich, L.E., Zhulin, IB. MiST: a microbial signal transduction database. *Nucleic Acids Res.* 35:D386-390, 2007

Workflow 5: Structural Biology

Summary

The nucleic acid sequence of a protein gene encodes an amino acid sequence that typically folds to generate a specific three-dimensional shape. This structure is often vital for the protein's function. In enzymes, the structure serves to keep key catalytic residues in a unique geometry, poised to act on substrate molecules. As such, the relationship between primary sequence and tertiary shape is central to our understanding of molecular biology. There is a wealth of information that is ripe for analysis in the context of the Knowledgebase. For example: the relationship between sequence and fold (for proteins and nucleic acids), the assembly of single molecules to form larger complexes, and the evolutionary relationships within and between protein families. Rapid progress can be made in providing functionality to researchers via the Knowledgebase. Initial workflows can focus on visualizing the linkage between sequence and structure (see first and second workflows below) and dissection and visualization of cellular compartments (see third workflow below). These will provide users with powerful tools to probe sequence/structure relationships, which otherwise are limited to experts.

Three structural biology workflows were submitted:

1. Locating and visualizing an enzyme active site
    a. Goal: Assign, then visualize the amino acid residues in a protein sequence involved in enzymatic activity
2. Determine and visualize the oligomeric state of molecular complexes
    a. Goal: Determine and then visualize the oligomeric state of a protein complex
3. Locating and visualizing a cellular compartment
    a. Goal: Locate (segment) and visualize one or more cellular compartments in a microbe.

For each research goal, the Inputs, Analysis process, Outputs, Tools, and Knowledgebase context were provided.

1. Locating and visualizing an enzyme active site

Goal: To assign and then visualize the amino acid residues in a protein sequence involved in enzymatic activity.

Inputs

- Sequence of a protein
- High resolution protein structure (from X-ray crystallography or NMR) or high fidelity homology model
- One or more related sequences/structures with known active site residues

Process

- Perform alignment (most likely using multiple proteins) of sequence with unknown active site residues (may also include multiple family members) against known residues
- In cases of high sequence similarity, the active site residues in the unknown can be identified by sequence conservation
- In cases of remote similarity, more complex models (e.g. hidden Markov, sequence motifs, combined sequence/structure alignment) may need to be generated to infer the likely equivalent residues in the unknown
- Predictions of active site residues can be validated against any prior biochemical data and/or phyologenetic information

Outputs

- Protein sequence with active site residues highlighted
- Visual representation in standard molecular viewing software with active site residues highlighted

Tools required

- Parsing protein structure and sequence
- Single and multiple sequence alignment
- Combined sequence/structure alignment
- Sequence display
- 3D structure display

Knowledgebase context

- Provides linkage to and automatic retrieval of related structures in the Protein Data Bank
- Performs complex sequence and sequence/structure analysis without detailed user learning
- Cross validates against other experimental data within the Knowledgebase and in other outside resources
- Displays results in easy to understand visual forms and for download and subsequent analysis

2. Determining and visualizing the oligomeric state of molecular complex

Goal: To determine and then visualize the oligomeric state of a protein complex.

Inputs

- Sequence of a protein
- One or more related sequences/structures with known oligomeric state
- Optionally experimental data to define oligomeric state, such as small angle X-ray scattering (SAXS)
- Optionally high resolution protein structure (from X-ray crystallography or NMR) or homology model

Process

- Perform alignment (most likely using multiple proteins) of sequence with unknown oligomeric state (may also include multiple family members) against sequences of known state
- In cases of high sequence similarity, the likely oligomeric state can be identified from the nearest similar sequence
- In cases of remote similarity, more complex models (e.g. combined sequence/structure alignment) may need to be used to determine if structural features involved in oligomerization interfaces are likely to be conserved
- Predictions of oligomeric state can be validated against any prior experimental data (e.g. SAXS), biochemical data and/or phyologenetic information

Outputs

- Three-dimensional model of oligomer
- Protein sequence with residues involved in oligomerization highlighted
- Visual representation in standard molecular viewing software with interface residues highlighted

Tools required

- Parsing protein structure and sequence
- Single and multiple sequence alignment
- Combined sequence/structure alignment
- SAXS data analysis
    - Calculation of standard distributions
    - Comparison of distributions to those calculated from 3D models
    - Searching of known structures for similar SAXS curves

- Protein structure writing
- Sequence display
- 3D structure display

Knowledgebase context

- Provides linkage to and automatic retrieval of related structures in the Protein Data Bank
- Performs complex sequence and sequence/structure analysis without detailed user learning
- Cross validates against other experimental data within the Knowledgebase and in other outside resources
- Displays results in easy to understand visual forms and for download and subsequent analysis

3. Locating and visualizing a cellular compartment

Goal: To locate (segment) and visualize one or more cellular compartments (e.g. mitochondria) in a microbe.

Inputs

- Three-dimensional reconstruction of one of microbes of interest (e.g. from EM-tomography or soft X-ray tomography)
- Characteristics describing the compartment of interest (e.g. shape, density, proximity to other features), or a human-generated training set
- Optionally a visual label indentifying the compartment of interest

Process

- Read 3D data
- Perform pattern matching analysis to identify likely compartments on the basis of input data
- Segment volume data to assign the identity of compartments (note that for some data, it is possible to *a priori* segment on the basis of density, but the problem of identifying compartments still remains)
- Calculate statistics (e.g. volume of cell occupied by compartment, standard deviations between samples)
- Cross validate against any other relevant biochemical data

Outputs

- Statistics of compartments segmented
- Visual representation in volume rendering viewing software with compartments highlighted

Tools required

- Parsing large 3D volume datasets
- Pattern matching algorithms to identify compartments
- 3D volumetric data display

Knowledgebase context

- Performs segmentation analysis without detailed user learning
- Cross validates against other experimental data within the Knowledgebase and in other outside resources
- Displays results in easy to understand visual forms and for download and subsequent analysis

Workflow 6: Imaging Bioinformatics

Summary

One of the major advantages of phenotypic characterization through microscopy is the ability to visualize cellular organization, morphology and ultrastructure, and localization. More importantly, microscopic imaging allows cell-by-cell measurements, revealing a cellular heterogeneity that is often lost when using OMIC data only. For example, *Desulfovibrio vulgaris* (Dv) is known to form micro-colonies at certain stages of development due to cell-cell communication, a complex mechanism that remains largely unknown. Such population phenotypes can then be interrogated at multiple scales through multiplexed imaging probes to identify changes in structure, morphology, and localization on a cell-by-cell basis. These morphometric features can then be linked to omic data to query molecular predictors of a specific phenotypic subset. The main challenge in managing image-based data is identifying a quantitative view for each assay, which can be integrated with omic data. These quantitative views are often represented as vectors and relationships between vectors. Figure 1 is an example of a typical workflow in *Imaging Bioinformatics*.

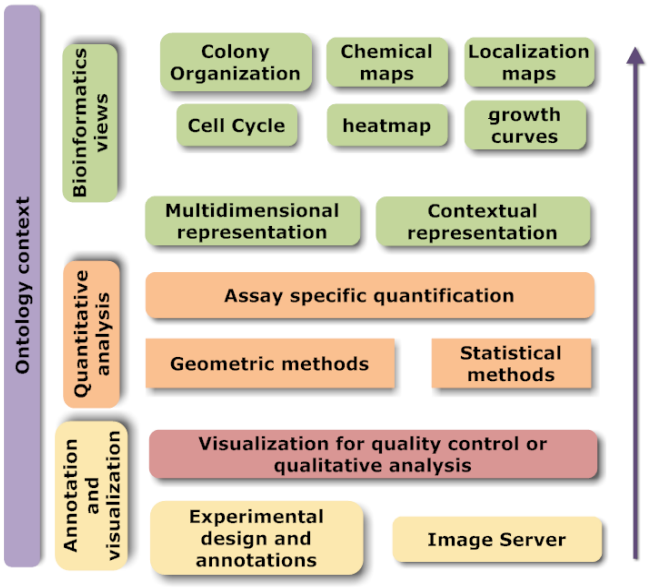


Figure 1. A typical workflow for image-based applications represents potential layers for image-based registration, quantitative analysis, and bioinformatics representation. Integration with omic data should take place through an ontology layer that provides a common set of vocabulary and ontology for exchanging bioinformatics views and their corresponding quantitative results.

Input Data

The input data consists of four types of information: (i) experimental design variables, (ii) imaging system parameters, (iii) raw image files, and (iv) queries used to target specific endpoints. (i) Experimental design refers to the model system, stress conditions, harvest time, imaging assay (e.g., labeling), etc. The main challenge has been reducing the number of user interactions needed to specify experimental design variables, since one rarely enters metadata at the granularity level that is often needed. There are no standards for capturing experimental variables; however, the microarray community has defined a complete protocol that can be leveraged. (ii) Most modern microscopes capture instrument setup information (e.g., optical path, illumination source) and store it as a header (e.g., in the form of a TIFF header) with raw data. Nevertheless, the Open Microscopy Environment (OME) has defined a schema for

specifying instrument configurations, and some vendors plan to support the proposed schema. (iii) Raw data are usually stored in a binary form, and the format varies between different vendors. However, LOCI and OME developed a transformer that can read any image format, parse it, and store it in a five-dimensional format. The end result is a homogenized representation of a diverse file format. (iv) The endpoints or biological queries have to provide a series of templates for guiding quantitative analyses. One can design a taxonomy that allows users to select from multiple templates.

Quantitative Analysis

Analytical requirements for image-based data are quite heterogeneous, and any computational pipeline must be extensible for new application software programs. However, common computational modules can be defined, integrated, and enhanced for a specific application. Nevertheless, there has to be a balance between excessive generalization versus specificity, as too much generalization adds to the complexity of a system and thus increases the learning curve required to use it efficiently. In general, image-based data analysis needs to incorporate a model to recover objects of interest in a robust fashion. Such a model can be expressed either geometrically or statistically. In some cases, model-free methods can be used, at a low level, to aggregate rich tokens for higher-level analysis. Once the images have been quantified, information can be composed and aggregated to form bioinformatics views (see "Output Data" below for more information). With respect to image analysis, the ITK image library provides a rich set of software and an extensible framework for adding new applications. However, it requires expertise in advanced software engineering, which may not be readily available at every institution.

Output Data

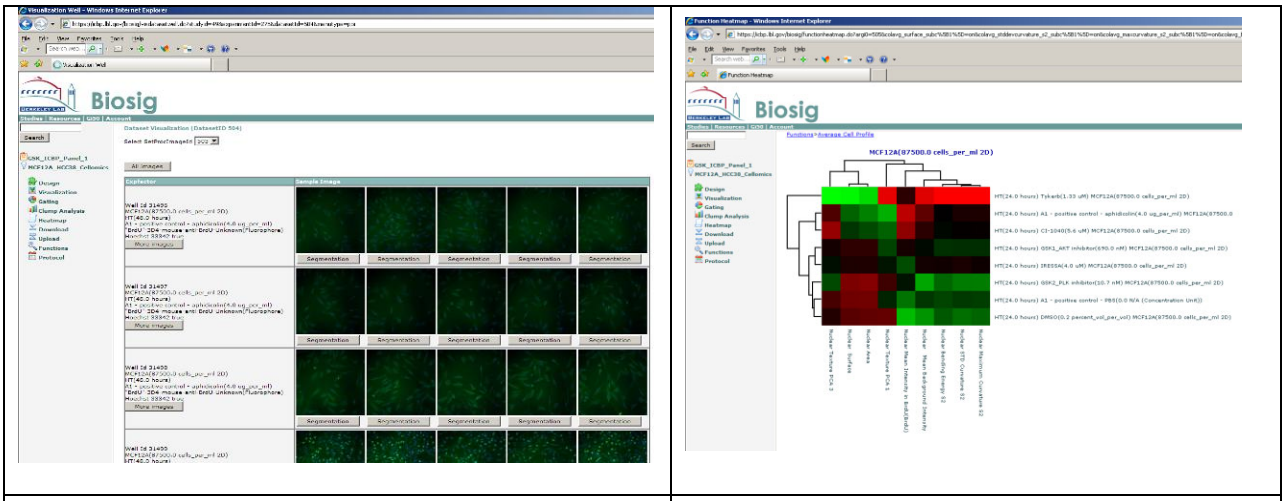
One of the characteristics of image-based assays is that a large number of data are often transformed into a very small amount of data. This is referred to as "bioinformatics views," which are often constructed by downloading computed information, and then processed further by using one of many statistical or data analysis stand-alone software packages. However, it is possible to integrate some basic capabilities into the bioinformatics platform. Examples include a dose-response curve, a growth curve, and co-localization frequencies. One of the advantages of imaging is that it maps cellular localization (or co-localization), chemical composition, and morphometric properties.

Current State of the Art

BioSig (ribo.lbl.gov:8080/biosig/home.do) is an example of an imaging bioinformatics system, which is being used for mammalian systems. BioSig builds on OME for image harmonization, leverages MIAMI (www.mged.org) standards for specifying experimental design variables, and has defined a number of tagged templates for assay-specific quantitative analysis. It also supports a schema for multidimensional profiling of cell-based assays for high-content screening, as shown in Figure 2.

The Next Generation

Current imaging bioinformatics platforms lack (i) an ontology and controlled vocabulary for microorganisms of interest to DOE, (ii) an integrated pipeline for bioinformatics and image analysis, (iii) an interface for integrating omic data, and (iv) the necessary analysis tools for mapping at multiple scales of different imaging modalities. The latter is quite important since it enables chemical mapping (e.g., Raman microscopy), localization mapping (e.g., electron or optical microscopy), and mass spectrometry imaging (e.g., MALDI imaging). Furthermore, having created these maps at multiple scales, one is also interested in correlative analysis between these imaging modalities for the model systems of interest under specific environment conditions. A potential correlative query would be how the chemical composition of the plant cell wall, visualized and quantified with Raman, is altered as a result of increase in biomass that is imaged with electron microscopy. In short, the next generation of breakthroughs in quantitative image analysis and imaging bioinformatics resides at the interface of different imaging modalities, and their integration with omic data.



(a) (b)

Figure 2. Imaging bioinformatics views. (a) Thumbnail visualization enables comparison of biological replicates (columns) for each set of experimental variables (rows). Each thumbnail is a hyperlink to a full- resolution version, where quantitative results can be overlaid on top of it. (b) Images in (a) are processed, and each cell is represented by multidimensional features. The user can select a subset of computed features, put them in a particular order, and view them through a heatmap. As a result, multiple phenotypic representations can be viewed simultaneously and compared in the context of experimental variables.

Section III: Strawman Knowledgebase Architecture

The preliminary diagram below was developed as a result of discussions held in conjunction with this workshop. Though this schematic will be refined in upcoming discussions, it is included in this report to indicate how the workflows (research protocols) relate to the ultimate system architecture. The workflows being developed by experimentalists to satisfy scientific objectives are critical to the development of many Knowledgebase architecture layers, such as data repositories (red), computing workflow management, and output visualization design.

The workflows provide information on data sources and types that must be accommodated by the Knowledgebase architecture. In-depth discussions will result in refinement of the workflows by the research and computing communities.

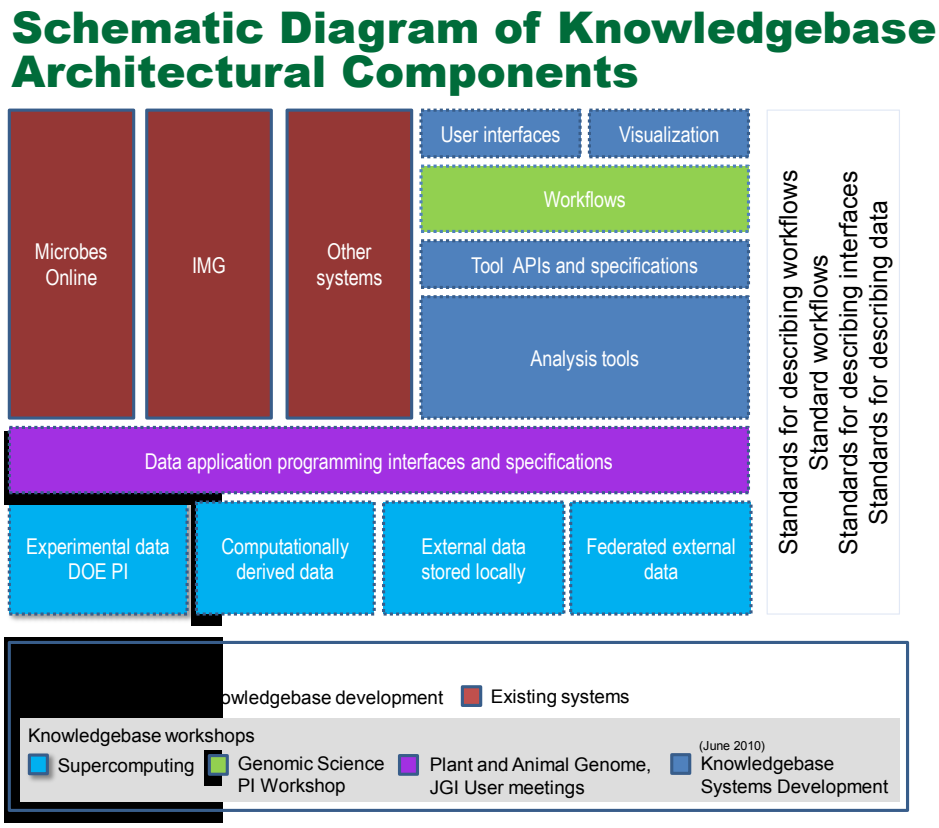


Fig. 3.1. Example Schematic of Knowledgebase Architectural Components.

Data

The data layer represented in the bottom of Fig. 2 illustrates several data components that will be important to achieving the goals of the Knowledgebase project. These data components will utilize several technologies. Relational database technology such as Oracle or MySQL will be used to manage data that is well structured and suited to relational technologies. Examples include the storage of account information, user configurations, and certain data and tool-related metadata. More recently developed technologies for representing data, such as Semantic Web, will be used for biological data with complex data model characteristics.

An important component of the data layer is data available from other sites that are remotely accessed and used singularly or in a federated manner. Federating external data sources will make heavy use of web services technologies. Web services are newer technologies allowing interoperability between software systems located at distinct sites. Federation will allow us to leave stable data at remote sites (i.e., NCBI Taxonomy) when a façade (wrapper, adaptor, bridge, etc.) can be constructed around the access routines provided by the remote site. The façade will serve to standardize access to data provided by multiple, distinct remote data sources.

Experimental data derived from DOE-funded work that is not available in other data sources in a suitable format will be structured and shared appropriately as part of the data layer. This data generally is thought of as the results of experiments funded by DOE. The data should not be limited to DOE-funded work; if others outside DOE wish to contribute, all the better.

The data layer also will contain data that exists remotely but is aggregated locally. Local aggregation can enhance data usefulness by putting the data in a modified format that corrects for missing metadata, incompatible formatting, or because internal computation integrates additional data. Pathway data, genome data, transcriptome data, and regulatory network data all stored in a suitable form for mash-ups are examples of data that likely will be found in the data layer component that represents locally aggregated external data. Another example of why external data is aggregated locally is because there will be external published data of use and specific data derived from DOE-funded work that is not available in the public domain. Other examples can be driven simply by the fact that computations such as similarity searching require local data sources for performance reasons.

Computationally derived data should represent another component of the data layer. Computations often can produce entirely new datasets rather than just adding value to existing ones. These computations may operate on existing datasets but generally produce a new type of data. For example, a computation on RNA sequencing–based gene expression data might produce a histogram of coverage statistics. This histogram is a new data type linked to the RNA sequencing data through descriptive metadata technology.

Analysis

The analysis component of the architecture will allow for development of both libraries and interfaces that promote the integration of analytical tools into the recognizable Knowledgebase. This component also will provide the facilities needed by the community to develop new algorithms and applications enabled by Knowledgebase infrastructure and data

access layers. Goals of the Knowledgebase project are to build and promote an environment where analysis tools are primarily derived from open contributions.

In direct support of the data layer, semantic-enabled search algorithms will allow metadata—information about the data stored in the data layer—to be more than just an attached publication or protocol list. Metadata is extremely valuable when making new scientific discoveries. Representing, integrating, searching, and performing logic on metadata can be challenging enough when the object being measured is sequence or crystal structure. Complex and conditional data derived from functional measurement of molecules, cells, and communities will make this an exceptional challenge. The rapid development of new technologies for making such measurements further increases the need to track the key information about experimental and analytical protocols for producing data and the processing applied to data before it is stored in accessible formats. A key goal for this effort will be developing tools that attach such information to data as easily as possible, identify the most important pieces of this data for scientific purposes and searching, capture experimental design and goals, and allow queries of this information.

Existing and New Systems and Projects not Developed by the Knowledgebase

Existing systems such as MicrobesOnline, the collection of IMG systems, the RAST systems, and others are expected to continue and benefit from the centralized or virtually centralized (federated) data stores and from direct programmatic access to the open methods developed as part of the Knowledgebase. We also anticipate that new systems will emerge.

It is expected that existing system developers can and will create application programming interfaces (APIs) to their systems and publish the specifications of those interfaces as part of Knowledgebase API specifications. These API specifications are analogous to the Sun Java Docs for the Java APIs. These interfaces may be used by other existing system developers or by the Knowledgebase development community.

As new systems emerge, embracing and nurturing them will be important. Guiding such projects so that they become important components of the Knowledgebase also will be necessary.

Workflows

Scientific workflows can help scientists, analysts, and computer programmers create, execute, and share experimental and analytical processes. These workflows can be captured as free text use cases or more formally represented using workflow languages. Regardless of whether a workflow is captured in a structured or unstructured manner, an important part of the Knowledgebase system architecture will be a graphical user interface that is available to the community so that anyone can access existing workflows and develop new ones.

User Interactions

The user experience will primarily take place through what is known as a horizontal web portal. These portals deliver an integrated front end to what is commonly thought of as several independent websites that allow users to easily search, visualize, and run analytical software on Knowledgebase information. Standard browsers, plugins, and web portal technology will enhance the user experience when command line or other existing user interfaces are not

suitable. This will allow members of the research community to have a customizable entrance to the Knowledgebase.

User interactions can be thought of in two parts—user interfaces and visualization—because these two components use markedly different technologies. For an effective Knowledgebase user experience, we will need to focus on both the user interfaces and the more challenging aspects of scientific data visualization.

Relying heavily on web technologies such as HTML, Javascript, and their derivatives, user interfaces allow a user to navigate the system, configure analysis environments, input data into the system, and retrieve results from it. These user interfaces can leverage the latest advancements in social networking to provide tools to the community for shared annotation and quality assessment and provide forums that easily reference Knowledgebase information.

Visualization (referred to as scientific visualization in some communities) relies heavily on graphics packages. The goal is to present data in a form that is useful for scientific discovery. There may be no interaction required when a visual representation of data is generated and presented.

Standards

Standards will be instrumental in achieving many aspects of the Knowledgebase project. These aspects range from scientific to engineering. From a scientific perspective, describing biological data and the relationships between data in a standardized way is critical to advancing our ability to interpret it. In relation to engineering, standards will enable healthy, continued evolution and growth of the system.

Several objectives will provide efficient and necessary utilization of standards. These objectives include embracing community standards when they are adequate, engaging in the community-development process of a particular standard when there is an existing standard that might be considered inadequate in its current form, and helping the community by initiating standards development where gaps exist.

Although standards for describing data and workflows will be critical, other types of standards will be important as well. Having community standards for data sharing is just one example of what will have to be supported in the Knowledgebase project. Another such example is developing standard workflows and benchmarking data that can be used by the community to facilitate a higher level of exchange among scientists.

In support of an open environment, standards for describing analytical tools, software libraries, data schemas, and other technical artifacts used to build the Knowledgebase will be essential for broad acceptance and use. Software tools and libraries implemented in the Java programming language benefit from a Java community–accepted standard on how to describe APIs. Requiring the use of these standards in code libraries will result in a solid documentation base that is needed for general acceptance and further use of the library by the community. Other programming languages such as Perl have similar standards.

Section IV: Workshop Summary and Conclusions

Workshop participants discussed the need for some level of individual research privacy, which could be achieved with user accounts. Data and code could be held in private, and analyses conducted in a nonpublic environment. The Knowledgebase also will need to allow users to track version history and provenance so that new analyses can be usefully compared with previous ones. Other important capabilities workshop participants discussed include:

- Curation not only of data, but also of models and representation of scientific concepts
- Comparison and analysis of methods and results over time
- Simulation, including the ability to modify and improve models
- Predictions based on simulation and analysis to form new hypotheses
- Comparison of predictions and results to guide experimental design

Only a few researchers today have comprehensive access to such computational capabilities, yet these tools are necessary to conduct research that will lead to important scientific innovations in energy and environment.

Also envisioned for the Knowledgebase are high standards for usability, understandability, discovery, and contribution. System design should be intuitive so that researchers can use it with minimal training. Knowledgebase components also need to be understandable. Although able to use a given software package, many people often do not understand the process by which the software derived its results (e.g., BLAST). Understandability implies that there is a good foundational basis for knowing that results returned to a user are based on robust scientific knowledge or assumptions. If results are not understandable, system features should allow the user to drill down to acquire information about how results were obtained. The Knowledgebase also should promote an environment of discovery, leading to new rounds of experiments or lines of research. Finally, engaging the entire research community in Knowledgebase contribution is critical. Any system being used by scientists ultimately should be measured on how well it accomplishes these concepts, advances research, and supports the scientific method.

Future Considerations for Workflow Definitions. Here we see a range of styles and level of content in the workflows. For the future final report of the Knowledgebase R&D Project, we will need to settle on a style. The Structural Biology workflow is very terse when compared with the others, but it is also very clear. In developing a standard for future workflows, this should be considered. An important question to raise: Do these workflows provide sufficient detail to allow requirements to be established that can drive the Knowledgebase Implementation Plan, and if not, how much more detail is needed?

Appendix 1: Agenda

DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop

Crystal City, Virginia
Tuesday, February 9, and Wednesday, February 10, 2010

February 9

2:00 – 2:30 p.m.	Robert Cottingham, Oak Ridge National Laboratory "Microbial Systems Biology Knowledgebase: Scientific Objectives and Current Prospects"
	Focus on examples of scientific objectives, benefits, and outcomes
2:30 – 3:00 p.m.	Discussion
3:00 – 3:30 p.m.	Robert Kelly, North Carolina State University "Near-Term Prospects for Functional Microbial Genomics: Moving Beyond the Monoculture Paradigm"
	One organism to two organisms, adding complexity
3:30 – 4:00 p.m.	Discussion
4:00 – 4:30 p.m.	Adam Arkin, Lawrence Berkeley National Laboratory "From Pathways to Populations and Back Again: Long-Term Prospects for the Microbial Systems Biology Knowledgebase"
	Much larger complexity of systems, data, models, and impacts
4:30 – 5:00 p.m.	Discussion
5:30 p.m.	Adjourn

February 10

1:00 – 3:00 p.m.	Impromptu follow-up session focusing on workflows

Appendix 2: Participants and Observers

Adam Arkin (LBNL)
Nitin Baliga (Institute for Systems Biology)
Jill Banfield (University of California, Berkeley)
Chris Bare (Institute for Systems Biology)
Ben Bowen (LBNL)
Tom Brettin (ORNL)
William Cannon (PNNL)
James Cole (Michigan State University)
Robert Cottingham (ORNL)
Brian Davison (ORNL)
Mitch Doktycz (ORNL)
Ronan Fleming (University of Iceland)
Cheri Foust (ORNL)
Hector Garcia Martin (LBNL/JBEI)
George Garrity (Michigan State University)
Adam Godzik (Burnham Institute)
Susan Gregurick (DOE)
Loren Hauser (ORNL)
Alyssa Henning (Cornell University)
Kimberly Keller (University of Missouri)
Robert Kelly (North Carolina State University)
Julia Krushkal (University of Tennessee, Memphis)
Libbie Linton (Utah State University)
Yukari Maezato (University of Nebraska)
Betty Mansfield (ORNL)
Victor Markowitz (LBNL)
Lee Ann McCue (PNNL)
Folker Meyer (ANL)
Jonathan Millen (University of Rochester)
Aindrila Mukhopadhyay (LBNL)
Kristen Munch (NREL)
Ambarish Nag (NREL)
Chongle Pan (ORNL)
Nicolai Panikov (Northeastern University)
Morey Parang (ORNL)
Charles Parker (Names for Life, LLC)
Bahram Parvin (University of California)
Amanda Petrus (University of Connecticut)
Madeleine Pincu (University of California, Irvine)
David Pletcher (JBEI/LBNL)
Iris Porat (ORNL)
David Reiss (Institute for Systems Biology)
Dmitry Rodionov (Burnham Institute)
Blake Simmons (LBNL)
Steve Singer (LBNL)
Marvin Stodolsky (DOE)
Ines Thiele (University of Iceland)
Judy Wall (University of Missouri)
Sharlene Weatherwax (DOE)
Steven Wiley (PNNL)
Jian Yin (PNNL)

Acronyms

ANL	Argonne National Laboratory
DOE	U.S. Department of Energy
JBEI	Joint BioEnergy Institute
LBNL	Lawrence Berkeley National Laboratory
NREL	National Renewable Energy Laboratory
ORNL	Oak Ridge National Laboratory
PNNL	Pacific Northwest National Laboratory

DOE Systems Biology Knowledgebase Workshop Report from the 5th Annual JGI User Meeting

Tuesday, March 23, 2010, 8:30 a.m. – 5:00 p.m.

Convened by

The U.S. Department of Energy Office of Science as part of the DOE Joint Genome Institute's (JGI) Genomics of Energy and Environment 5th Annual User Meeting, Walnut Creek, California

Workshop Organizers: Susan Gregurick (DOE) and Bob Cottingham (Oak Ridge National Laboratory)

Workshop Cochairs: Victor Markowitz (JGI, Lawrence Berkeley National Laboratory) and Jill Banfield (University of California–Berkeley)

Table of Contents

1. Introduction

2. Background

3. Topics Discussed at the Workshop

3a. Proposed Science Objectives for Microbial Community Analyses

3b: Standards: The Role of Standards-Setting in the Knowledgebase

3c. Tool Builders and Data Generators: The Need to Engage the Various Scientific Communities

3d. NIH Interactions: Data Resources and Leverage

4. Expanded Discussion from Workshop: Assignments

4a. Workflows and the Systems Biology Knowledgebase

4b. Metagenomics Systems Biology Knowledgebase: Workflows, Background, Design Goals, and Recommendations

4c. Toolkit Registry Development

4d. The Knowledgebase as an Open Development Platform

4e. Institutional and Career Considerations Surrounding Open Source Development

4f. Other Potential Science Objectives and Knowledgebase Features Drawn from Responses to Preworkshop Charge Questions

Appendix 1: Agenda

Appendix 2: Participants and Observers

Introduction

This report covers discussion and material from the Department of Energy (DOE) Systems Biology Knowledgebase workshop held on March 23, 2010, prior to the 5th Annual DOE Joint Genome Institute (JGI) User meeting. The focus of this knowledgebase workshop was to discuss scientific objectives and challenges for data handling and knowledge integration specific to the study of microbial communities or metagenomes. The topics also included some discussions and items pertinent to all development and initial implementation of knowledgebases for the broader biological community.

A brief table of contents for this report is provided above. First, there is a background summary of the purpose of the DOE Systems Biology Knowledgebase planning project. Next is a summary of several topics presented and discussed during the workshop. Many of these topics require more discourse than could be fully covered during the meeting itself. Several groups and individuals were assigned to elaborate on these topics for inclusion within the report. These expanded topics are in the next section but directly refer to topics in the preceding section. For example, in the discussion of science objectives, having illustrative examples of workflows for the study of microbial communities was desired. Finally, there are appendices containing the participants list and agenda.

Prior to the workshop, participants were asked to consider the following **charge questions**:

1. What are key experimental and computational next steps that build on the sequencing data and information provided by JGI and that are feasible for an initial Knowledgebase implementation associated with research in microbial communities?
2. What types of data and information are currently available or required to accomplish these objectives?
3. How are these research goals hindered by an inability to access and integrate data from various sources or of other types?
4. What are the bottlenecks in bioinformatics and computational algorithms that need to be addressed to accomplish these goals? Specifically, is there a benefit to closer collaboration between sequencing analysis and downstream analysis?

As part of the Knowledgebase planning project, DOE is sponsoring a series of community workshops to establish the requirements for the Knowledgebase and to outline a plan for implementing them. Previous meetings include the following, and the output from each is available online at www.systemsbiologyknowledgebase.org/workshops.

1. **Knowledgebase workshop at the Supercomputing conference** in Portland, Oregon (November 2009). Explored the potential for applying the cloud computing approach to systems biology research.
2. **Joint USDA-DOE Plant Genomics Knowledgebase workshop** at the Plant and Animal Genome meeting (January 2010). Addressed the Knowledgebase requirements necessary for developing data capabilities for plants.

3. **DOE Genomic Science Microbial Systems Biology Knowledgebase workshop** at the DOE Genomic Science Contractor-Grantee (PI) meeting in Crystal City, Virginia (February 2010). Outlined workflows and data integration methods pertaining to microbial sciences that can inform Knowledgebase specifications and requirements.

Since the goal of the Knowledgebase planning project is to develop an initial prioritized plan for a useful systems biology knowledgebase, there is a continued consensus that these initial efforts cannot be all things for all users. It is better to show strong success in a few areas than minimal progress in many areas. That this needs to move forward is also reflected in the standards discussion below. Having too broad an approach has stymied and slowed past efforts.

2. Background

The Department of Energy Genomic Science program, within the Office of Biological and Environmental Research (BER), supports science that seeks to achieve a predictive understanding of biological systems. By revealing the genetic blueprint and fundamental principles that control plant and microbial systems relevant to DOE missions, the Genomic Science program (genomicscience.energy.gov/) is providing the foundational knowledge that underlies biological approaches to producing biofuels, sequestering carbon in terrestrial ecosystems, and cleaning up contaminated environments.

Knowledgebase Vision and Background

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. To manage and effectively use this ever-increasing volume and diversity of data, the Genomic Science program is developing the DOE Systems Biology Knowledgebase—an open, community-driven cyberinfrastructure for sharing and integrating data, analytical software, and computational modeling tools. Historically, most bioinformatics efforts have been developed in isolation by people working on individual projects, resulting in isolated databases and methods. An integrated, community-oriented informatics resource, such as the Knowledgebase, would provide a broader and more powerful tool for conducting systems biology research relevant to BER's complex, multidisciplinary challenges in energy and environment. It also would be easily and widely applicable to all systems biology research.

In general, a knowledgebase is an organized collection of data, organizational methods, standards, analysis tools, and interfaces representing a body of knowledge. For the DOE Systems Biology Knowledgebase, these interoperable components would be contributed and integrated into the system over time, resulting in an increasingly advanced and comprehensive resource. Other elements of the Knowledgebase vision are defined in a March 2009 report (genomicscience.energy.gov/compbio/) based on a DOE workshop that brought together researchers with many different areas of expertise, ranging from environmental science to bioenergy. The report highlights several roles the Knowledgebase will need to serve.

3. Topics Discussed at the Workshop

This section attempts to briefly summarize the wide-ranging discussion during the meeting. Where there appeared to be a general consensus, this is indicated. The level of discussion detail was not the same for all topics, and thus the level of detail in this report is uneven. Many of the topics were assigned to participants to develop further details after the workshop for inclusion in this report. Discussion of science objectives and the resulting workflows (Section 3a and related Sections 4a and 4b) was the primary focus of the meeting.

3a. Proposed Science Objectives for Microbial Community Analyses

The earlier 2009 report summarized needs and visions for knowledgebases. Here we are challenged to define precise science objectives: What do we want to accomplish in the science now? These prioritized science objectives will be a mix of priorities for importance and for current feasibility. Most of these objectives will require the exchange of data and insights (i.e., knowledge). To drive this interoperation, the Knowledgebase must have challenge problems that require cooperation and integration. A number of science objectives were described and discussed at the workshop. More potential objectives were gathered from online input to the charge questions. It will be obvious that there are common themes within objectives articulated in the report and in the earlier workshops. However, there are some unique aspects with respect to metaomics, or microbial community studies.

Some of these unique aspects with respect to microbial community studies are:

- *Massive amounts of data.* There will be terabytes of data resulting from genomic sequencing and increasingly from other techniques.
- *Datasets that never "close."* Unlike a genome for a microbial isolate, one can never finish—more data will just provide deeper details and resolution without reaching an inherent endpoint.
- *Experimental protocols will continue to develop and rapidly change.* An example is the increased application and development of RNA sequencing technologies.
- *All studies are studies of populations.* Even a species within a natural community must be considered ultimately as a population of genetic individuals that will change and evolve.
- *Natural communities are closely linked to their environmental context.* Unlike a laboratory study, this environment will not be controlled and must be observed. Despite the best available knowledge to capture the most important measurements, these observations will be incomplete. This provides a serious metadata challenge.
    - Note: Metadata is the associated data and information that provides context for the primary dataset. For example, a microbial community is analyzed for its metagenome by 16SRNA (the genomic sequences are the primary dataset). The metadata would be, for example, the location, time, environmental conditions, method of genomic isolation,16SRNA.

Four science objectives for initial study of microbial communities were proposed and discussed during the workshop. These were broadly affirmed as valuable by workshop participants. However, these and the expanded list were not prioritized during this meeting. The prioritization of these and other objectives will be a primary goal of the final Knowledgebase workshop in June 2010. The objectives discussed were:

- Metagenome analysis workflows
- Genome-based prediction of culture conditions
- Linkage and feedback from transcriptomic and proteomic data to gene calls
- Expanding metabolic pathways from metabolomic data and linking to other datasets

Metagenomic analysis workflows were seen as important in both this workshop and the one held in conjunction with the Genomic Science PI meeting. This workflow discussion has been given its own section below (Section 4a). One example of this challenge problem is that the first phase in analyzing a metagenome is done at one site, export to the binning into analysis of organisms at another site, exporting for pathways analysis at another site, followed by regulatory analysis at another site. This would drive interoperation and connections between the different groups, resulting in great science. More participants liked this collaborative model, but some preferred an approach where analysis tool needs are identified *a priori*, the tools are developed and distributed via the Knowledgebase, data is analyzed using those tools, and feedback is provided to the developers.

The need to develop expanded workflows relevant to the science community studying the microbial communities was recognized at the PI meeting and at this meeting. A small group was assigned to work offline on describing such workflows—both present and needed. Their effort is almost a stand-alone report and is presented in Section 4b and briefly summarized below. *The recommendations from this sub-report should be expanded upon to create a more detailed initial guidance in the final workshop.*

From the perspective of the metagenomics community, the DOE Systems Biology Knowledgebase will need to fulfill a range of requirements to achieve the research community's envisaged goals. These include:

- Providing a common mechanism for collecting, organizing, annotating, analyzing, and distributing data that enables easy data sharing and comparative analyses.
- Facilitating ***dynamic*** interconnection of data types, data sources, applications, and workflows to allow ***data integration*** for biological insight.
- Enabling researchers to identify, assess, and access all relevant ***datasets*** worldwide.
- Allowing scientists and facilities to "publish" their data, applications, and workflows into the "live data network."
- Providing space for larger-scale data integration, analysis, and publishing.

- Providing scientifically accepted rewards for researchers who "publish" well-annotated, good quality data, applications, and workflows.

We suggest that this could be achieved through the development of:

- A set of community-accepted semantic description formats (ontologies)
- A peer-to-peer based system of data, metadata, ontology, analysis tools, and workflow registration repositories that are integrated in discovery, access, and utilization through common semantics.
- Guidelines and software libraries that allow scientists and facilities to "publish" their data, applications, and workflows into the Knowledgebase in a set of agreed forms.
- A mechanism that allows scientists and facilities to easily and rapidly annotate, change, and correct research results and annotations in the Knowledgebase, capturing source, reason, quality, and proof for changes.
- User-friendly interfaces (APIs and people) to access data and application modules, ***as well as*** derived data products, enabling other users to build novel solutions with the data.
- A framework of citable, unique identifiers for data, applications, workflows, and researchers.
- Guidelines, training, and workshops for all new products and concepts provided by the Knowledgebase.

Genome-based prediction of culture conditions. Here the challenge is: Using a partial single microbial genome found within microbial communities, can we predict how to cultivate (and isolate) this target species? Put another way, can we predict culture conditions from genomic information? This Knowledgebase tool will be very valuable in rapidly culturing currently "unculturable" isolates from microbial communities. This would expand the study of difficult-to-culture or new microbes with interesting properties. This could lead to better integration or new experiments where one could envision testing 500 isolates a day to achieve a goal of studying newly discovered organisms with unique properties faster and cheaper.

For example, if the genome identifies heterotrophic metabolism features, will this organism grow on lactate? Is it an auxotroph, or will it require some amino acid supplement? This tool would tell you what experiments are necessary to test the proposed metabolism hypotheses. Further development of this concept would be needed including: What aspects of this tool could be automated? After the success or failure of the initial experimental cultivation tests, what information should come back to you? How do you incorporate knockout data, and can you predict the effects of knockouts? This becomes a capability tools and challenge for both the informatics and experimental communities.

The prediction of culture condition is the initial goal, but this scientific objective can be seen as the first step to a broader scientific goal in the area of genome-based functional prediction. This high-level goal would move knowledge from genetic information (which is more and more easily available compared to other data) into molecular or protein function, then to organismal function, and on to community function. It is complexity across scales. These studies are a prerequisite for investigating the function of both microbes and microbial communities. At a higher level, this would also provide potential data to feed back into improved annotation and validation. However, as stated in many other objectives, the consensus among workshop participants was that this initial effort will move most rapidly if used to address a specific problem. Each of these objectives would require detailed workflows to be developed.

Linkage and feedback from transcriptomic and proteomic data to gene calls. This scientific objective is a subset of the broader need to improve gene calls or annotation. The higher-level needs to move annotation beyond simple homology inferences were well described throughout the 2009 Knowledgebase report. The challenge here is using the massive amounts of data from transcriptomic and proteomic measurements to improve gene calls. This data is already used in the most straightforward manner—to promote gene calls from hypothetical to putative when a transcript or protein signature is observed. However, even this use does not often extend beyond the specific metagenome or genome under study. We need to find ways to draw further functional confirmations to improve gene calls, to invalidate and correct false calls, and to provide better descriptions for use in further homology searches.

With the rapid improvement of techniques such as RNA sequencing, it is clear that transcriptomic data for metagenomic communites soon will not be limited by the current requirement for an *a priori*—determined metagenome for that community. This will also enable better proteomic data analysis. This will require improved cluster analysis and the inference of pathways and function. Localization data from parts of the community (such as using laser dissection to gather small samples) will be needed to create estimates of community structure and function.

Expanding metabolic pathways from metabolomic data and linking to other datasets. There is a clear, if sometimes difficult, path from genomic to transcriptomic and proteomic datasets. Each is linked by the underlying gene. There is a different challenge in taking metabolomic data and validating and expanding metabolic pathways, as well as linking these pathways to the proteins and regulation. Since metabolites are pathway oriented, not genome oriented, the challenges of metabolomics will be largely similar, whether dealing with single microbes, communities, or plants. A related issue and challenge is extending metabolite concentration data into flux estimates. Due to tightly controlled multistep pathways, key intermediates can be present at very small levels, while the flux through that intermediate is large. With metabolomics, thousands of metabolites might be detected. However, there may be no final answer, and the dynamic range issue can confound the depth of analysis (concentrations can range from mM to single molecules). On the positive side, while there are thousands of metabolites potentially present, most experimental research targets, particular processes, or pathways (with the identification and quantification of tens of metabolites) are all that is

needed. Still most metabolomic techniques are untargeted (i.e., they try to measure everything).

Challenges here include the positive identification of detected metabolites. For example, in the synthesis of lignocellulosic biomass, there are many similar compounds such as sugar isomers. The gold standard is the purification of synthesis of a compound for use as a standard for identification. As identification libraries continue to expand, do we need to save raw data to allow later identification of metabolites from saved data?

Another challenge is to link confirmed metabolites with the measured proteins that catalyze that reaction. (Note that this requires the correct functional identification of the protein.)

Clustering, visualization, and other tools are needed to extract insights from metabolomic data. We need to have these both for microbes and for observing the change of function within a community. This is needed to determine how the rest of the microbial community environment influences the pathways of member organisms and how they utilize their genetic potential. These tools should also highlight apparent "gaps" in pathways where either metabolites or enzymes do not appear to be present. This can help identify needed experiments to fill in important pathways.

Other potential science objectives have been proposed from several other sources. Workshop participants were reminded to return to the broad objectives in the 2009 Knowledgebase roadmap. There were also a number of potential science objectives suggested in response to the charge questions and posted by participants on the Knowledgebase wiki site (www.systemsbiologyknowledgebase.org). We are continuing to extract these objectives and will place them in the final report in Section 4f.

3b: Standards: The Role of Standards-Setting in the Knowledgebase

Standards to expedite data and file sharing are important. Gene sequence data is relatively established as a standard. mRNA expression (MIAME) and other standards are being developed. However, participants had a range of opinions on the priority of standards (i.e., when do we focus on the standards?). Historically, standards development committees by community consensus have taken a very long time, and there is a need for this effort to move faster. Part of this long duration is driven by the desire to make the standards do all things for all people and uses. For example, required metadata lists quickly become wish lists of all possible information. There have also been "dictatorial" attempts at setting standards. These can lead to frustration as they are outgrown, such as in the file formats used for annotation for the last decade. Nevertheless, at a minimum, there was agreement in the need to have some standards for file-sharing formats to expedite transfer (I/O protocols). On the other side, there is the sense that if we do the needed work, the standards will sort themselves out. If the data exists, and there is a need to share, "someone" will create a protocol for sharing, which in effect is a small *de facto* standard. The challenge here is that this leads to duplication and balkanized tools. Within the context of this workshop, the range of consensus was narrower after the discussion. *Standards are important, but standards-setting is not the first task or top priority of building a*

Knowledgebase community. This workshop, the developed workflows and the final workshop report need to focus on science needs and what the initial Knowledgebase version 1 will do. If some standard setting is required as part of this implementation, it can be addressed at that time. There was an agreement that this group not be distracted into spending time in the actual standards discussions. Beyond the need for I/O, it was not clear that major effort was required in standards-setting in the first year or two. Broadly, the first two years of the Knowledgebase should focus on implementation data and tools to enable specific science.

This I/O-focused approach is re-emphasized below in the API interface discussions and workflows. There is a minimalist view that standards are actually formalized file formats, but the discussions of required metadata move beyond that interpretation.

3c. Tool Builders and Data Generators: The Need to Engage the Various Scientific Communities

Another consensus was on the goal of knowledgebases. The Knowledgebase will enable better understanding and interpretation by the "experimental" biologist and will enable testing and development of new analysis tools by the computational biologist. This reaffirms the goals stated in the 2009 report and showcases two critical science communities essential for the Knowledgebase: (1) the computational biologist or bioinformaticians who build the tools and (2) the systems biology data generators who design and run the experiments and usually provide initial interpretations. Both need to provide insights and inferred knowledge to each other through the Knowledgebase then out to the broader scientific community. This concept is presented at a high level in the Fig. 1. A challenge for both groups is the need for confidence versus just information. This was well articulated as: "I'd rather have less data but be more confident that the data is "real. I'd rather see less data with higher quality." This data would be used to create processed interpretations, like the calling of a gene. This is a challenge in assessing quality and confidence in the sea of data. For example, it is hard to assess and utilize negative experimental data because publications release only what worked. Elsewhere, frustration was expressed at the loss of underlying information when the data is processed. For example, more information goes into the calling of a gene than is saved in BLAST (i.e., intermediate analysis is lost). Also, the identification of a protein from three peptide fragment hits will lose the possible post-translational modification data hidden in an MS spectra from a "missing" peptide fragment. The combination and cross-correlation of multiple datasets from different sources into a synthesizing computational analysis struggle with different qualities of data and unreported conditions. An example recent work shows that errors in genome annotation are propagating.[1]

[1] Schnoes, A. M., S. D. Brown, I. Dodevski, and P. C. Babbitt. 2009. "Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies," *PloS Computational Biology* **5**(12), e1000605. doi:10.1371/journal.pcbi.1000605.

Fig. 1. Relationship between the DOE Systems Biology Knowledgebase and the Larger Scientific Community

There are two communities that must be both served and enabled by the Knowledgebase. One focus needs to be on the biologists sitting at their computers having done an experiment, who want to understand their results. Another focus is enabling tool builders. All agreed that this is not an ever-growing but static archive. *It is a combination of new* experimental data and tools that accesses a growing reference data. By having common access to quality data, tool-builders will also have the transformations of the data products in one place. This should accelerate the evolution of transformations and provide a better process for designing new data products. Some innovative ideas in this arena were suggested. These included tools registries, challenges and challenge grants to answer "tools needs," and Facebook-style entries of "my experiment" to advertise. There is a more detailed section on how a tools registry might function (see Section 4c). A vision is to improve data analysis sufficiently that experimental sample generation becomes a bottleneck, despite the massive amounts of data generated per experimental sample.

This can start with well-defined workflows leading to a mapped interface with access to data and tools. Then we can string tools together to do powerful operations without having to worry about the data formats and come up with an answer. This answer may be a better

interpretation or a better designed experiment. This may link into web services and other models. This interface or API should be practical, with not much investment early on.

An early priority should be to develop initial APIs to get to the data with an interface to the tools. The API should make it easier to develop and test new tools for biologists to use and to add them into the interface. There was broad consensus on the importance of the API—that it should be modular and interchangeable. An open question for the architecture and the data transfer challenges relates to how much analysis is done where the data resides versus at the viewer's site (download and I/O concerns).

There is also the need to consider architecture that relates to the massive data transfers that we are potentially considering, especially in a federated or distributed approach. This can be considered as the datasets increase; it is more and more difficult to perform "all versus all" comparisons. A number of web-based applications for metagenomics exist that do not currently support large-scale sequence analysis, including, but not limited to, on-demand clustering of user-provided datasets. Thus, the evolution of centralized data repositories and analytical services in metagenomics is currently not in sync with the accumulation of next-generation sequence data as it relates to end-user capabilities. To put things in perspective, consider the initial ScalaBLAST calculation in which comparing 1.6 million proteins in the IMG 1.6 database against the 3.2 million proteins in the nonredundent database consumed approximately 5 years of CPU time. Individual investigators simply cannot achieve these calculations due to technical or infrastructure limitations, and even if they could, the visualization tools needed to interpret and compare next-generation metagenomic and metaproteomic datasets do not scale with data volume and complexity. While good progress has been made in developing tools to inventory and, to a lesser extent, to compare microbial community structure and function, there is no comprehensive tool that allows integrating and comparing multimolecular datasets (e.g., DNA, RNA, protein, and metabolites), which are needed to fully realize the vision of microbial systems ecology.

There is continued consensus for a federated model. However, this federation cannot be the current system of separate unconnected sites. Here, federated means distributed resources, data, and tools but integrated and coordinated in a manner to be apparently seamless to an outside user. Some very mature examples are the current genome data repositories (e.g., GenBank), which actually are distributed in three sites (the United States, Europe, and Japan) but appear as one to the science community. Of course, reaching this level of integration will take a long time and effort and is beyond an initial plan. The use of a federated model brings with it the underlying challenge of how much centralization is required in deposition, curation, or "advertizing." (Note: "advertising" was discussed as a possible mechanism to draw attention to new datasets or tools in the ongoing development of the Knowledgebase.) A possible consensus in this group was that this does not matter as long as the access and the goals can be accomplished.

The development of an API allows the potential of an "open-source" system. The potential and challenges of "open" systems are discussed in more detail in Section 4d.

This use and development and data deposition in knowledgebases must be balanced with the need for some level of public/private embargo and the need to further the careers of

bioinformaticists and experimentalists. This was deemed important and is covered in more detail in Section 4e.

3d. NIH Interactions: Data Resources and Leverage

There was discussion about the need for awareness, linkage, and leverage with NIH-led efforts, in particular NCBI. Current and planned NCBI efforts are described elsewhere. Workshop consensus was that we should leverage resources as much as possible. In particular, we should use both existing and under-development NCBI capacities as an archive and repository as much as possible. But there will always be a gap in filling current BER Genome Science needs and challenges, therefore we will also need our own efforts and to link them with other projects.

4. Expanded Discussion from Workshop: Assignments

4a. Workflows and the Systems Biology Knowledgebase

In bioinformatics, complex biological analyses frequently require large-scale computations that compose standard tools and methods into a pipeline, or workflow, that runs a series of tasks to achieve a specific outcome. There are two major types of workflows, namely:

1. **Ad hoc Interactive:** In ad hoc interactive workflows, the biologist is fundamental in driving the steps involved in the workflow. Interactive tools (e.g., Cytoscape, DMV, R scripts) are used to analyze and visualize data, and the results from one tool become the inputs to the next tool in the workflow. The biologist typically drives the transition between tools based on his or her observations of the state of the analyses, and data is moved between tools either manually (e.g., saving files in specific formats) or by using a lightweight data transfer tool like Gaggle (www.systemsbiology.org/Technology/Data_Management/Gaggle).

2. **Automated:** Automated workflows, also called pipelines, take a set of input data and apply a series of analyses to the data to produce outputs. No human intervention is necessary to invoke the next step in the workflow and to transfer data between computations. Automated workflows can take anything from seconds to weeks to execute, and the steps in the workflow are commonly controlled by scripts or workflow tools like Taverna (www.taverna.org.uk/).

While the precise software mechanisms used to coordinate the steps in a workflow vary between the interactive and automated cases, the ease of construction of workflows in both cases is hampered by two fundamental technical issues:

- **Tool heterogeneity**: Standard tools and algorithms are not created using a common software framework so that they can be readily "plugged together" to form a workflow.

- **Data heterogeneity:** Standard tools and algorithms consume and produce data in a variety of different data formats. Feeding the outputs from one tool into another commonly requires data transformations to produce inputs in a format that a given tool is expecting.

For these reasons, creating effective bioinformatics workflows is non-trivial and requires considerable effort from biologists and software engineers alike in order to meet scientific objectives.

The Systems Biology Knowledgebase is an opportunity to address the current complexity of creating both interactive and automated workflows. The Knowledgebase can create a lightweight, flexible software infrastructure that enables tool developers to "componentize" their existing and new algorithms, providing standard interfaces that can be used to compose tools into workflows. In addition, the Knowledgebase infrastructure can support the flexible, discoverable definition of data formats that tools produce and consume. By describing a given tool's data requirements using metadata, converting data from one tool to another becomes simpler, and potentially automatable.

We therefore recommend the Knowledgebase implements a set of simple programming interfaces that enable much more effective workflow construction and reliable execution. By reducing the "levels of pain" experienced by biologists and software engineers in creating workflows, we envisage the creation of a software ecosystem in which useful workflows can be rapidly built, deployed, and shared with the community through the Knowledgebase infrastructure. This would be analogous to social networking sites such as Facebook, which encourage development and sharing of new applications based on the software infrastructure and programming tools that Facebook makes available. This model, which is expanded upon in the next section, is designed to (1) encourage development of many tools that provide multiple approaches to solving a particular problem and (2) enable the end-users to determine which approaches survive. Applications that accurately solve a problem in a particularly elegant or succinct manner will become highly adopted, and others will slip into oblivion.

A primary issue to be addressed under the Knowledgebase plan is the motivation of the developer. Platforms like Facebook, Twitter, or the Apple App Store can provide a financial incentive for *de novo* application development. Knowledgebase infrastructure and early applications will need to be developed under more conventional models. But as the programming platforms become established, the project will need to consider funding models designed to maintain and expand innovation over the long term. This may include a combination of standard funding models and models designed to reward *de novo* application development. Failure to address this basic issue will almost certainly result in stagnation of the development cycle.

4b. Metagenomics Systems Biology Knowledgebase: Workflows, Background, Design Goals, and Recommendations

Integrating Metagenomic-Enabled Workflows

Most metagenomic data come from microbial ecosystems. Data derive from a broad range of environment types—from the deep subsurface to the human gut—motivating many questions such as how are microbial communities structured, and how do they function? Do genetic profiles vary across environment types? Using metagenomics to answer such questions will

require the effective integration of information about metabolic potential (genomic sequence); metrics for function (proteomics, transcriptomics, metabolomics); contextual information; data that define the physical and chemical environment (metadata); and methods to consistently and accurately update annotation as new evidence becomes available (see Fig. 2). Metagenomic data may be collected from one or many samples, whereas proteomic, metabolomic, and transcriptomic data typically stem from a diversity of experiments such as time series, environmental perturbation, and genetic manipulation.

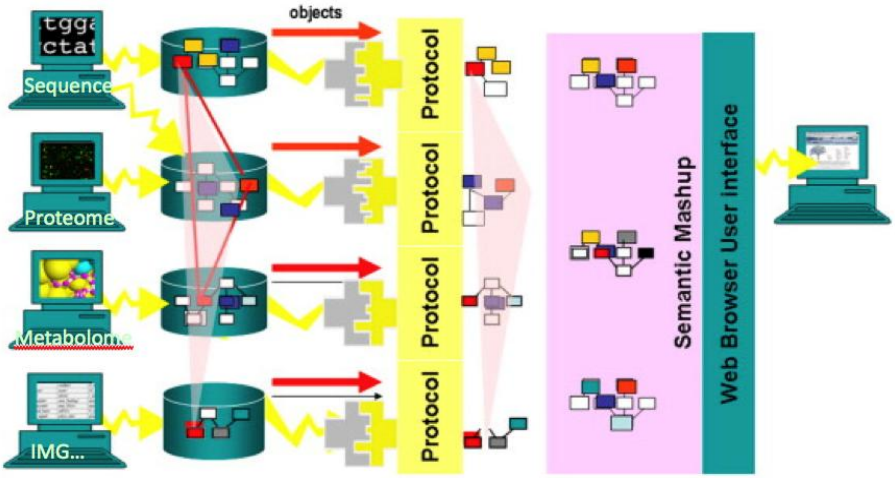


Fig. 2. Data Warehouse. Data are linked by common or discovered identities and shared annotations (tags) drawn from controlled vocabularies, and managed by identity and ontology authorities. The data are accessed through simple application programming interfaces (APIs) and aggregated through browser scripting based on common identities and tags. [Slightly modified from Goble and Stevens 2008.]

One of the most important aspects of metagenomic investigations is that sequence information is (or will be) intimately linked with proteomic, metabolomic, and transcriptomic data. Commonly, metagenomic workflows begin with sequence information, but a Knowledgebase—an artificially intelligent tool that provides a mechanism for collecting, organizing, analyzing, and distributing data—must be designed to facilitate dynamic interconnection of these data types to allow data integration for biological insight (see Fig. 3). The need for dynamic interconnection is underlined by the observation that data can exist in many states: "there is live data, living data (more live than live), stale data (archived?), dead data (archived?), lost data, vandalized data (valid data overwritten by non-valid data)."[2] For example, when data consumers download a specific dataset from a resource and put it into a new form for their own purposes, the data become disconnected from the original source in the absence of dynamic linking or a provenance system. It therefore cannot benefit from changes or upgrades in the source (i.e., it becomes "stale" or "dead"). Examples of changes at the source include reassignment of a gene function, re-searching of a proteomic spectra database with new genomic sequence, and changed identification of a metabolite due to the addition of new, standard metabolite profiles to reference databases. These types of data insertions, deletions, and mergers represent problems for all subsequent users of a resource.

[2] Web commentary, www.ted.com/talks/tim_berners_lee_on_the_next_web.html

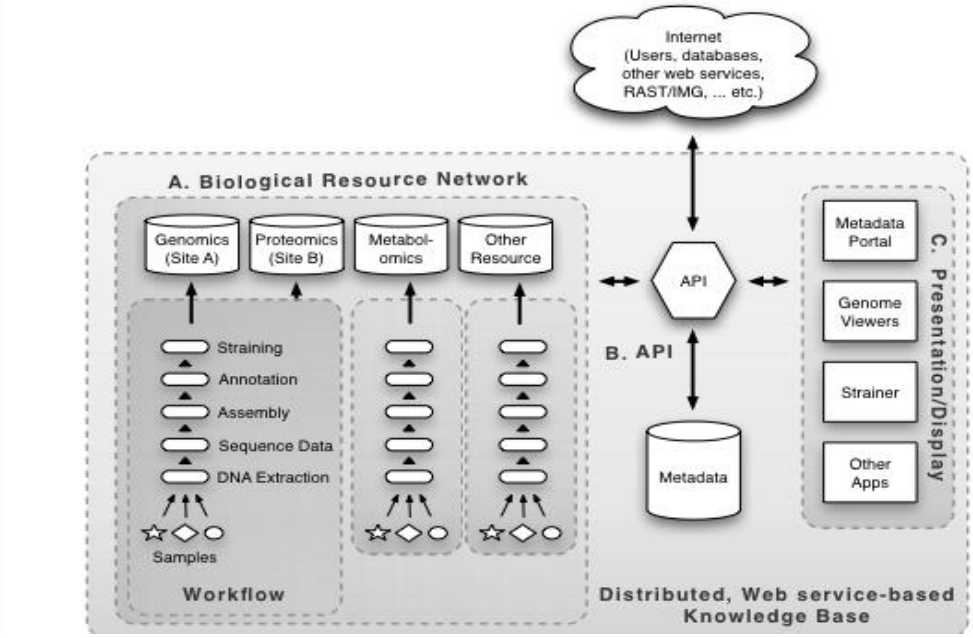


Fig. 3. Overview of Metagenomic-Enabled Workflows Feeding into a Knowledgebase System. This figure provides examples of data-generating experiments, data types, file formats, and processing steps. Note that not all studies include other "omics" methodologies. Data derive from a common source (a natural sample or series of samples, an experiment or series of experiments), and data are integrated via tools to answer specific questions. Lines with arrowheads represent data flow, cylinders indicate a data resource, and rectangles indicate an application (stand-alone or web-based). Data analyses can draw on a wide variety of existing and newly developed tools (e.g., assembly programs, gene prediction, functional annotation, analysis of regulatory structure), as well as tools developed specifically for metagenomics (e.g., example methods to visualize and analyze strain variation).

Metagenomic Data Challenges

Five key attributes associated with metagenomic data pose challenges that require special consideration.

(1) Volume of Data Generated. Sequencing of metagenomes generates somewhat to highly fragmentary datasets, often with low redundancy levels and potentially high error rates (due to low genomic coverage with error-prone sequencing).

The growth in sequencing capabilities has led to a flurry of metagenome sequencing and analysis projects in recent years. Interestingly, computational analysis costs are now quickly outpacing data generation costs (Goble and Stevens 2008). As shown in Fig. 4, running similarity searches (BLASTX) for data generated by one run on an Illumina GS-FLX instrument (costing approximately $15,000) will take 60,000 hours of compute time on a recent machine (or cost approximately $120,000 if run on Amazon's EC2 service).

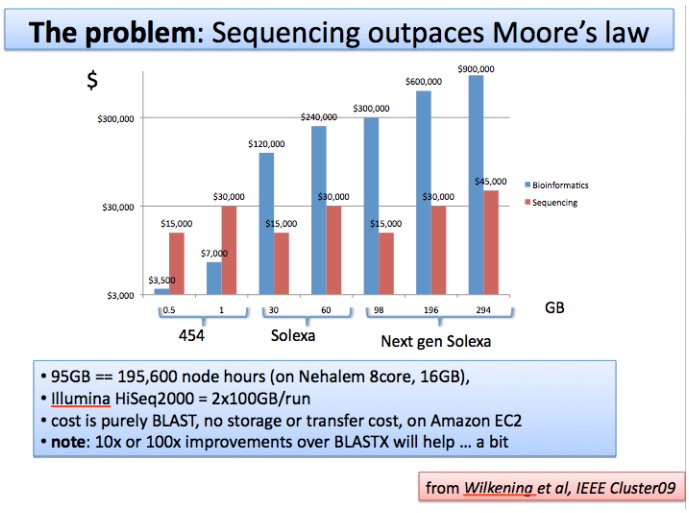


Fig. 4. Cost Comparison. The computational costs (blue) are rising significantly with growing sequencer yield. Already they can be up to 10 times the sequencing cost (red) on some instruments. Costs shown are for running BLASTX against the current National Center for Biotechnology Information (NCBI) non-redundant protein database. [From Goble and Stevens 2008.]

An associated problem is that while hundreds of metagenome datasets are publicly available, these are widely distributed across the world, and it is at times challenging to identify and access all relevant datasets. Many more experimental data sources are not publicly discoverable. Therefore, few high-profile studies have emerged that attempt the systematic comparison of available public data because this task is difficult (and, in part, because currently the primary focus of most investigators is on their own datasets). One way of enabling both general discovery and access (e.g., via web services, as well as "local" access to the data) is via a peer-to-peer based framework of dataset registries. Unique identifiers such as Digital Object Identifiers (DOI) known from publications, URLs, or persistent URLs (PURLs) could enable the citation of high-quality, well-annotated datasets, and offer identification and linkage between datasets. Any such framework also would encourage metadata and semantic enrichment, enabling queries such as "Display all soil metagenomes from the Midwest" or even "What is the number of short-read metagenome datasets currently available from a specific sequencing platform."However, without accompanying metadata, the sequences may be next to useless for some of the broader purposes.

(2) Never-Ending Analysis. With metagenomic data investigations, the analysis is never finished. Unlike an isolate genome where a final set of open reading frames and relatively stable functional annotation is generated, metagenomic datasets are prone to continual revision resulting from new sequence data, additional manual curation, new reference genomes, and more. Gene numbers, for example, may change multiple times, and so can organism assignments of genome fragments. Although curated metagenomic datasets share many features with isolate datasets, there is a major difference: there is no single answer in almost all cases because populations are not clonal cultures. In many current studies, the extent of metagenomic sequence curation is minimal, but this must change. A Knowledgebase system must be ***dynamic*** in order to be able to deal with this key attribute.

(3) Sequence Variation. One of the most important features of a metagenomic dataset is the sequence variation that is captured via the sequencing reads. Typically, each sequencing read derives from a different individual, and thus a metagenomic dataset provides information about sequence heterogeneity. Although it is arguable whether or not the primary sequence data must be retained (or just the base-calling scores), ***a system to tie read-based data to the composite, assembled sequence is essential***. In addition to providing access to such data, an annotated and permanent record of research steps taken during curation will be essential to establishing a powerful, relevant, and long-lasting scientific Knowledgebase. To our knowledge, this capacity is not available via any shared resource. Sequence-variation analyses will be critical to population genetics and evolutionary studies, for efforts to identify the reasons for fine-scale variation in functional attributes, and to locate potentially interesting gene variation for targeted bioengineering applications.

(4) Tools and Computational Infrastructure Are Required for Data Sharing and Comparative Analyses. As the number of sequencing and "omics" instruments available for metagenomic research grows, ever increasing the volume of available data, the community needs both the computational infrastructure to analyze and compare the data as well as the tools to analyze the results.

An important approach to tool development is the generation of a series of modular components (as opposed to large, integrated pipelines suited to run on a centralized computational facility). In addition to portability, an advantage of the "components"-based approach is that the user retains considerably more flexibility with regard to the way in which the data are processed. An example of current interest involves software for correcting homopolymer errors in 454 sequences. Currently, this capability is contained within a complex package within a pipeline. For practical reasons, it is undesirable to send the entire dataset to a staff member at the centralized facility for homopolymer correction, and, more importantly, data reprocessing in a new pipeline will unlink information already associated with the sequence.

(5) Diversity of Data Types, File Formats, and Processing Steps. Scientific research itself has become more specialized on the individual level and more collaborative and international on the community level, making it desirable and necessary to relate one's own local research results meaningfully to the geographically distributed, multifaceted results being compiled elsewhere in the world. Systems biology has a long tradition of utilizing diverse research results from experimental and computational methods that stem from varied and distributed sources. Commonly, these research results are locally integrated and synthesized with the scientist's own findings, then published as yet another valuable source of information. Over the years, the community has created a wealth of outstanding data sources and tools for access, integration, and analysis. Unfortunately, these ***sources of scientific knowledge and analysis are mostly characterized by a diversity of data formats, data representation, metadata, and access methods*****,** making it difficult to identify all of the relevant data sources for a given topic, assess their quality, and integrate them into the scientific research process.

Architecture Design Goals and Knowledgebase Adoption Strategies

As experiences in other scientific communities have shown, it could take many years to change working practices and move to a central deposition system based on community-agreed, standard data formats, metadata, and semantic data descriptions, with associated discovery and analysis tools. This type of integration has been most successful in slower-moving fields with less diversity in their research methods than metaomics, fewer data sources, and more standard analysis software. In these fields, such integrations have been very successful in terms of making data more widely known, used, and effectively analyzed by the community. Despite the larger challenges faced by our community, such central deposition systems have their place in the Knowledgebase, but it appears unlikely that the more rigid structures of a central data and applications repository could effectively meet all systems biology needs. Instead, a more flexible approach is needed. We advocate combining the benefits of the more rigid, standardized frameworks with a "live data network" of shared experimental results.

Ideally, the live data network component of the Knowledgebase would be compatible with current, more distributed working practices, while at the same time assisting with greater integration of resources. A peer-to-peer based system of data, metadata, analysis tools, and workflow registration repositories, integrated through common semantics would seem desirable. In this scenario, users could "publish" their data, applications, and workflows into the live data network by describing their provenance, content, location, access, and usage methodologies (both for other users and computer applications) in one of a set of community–agreed semantic description formats (ontologies). In designing these semantic descriptions and underlying metadata, it will be important to focus on the data content and ontologies. Important concepts that must be included, rather than the particular local implementation, need to be identified. Semantic ontologies will allow the mapping between different expressions of the same concepts (within reason), as well as linking concepts where their expressions do not overlap but are related.

If a "local" format contains desired information, it is relatively simple to map these data to the shared semantic description and make the local knowledge available in a community format. These conversion interfaces are not difficult, but they can be time consuming to establish. Therefore, offering different levels of community participation would be desirable. Initially, one might submit only enough information to allow others to discover one's resource. More functionality can be added over time to support full integration. Alternatively, central data centers might offer services to smaller groups to integrate their data into the archive and "publish" it for them (including data analysis, such as normalization and filtering, annotation, and more). This type of tiered approach to Knowledgebase participation is critical because it would help to bring about adoption by the wider community by supporting the twin (and sometimes competing) goals of (1) defining relevant data standards and formats for participation in a more centralized repository and (2) the necessity of allowing dynamic integration to be done at a smaller, local, and scientific inquiry-driven level.

Similar to the underlying resources (see Fig. 5), links between people, data, applications, and workflows could be explicitly recorded and published, or discovered following the semantic description "trail." This approach is analogous to the LinkedData Web proposed by Tim Berners Lee but is extended with more domain-specific information about the data, especially the

applications and workflows, to aid directed scientific discovery and experimentation. An added benefit could result if the Knowledgebase provided space for larger-scale data integration, analysis, and publishing—following the same format described above—to aid smaller institutes and preserve important results after funding for their further curation elapses.

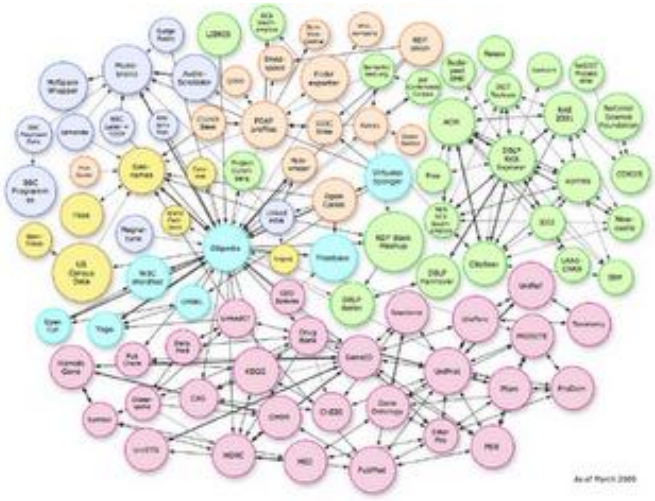


Fig. 5. Linking Open Data. This cloud diagram gives an overview of published datasets and their interlinkage relationships.

When designing the community-agreed metadata and semantic formats to be utilized, advice from the data curation community should be considered. Their aim is the development and provision of methodologies and tools that allow the perpetual reuse of data by its designated user communities, by keeping it "living" (i.e., adapting its representation to changing community trends without affecting its integrity). Data policy examples, data life cycle models, archival reference models, and more will help to define critical components of the data, analysis software, and workflow descriptions.

The Knowledgebase's success will depend strongly on the quality of the data, tools, and workflows that are available through it, as well as the ability of researchers to identify, assess, and integrate resources quickly. Much of the latter will depend crucially on the quality and extent of the metadata and semantic information about the resources. Herein lies a potential problem. Although good data or tools will usually lead to direct scientific rewards in the form of publications and resulting citations, the time-consuming annotation and preparation of sharable datasets is not directly rewarded by the community in similar fashion. This often results in a lack of motivation to provide this vital information, or to provide it with the required due diligence. One way to resolve this issue might be to enable scientists to "publish" their datasets, following similar style, content, and peer-review standards that they are required to follow throughout the publishing world (e.g., see DataCite at www.datacite.org/). This way, datasets would be citable and earn the owner well-deserved recognition and rewards.

Similarly, to encourage participation in the Knowledgebase, it will be critical to ensure that utilized data, analysis tools, and workflows are correctly attributed to their owner, an often difficult task due to similar names and changing names and affiliations. Recently, the research

community has initiated more concerted efforts to develop unique identifier systems for researchers, although most of them are focused on publications (e.g., International Standard Name Identifier (ISNI) being developed by the International Organization for Standardization (ISO) as Draft International Standard 27729, and Open Researcher Contributor ID (ORCID) being organized by the ORCID Initiative) and not data elements. Additionally, protocols for maintaining permanent data identifiers are being developed and increasingly deployed in the biological domain. The combination of Uniform Resource Names (URNs) and Uniform Resource Locators (URLs) are one example. URNs represent a persistent, location-independent identifier, and they promote mapping to other namespaces. A URL is the specific location of a URN. For example, a specific protein name and its annotation from the Kyoto Encyclopedia of Genes and Genomes (KEGG) can be represented by both a unique identifier and a location. Using the Life Science Identifiers system (lsids.sourceforge.net/), a URN in the Knowledgebase might be "urn:lsid:doekb.gov:eco:b1743", which uniquely names a specific gene in *E. coli*. The location of the data element is accessed using the URL www.genome.jp/dbget-bin/www_bget?eco:b1743. This combination of URN and URL provide an unchanging name (the URN) and a location (URL) of data about the URN. Using a common naming system allows for data linkages and for the development of rich semantic descriptions. Another identifier system is the persistent URL system, or PURL. This system achieves the same goal of specifically naming and locating data, but it does not directly describe the location. Instead, it references an intermediate location that redirects the request to the proper location. PURLs depend on a master system for redirection and on contributors to maintain their links. As an example, the same *E. coli* protein described previously can be accessed from the UniRef Database using a PURL system from this link: purl.uniprot.org/uniprot/P77754.

Using unique identifiers (DOI, URN/URL, or PURLs) for data, and potentially applications and workflows, will make it possible to utilize the services the library and web technology communities offer for information discovery and access. Furthermore, different data publications and publishers could be easily linked through citations, as could the history and connectivity of data elements in the Knowledgebase.

Some Data Sharing and Analysis Needs

As data sharing is becoming more common, data storage is essential, but as with the World Wide Web, it need not be centralized. As noted previously, repositories must be dynamically linked to information sources; otherwise, we run the risk of "stale data." To enhance research, each repository should consist of user-friendly interfaces (APIs and people) to access data and application modules, as well as derived data products, enabling other users to build novel solutions. One such product provides genomic neighborhood views across multiple genomes. Both the DOE Joint Genome Institute's Integrated Microbial Genomes (IMG) system and Argonne National Laboratory's Rapid Annotation Subsystem Technology (RAST) offer these views (see Fig. 6). However, such foundational capabilities need to be extended to meet real systems biology needs. For example, a user may need to integrate proteomic, metabolomic, or transcriptomic data in a display such as that shown in Fig. 6. Semantic data interfaces could provide vital support in this endeavor, as they insulate tool developers (including those of user-friendly interfaces and data products) from differences and changes in local data formats, data organization, and access mechanisms. Based on the community-agreed semantic descriptions

(ontologies), any software can search and request data using semantic concepts; for example, "Give me the first protein of the second metabolic pathway that the system identified in KEGG for *Shewanella*." This type of request would still continue to work even when the underlying data sources change, as long as the data sources maintain their semantic interfaces to the Knowledgebase. Therefore, any tool development effort can focus on new functionality, rather than redundantly expending effort on data search and access methods. Similarly, it will be much easier to combine development efforts that previously would have only benefitted selected data sources. Application maintenance will again require reduced resources, as changes to underlying data sources must not affect the usability of any tools, unless they want to benefit from any additional information (new concepts, not more data in the same structure).

In general, we anticipate building upon tools such as those shown in Fig. 6 to address new needs and provide additional capabilities. For example, a user may need to answer a specific question such as the temporal distribution of strain genotypes. Thus, the "display" tool may need the ability to output specific data characteristics as inputs to other programs (e.g., the library of origin of a set of sequencing reads, or the environment type from which specific gene contexts derive). A benefit of a web service—based system is that it allows for the development of tools for such specific questions. In general, continually capturing ***innovation*** by the broader community for purpose-specific application development provides a way to address the unavoidable limitation that no program can ever meet all possible user needs.

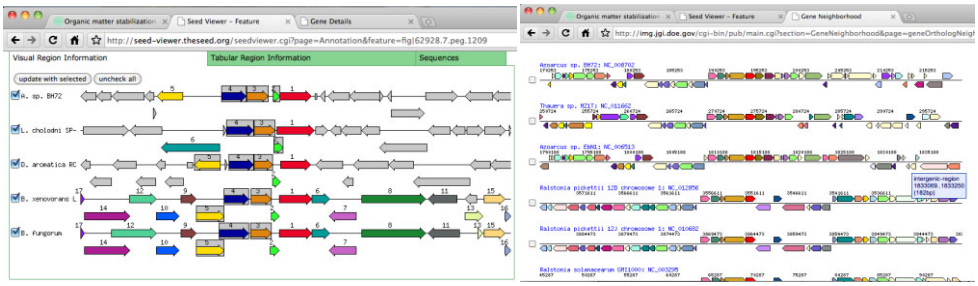


Fig. 6. Data Displays. The chromosomal context of a protein in *Azoarcus sp.* BH72 as shown by the SEED-Viewer (left) and IMG (right). Both systems provide a similar view of data that are computationally expensive.

With the growing volume of data, there is an opportunity to introduce "lightweight standards" to allow the exchange of many sequence datasets and to reduce the cost of computational analysis. The Genomics Standards Consortium (GSC) has presented the Minimum Information about a Metagenomic Sequence/Sample (MIMS) standard, which allows the exchange of contextual data for metagenomes (e.g., location, sampling method, and biome description). This provides an initial description of the sample, but only minimal information about computational sample processing is included. The GSC's M5 working group has presented a draft metagenome interchange standard (MTF) that includes computational results and MIMS metadata (see Fig. 7). However, as important as standards are, they can have a downside. It is also important to ensure that standards are flexible enough to adapt to cutting-edge advances in the field, allowing specific users to add additional information.

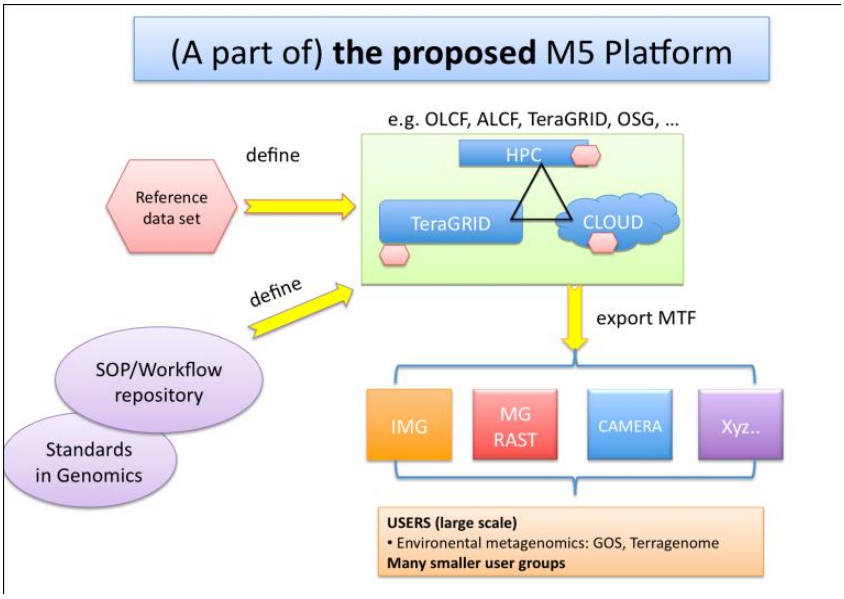


Fig. 7. Standard Development. The proposed M5 platform will include a standardized processing pipeline that can be executed by third parties (e.g., large supercomputer facilities), and, via a reference non-redundant protein database, it will enable many groups to use the results. JGI's IMG/M and MG-RAST, as well as the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA), are working on the M5 standard.

Integrating data across multiple disciplinary Knowledgebase subcomponents. An important concept for a DOE-wide Knowledgebase is that it will interlink components tailored to specific areas of systems biology investigations that have been or will be designed by those with special expertise with the various data types and needs. A possible solution is a Metagenomics Knowledgebase (MKb) developed on a database structure flexible enough to be ported to other laboratories and populated with data from any of a wide diversity of systems.

Such an MKb must enable free flow of data into and out of a larger DOE-wide Knowledgebase structure, as well as into and out of many other publically available databases (e.g., KEGG and Pride). One component will be data repositories (e.g., from a specific research group or team) with one or more data resources. Many studies will include both metagenomic sequence and other data forms (e.g., proteomic, metabolomic, or microarray data). Each data repository will have specialized tools, so that one might visualize links from an MKb to, for example, a project-specific transcriptomic dataset in the Transcriptomics Knowledgebase (TKb), or the Proteomics Knowledgebase (PKb). Developing guidelines for assisting local groups in creating resources such as these is an important goal for the DOE Knowledgebase. A system that describes and publishes data dynamically will greatly aid in this step, and incorporating the tiered participation option mentioned previously will ensure that all stakeholders are represented.

One of the most important obstacles to overcome is the lack of data integration in a form that enables data mining by groups not previously involved in specific kinds of experiments. To some extent, web-services type interfaces to the data and a comprehensive dataset registry will allow only a certain subset of queries and satisfy just a subset of researchers. "Web services" typically means an API that can be accessed over a network and executed on a remote system hosting

the service. It is usually in one of two forms: "big web services" (e.g., KEGG) and representational state transfer services (RESTful services; e.g., UniRef and GO).

Currently, "the integration requirements of biologists working with unpublished data are not being widely addressed by the community."[3] To illustrate the opportunities associated with data integration across experiment types, the different data types generated in different laboratories by different methods (e.g., metabolomic data and proteomic data) must be considered. Another researcher might be interested in a specific protein of unknown function suspected to play a role in a certain metabolism. So long as the data are accessible, patterns may emerge through the integration of this information with information from other research groups. This can be accomplished as long as the experimental groups use a consistent "resource description framework" (RDF), a data description document that describes and links the data (e.g., gene A is translated into protein B).

Both centralized, high-performance computing (HPC) services and dynamic, "local" web services can be envisioned as important coexisting and linked parts of the MKb. For example, each of the experimental platforms (e.g., metagenomic sequencing, proteomics, and transcriptomics) requires extensive HPC, but the calculations and analysis only need to be done once, and the results then are shared (e.g., by using a common data description and a RESTful service). In fact, the raw data could be made publically available upon generation, and as the local system converts raw data to processed data, the broadcast version is continually updated ("live data"). This should be feasible within the framework of the laboratories that participate in data sharing, although it is likely necessary that central data warehouses will create new releases on a regular basis.

In the specific case of metagenomics, raw sequence is generated in vast quantities. Over time, these sequences are assembled and annotated. In parallel, comparisons among reads and assembled fragments reveal within populations variation in gene sequence and gene content. A whole suite of computational tools is required to collect and present the data as part of the in-house analysis, but this work product (the added value) currently is never distributed. For mass spectral datasets, a resource description could be generated that enables a researcher to access the up-to-date analysis and download components (e.g., reads, contig sequence, and single-nucleotide polymorphism (SNP) concentrations). In this way, long-running, CPU-intensive analyses can be part of a Knowledgebase approach that allows biological data integration via web service protocols. This will accomplish the important goal of keeping those responsible for data generation and data upkeep ("live data") connected to widely distributed data analysis tasks. Furthermore, the data are continuously enhanced as the network of links to new experiments expands.

[3] Anwar, N. and E. Hunt. 2009. "*Francisella tularensis novicida* Proteomic and Transcriptomic Data Integration and Annotation Based on Semantic Web Technologies," *BMC Bioinformatics* **10**(Suppl 10:S3), PMCID: 2755824. Electronic resource.

Annotations change as our understanding of protein families and complete genomes improves over time. Work on specific genes or proteins builds a body of functional evidence around each of these entities and their families. An optimal annotation system should permit easy and rapid updating of specific sequence annotations in response to new experimental data, and provide a straightforward method to record the ***source and quality*** of the updates. Automated annotation systems should be able to use the quality data to inform new annotations and update overlapping datasets. Such a system would dramatically reduce the time from discovery to annotation and enable very richly annotated descriptions of genes, pathways, and organism function. It also would facilitate collaborative research by disparate laboratories focused on a common genetic system and reduce propagation of erroneous annotations into new datasets.

The primary problem with such a system is ensuring the accuracy of changes made by the broad research community. In the 1980s, GenBank dealt with this problem by restricting annotation changes to the individual that submitted the sequence. However, the MKb must move beyond this static model to embrace systems used by organizations such as Wikipedia to validate changes. Therefore, in addition to recommending a MKb that focuses on sharing of "live" experimental and associated data, our working group recommends finding a mechanism in which new knowledge deposited within the data warehouses (e.g., NCBI) can be updated easily.

Role of grand challenges in linking community, shaping Knowledgebase development. Transition to a scientific framework in which data sharing and data integration is facile will present many challenges. At this time, small groups are beginning to address parts of the problem (e.g., metabolomic-metagenomic-proteomic data integration on a small scale), but the effort at the "omics" community level has a long way to go. A recommendation of our working group is to motivate the formation of linkages and overall architecture of a Knowledgebase with this goal via "grand challenges" that require data integration and sharing. We find this approach preferable to tasking a group of bioinformatics experts with establishing a system that will be later used by the community.

As one example of a grand challenge, consider the potential for data integration to improve protein annotation [g1]. Currently, most genomes encode a significant number of lineage-specific proteins that have not been studied biochemically. These proteins may hold considerable significance for DOE efforts in environmental remediation and bioenergy, as they may be involved in novel pathways for metal redox transformations or degradation of complex organic compounds. Similarly, there are probably many small non-coding RNAs coded on genomes and genome fragments for which annotations are lacking in public databases. Consequently, many gene predictions are uncertain, and a significant number of predicted proteins discovered via short-read sequencing may be corrupted by frameshifts. A single confident identification of a hypothetical protein via proteomics (or transcriptomics) converts a "hypothetical" to a "protein of unknown function" (an annotation that could be amended with the words "validated by proteomics"). If all the curated proteomic data from all samples worldwide could be integrated in a single analysis, many annotations in public databases could be updated. In addition, detection of the first peptide in a protein can confirm either the start site or the truncation status of the mature protein (e.g., due to cleavage of signal peptides). High-throughput improvement of start-site information from either proteomics (or transcript

sequencing) will provide important constraints for better gene prediction and regulatory models. Such tools and models are essential for confident systems biology studies.

Recommendations

The Systems Biology Knowledgebase will need to fulfill a range of requirements to achieve the research community's envisaged goals. These include:

- Providing a common mechanism for collecting, organizing, annotating, analyzing, and distributing data that enables easy data sharing and comparative analyses.
- Facilitating ***dynamic*** interconnection of data types, data sources, applications, and workflows to allow ***data integration*** for biological insight.
- Enabling researchers to identify, assess, and access all relevant ***datasets*** worldwide.
- Allowing scientists and facilities to "publish" their data, applications, and workflows into the "live data network."
- Providing space for larger-scale data integration, analysis, and publishing.
- Providing scientifically accepted rewards for researchers who "publish" well-annotated, good quality data, applications, and workflows.

We suggest that this could be achieved through the development of:

- A set of community-accepted semantic description formats (ontologies).
- A peer-to-peer based system of data, metadata, ontology, analysis tools, and workflow registration repositories that are integrated in discovery, access, and utilization through common semantics.
- Guidelines and software libraries that allow scientists and facilities to "publish" their data, applications, and workflows into the Knowledgebase in a set of agreed forms.
- A mechanism that allows scientists and facilities to easily and rapidly annotate, change, and correct research results and annotations in the Knowledgebase, capturing source, reason, quality, and proof for changes.
- User-friendly interfaces (APIs and people) to access data and application modules, ***as well as*** derived data products, enabling other users to build novel solutions with the data
- A framework of citable, unique identifiers for data, applications, workflows, and researchers
- Guidelines, training, and workshops for all new products and concepts provided by the Knowledgebase

References for Section 4b

Anwar, N. and E. Hunt. 2009. "*Francisella tularensis novicida* Proteomic and Transcriptomic Data Integration and Annotation Based on Semantic Web Technologies," *BMC Bioinformatics* **10**(Suppl 10:S3), PMCID: 2755824. Electronic resource.

Cheung, K. H., H. R. Frost, M. S. Marshall, E. Prud'hommeaux, M. Samwald, J. Zhao, and A. Paschke. 2009. "A Journey to Semantic Web Query Federation in the Life Sciences," *BMC Bioinformatics* **10**(Suppl 10:S10). Electronic resource.

Goble, C., and R. Stevens. 2008. "State of the Nation in Data Integration for Bioinformatics," *Journal of Biomedical Informatics* **41**(5)*,* 687–93.

Wilkening, J., A. Wilke, N. Desai, and F. Meyer. 2009. "Using Clouds for Metagenomics: A Case Study," *IEEE Cluster 2009.*

4c. Toolkit Registry Development

The development and management of a Systems Biology Knowledgebase will provide a unique resource to integrate experimentation, modeling, and bioinformatics across disparate levels of biological inquiry. Foremost, the successful implementation of a modular cyberinfrastructure to service the informatic needs of the systems biology community must actively recognize the broad potential user base of the resource. This recognition is critical to populate the resource with the appropriate data, tools, workflows, and corresponding literature consistent with the expectations of the user community and commensurate with the varying expertise of the Knowledgebase clientele.

Fundamental to the success of the Knowledgebase will be the bioinformatic services and resource sharing potential of the portal. At present, non-computational biologists seeking to incorporate high-level informatic investigations into their research program do not have access to a resource analogous to that which is envisioned in the development and deployment of the Knowledgebase. Most often direct consultation with expert bioinformaticians is required to initiate such investigations, effectively dissociating experimentalists with their data in the pipeline of biological discovery. To improve the efficiency of bioinformatic investigations, a centralized resource announcing the availability and utility of the various tools, applications, and algorithms available will stand as a tremendous advance toward minimizing the opacity of "–omics"-based data analysis and interpretation. To meet this need, it is recommended that a Knowledgebase-hosted toolkit registry be developed to provide a comprehensive inventory of software available to the user community. The construction of such a registry will also assist in defining the architectural strategy of the Knowledgebase engine, including compute resources, need for portage, development of APIs for web-based analysis, and demarcation of analysis subsystems specific to particular tasks and objectives.

Below is a brief list of thematic applications and additional resources that could initiate inventory within the Knowledgebase resource:

Sequence assembly

Sequence annotation

Comparative genomics

Phylogenetics

Cluster analyses

Statistical analysis

Metadata integration

Geographic Information System (GISP deployment

Data QA/QC

Transcriptomic resources

Proteomic resources	Genetic database resources
Metabolomic resources	Metabolic subsystem database

Strategically, each of these resources should contain a brief description of the type of data or analysis provided, its product, and rationale for why a user might be interested in using a given tool or database. Importantly, the Knowledgebase must also exist as a central forum where researchers and developers can exchange ideas to nucleate new tool development. Curation and expansion of the toolkit registry should be guided through an interactive Wiki environment where users can post comments and suggestions for including new resources into the Knowledgebase. Additionally, resource examples and workflows should be included to assist users. An exquisite example of a tool registry has been developed as part of the Neuroscience Information Framework accessible at neuinfo.org/nif_tools/nif_registry.shtm.

Ultimately, content within the Knowledgebase must be adequately indexed to allow integration across the multi-dimensions of available data. Key drivers of this need include incorporating genetic, genomic, transcriptomic, proteomic, and metabolic datasets with phylogenetic, metabolic, imaging, ecological, and geospatial information. Although such considerations are beyond the scope of the toolkit registry described here, it is critical that a composite inventory of analysis tools be available to augment discovery and seed the data integration process.

4d. The Knowledgebase as an Open Development Platform

Enable tool development and integration, by providing an open developer platform inspired by the Facebook Platform / Google Apps API. Allow outside developers to produce novel analysis and visualization tools that can query the database directly (with appropriate access controls) and display and exchange results through a common UI. There will always be disagreement between research communities on which analysis is the best for any particular data type. DOE should not be in the position of enshrining one type of analysis over another. It should provide the platform, let the individual researchers develop the tools, and let the community reach a consensus.

Platform Infrastructure. The foremost task for the knowledgebase platform is to provide the user to the underlying knowledgebase data, if necessary shielding the user from how that access is achieved (e.g. federated versus centralized, cloud-based versus central server, etc.). It should also provide the user with elementary analysis and visualization tools to apply to that data, a way to store intermediate results, data standards to allow data to be exchanged between tools, and ways to chain analysis tools together to create ad-hoc interactive workflows. In addition, it should provide a low-threshold infrastructure for tool development, reuse, and dissemination.

User Empowerment and Community Collaboration. Regardless of the size and quality of the Knowledgebase development team, there will inevitably be more developers, talent, and ideas (not to mention time to implement) “outside” than “inside.” We should aim to leverage the talent within the Knowledgebase user community to develop and choose the best tools. Many novel bioinformatics tools suffer from a “failure to launch,” never reaching beyond the initial

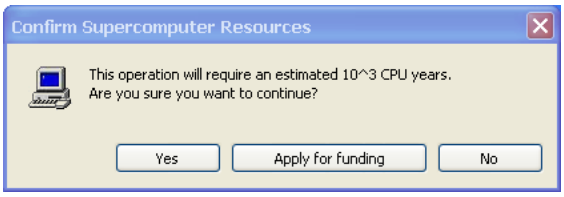
journal publication, due to a lack of a web-based implementation or lack of marketing skills of the developer. By enabling tool developers to integrate their tools with the Knowledgebase platform and tie directly into its user interface, we can expose a wider variety of tools to a wider variety of users and enable more users to become tool developers themselves.

Components, Scripts, and Open-Source Development. Individual tools may be as simple as calculating the GC content of a DNA sequence or displaying a matrix of numbers as a heatmap. More complex tools can then be constructed by combining these elementary components, piping data from one tool to another—similar to unix shell scripts that are composed of elementary data manipulations such as "grep" and "sort," with some control logic to pipe data between them. At the extreme end, entire processing pipelines could be encapsulated into a single tool, calling upon dozens of other analysis and visualization tools. By making these tools open source, we can enable even relatively novice programmers to tinker with and improve upon them (e.g. swapping out one statistical test for an improved one or adding a novel visualization tool).

Reputation, Attribution, and Credit. If we open up tool development to the world, some mechanisms are needed to enable the community to disseminate, vote, and prioritize the highest quality tools. The reputation of a tool may be based on various factors, including usage statistics (how many other tools incorporate this tool and how frequently is it actually called), direct votes by users, and the reputation of its developer. Developer reputation in turn would depend primarily on the reputation of the tools they have contributed. Community reputation may be a powerful incentive for contribution, especially for junior members. A credible mechanism for attribution and credit could potentially also be used to drive funding and even tenure decision for tool developers, on par with journal impact factors.

Some important issues which need to be resolved:

Computing Resources. Some tools can easily be run on the user's own computer, some should be run on the server side because of higher CPU or storage requirements (e.g., BLAST against NR), others may require substantial high-performance (or cloud) computing capabilities. How do we throttle processes to achieve an equitable distribution of resources? How do we keep users from making an expensive mistake? How should we deal with a "poorly behaved" tool? Can we estimate a priori (e.g., based on previous usage statistics) how much computing power a specific tool will require? How do we fund additional computing time on this system? Can users simply buy more compute power on the cloud?



Incentives for Contribution. How do we encourage an active and vibrant developer ecosystem? Some of the larger components such as the Knowledgebase platform and early applications will need to be developed under more conventional funding models. But as the programming platforms become established, the project will need to consider funding models designed to maintain and expand innovation over the long term. Significant attention should be paid during the design phase of the Knowledgebase platform to social engineering and design of interactions between tool developers and scientists. How do scientists share and evaluate

tools? How can we leverage existing networks of interaction to enhance community buy-in and involvement?

A suitable pilot project for an open Knowledgebase development platform might be to mirror part of an existing genome database (e.g. IMG, MicrobesOnline, SEED, etc.), implement rudimentary tool development infrastructure, make this platform accessible online, and then invite an external user to write a novel analysis or visualization tool not already found in the original database.

Sidebar: the Facebook Platform

Facebook released its "Facebook Platform" in May 2007, enabling users to "build the next generation of applications with deep integration into Facebook, mass distribution through the social graph, and a new business opportunity." The Facebook experience has shown that this is an excellent way to involve the community in the development of the platform. Users immediately jumped on the opportunity and started generating little tools and widgets—sometimes in direct competition with tools Facebook had already implemented. The Platform provides multiple integration points for apps to integrate seamlessly into the existing Facebook user interface. Many Facebook apps turn out to be useless or poorly designed and disappear into obscurity, but some are absolute hits and propagate rapidly throughout the community, resulting in far more high-quality tools than the Facebook developers could ever have implemented themselves. As of June 2009, two years after introduction of Facebook Platform, Facebook reported 350,000 active applications from over 950,000 developers. A significant part of the Platform infrastructure itself was open-sourced in 2008, and it is possible that some pieces of this could be leveraged, although the needs for a Knowledgebase platform are likely to be very different than for a social networking site like Facebook. Note, however, that the underlying Facebook database is much larger than existing genomic databases and has orders of magnitude more users and hits.

4e. Institutional and Career Considerations Surrounding Open-Source Development

This section describes some considerations with regard to institutional technology transfer philosophy and the career impact of open-source software development and use in the DOE BER Knowledgebase program.

Science assumes a clear account of methodology that is repeatable by others. Open source provides a definitive account of methods where software was used in analysis. Open source has become widely practiced in federally funded research. An Open Source Policy would be at least in keeping with the spirit of recently proposed legislation to provide free open access to all federally funded research within 6 months of publication. Open source does not have to mean immediate release. The Open Source Policy could be similar to the Data Release Policy where there is some allowance for limited access before being made public. At a glance, open source seems like an obvious choice for the Knowledgebase, but there are real issues associated with making all software immediately open source. Key issues are discussed below:

(1) Experience has shown that conflict arises between the open-source concept and the desire by home institutions to license IP in the process of facilitating technology transfer. This has in the past involved situations where the home institution licensed software to companies and

earned revenue from software developed under DOE grants. This issue can arise within universities, national laboratories, and private industry. While resolvable, this will need to be addressed. In the open-source scenario, where software is publicly available, licensing efforts are defeated, and the contractors performance and thus award fee depends partly on their success with technology transfer. This will depend in part as to whether the software has unique value or is merely routine in nature. While an open-source policy might seem in conflict with an expectation of technology transfer placed on institutions, such policy is intended to encourage rapid transfer of IP to provide a basis for new business development and benefit society. An open-source policy would accomplish that objective.

(2) The DOE Genomic Science program has a strong history of doing bioinformatics research—that is, developing new algorithms or solving the hardest computational problems in new ways. Examples include gene finders, transcription binding predictions, protein structure prediction, protein dynamics and function, metabolic pathway simulations, and large-scale cellular process modeling. To a researcher in bioinformatics, such products represent the publishable results of research, and such investigators have a right to publish their algorithms and their performance in the literature prior to open release in a manner similar to experimental scientists. As we work toward large infrastructure and in many current national laboratory SFAs (Science Focus Areas), we see the role of bioinformatics changing in part from research to that of programming and operational support, for example in building databases or websites for projects. It is clear that we have or perhaps should have two classes of bioinformatics tasks: (i) publishable research, which develops new algorithms or methods for key problems, and (ii) infrastructure support and development, which is likely to be much less publishable and where methods are more likely to be mature. The infrastructural element is analogous to core facilities for major experimental capabilities such as sequencers.

Fig. 1 illustrates the potential collaborative nature and continuum of interests and capabilities across the scientific community between the pure experimentalists and pure computationalists that become evident in the context of the Knowledgebase as a platform for future research that brings different groups and communities together. Publishable tools development is an aspect of research, while the infrastructure development is more linked with the development of the Knowledgebase itself. Infrastructure development and deployment are much more amenable to immediate open-source standards, with rewards to such individuals much less likely to be publications or novel results. A concern is that in large projects and in the push toward open source, we forget that the research mission in bioinformatics (tools development) is very important and the types of individuals that do such research are vital to the DOE Genomic Science program. There needs to be a strong research activity to generate solutions for next-generation problems in bioinformatics. We need to identify proper incentives for both paths and encourage top people in both for sustained careers. Ultimately both are required, working together, to attain the ambitious scientific objectives of the future.

4f. Other Potential Science Objectives and Knowledgebase Features Drawn from Responses to Preworkshop Charge Questions

These potential science objectives were drawn from the online responses to the charge questions. This section needs further expansion and revision for the final report. The development of these lists from this and prior workshops will speed the community's process of identification and then prioritization of the set of scientific objectives that will be developed in much greater detail needed for the final workshop and ultimately for the final report—the Knowledgebase Implementation Plan.

Regarding data quality and annotation:

- Use data quality indicators.
- Use experimentally verified data only.
- Create a clean computing system using some model organisms with only experimentally verified data.
- We need the ability to update annotations in genome sequence repositories.
- Need statistical correlations of datasets.
- We need standard quantitative approaches for dealing with the different data types for normalization and assessment of statistical significance.

Regarding microbial community omics data integration, we need the following capabilities:

- To integrate short read sequence data (Illumina data) and proteome data.
- To routinely integrate all omics data for every newly sequenced organism to minimally include: RNA sequences, proteomics, metabolic phenotype (Biolog) profile. Move beyond gene-based annotations to pathways.
- Comprehensive tools that allow integrating and comparing multimolecular datasets, which are needed to fully realize the vision of microbial systems ecology.
- To integrate sequencing data and downstream analysis into a common analytic pipeline that enables end users at different skill levels to interrogate their data in interactive ways in real time. Controlled vocabularies or ontologies to leverage metadata across different organisms or samples.
- To link genotype with biogeochemistry or biogeography.
- Under defined conditions, to compare proteins expressed in related strains of bacteria to predict metabolic potential of microbial communities and resolve physiological differences; use this information to identify biomarkers diagnostic for specific biogeochemical processes. Intercompare spectral libraries to identify unique peak profiles as new gene models are added to the protein database.

Appendix 1: Agenda

DOE Systems Biology Knowledgebase Workshop

Walnut Creek, California
Tuesday, March 23, 2010

8:30 a.m. – 8:40 a.m.	Welcome, Susan Gregurick
8:40 a.m. – 9:00 a.m.	"Introduction to Knowledgebase Initiative and Workshop Objectives" Bob Cottingham
9:00 a.m. – 9:30 a.m.	"Science Presentation on Metagenomics: Current Experience and Future Expectations for the Knowledgebase" Phil Hugenholtz
9:30 a.m. – 10:00 a.m.	"Science Presentation on Metagenomics: Current Experience and Future Expectations for the Knowledgebase" Jill Banfield
10:00 a.m. - 10:30 a.m.	Panel Summary and Audience Questions
10:30 a.m. – 11:00 a.m.	Break
11:00 a.m. – 12:30 p.m.	"Informatics Perspectives and Roundtable: How to Transition from the Present Towards an Open, Shared, Integrated Knowledgebase"
	Discussion Leads: Adam Arkin, Folker Meyer, Ed Uberbacher, Nikos Kyrpides, Peter Karp, Tatiana Tatusova, Victor Markowitz, Bob Cottingham
12:30 p.m. – 1:00 p.m.	Working Lunch
1:00 p.m. – 1:30 p.m.	"Presentation of Metagenomic Workflow Example" Jill Banfield
1:30 p.m. – 3:00 p.m.	Panel Discussion, Jill Banfield
3:00 p.m. – 3:30 p.m.	Break
3:30 p.m. – 4:30 p.m.	Panel Discussion, Jill Banfield
4:30 p.m. – 5:00 p.m.	Conclusions and Adjourn, Bob Cottingham

Knowledgebase Wiki: sites.google.com/a/systemsbiologyknowledgebase.org/kbase/

Appendix 2: Participants and Observers

Participants

Adams, Paul (LBNL)

Allen, Eric (University of California, San Diego

Allgaier, Martin (LBNL, JBEI)

Anderson, Gordon (PNNL)

Arkin, Adam (LBNL)

Baker, Scott (PNNL, JGI)

Banfield, Jill (University of California, Berkeley)

Benton, David (University of Wisconsin)

Bhaya, Devaki (Stanford University)

Bowen, Ben (LBNL)

Bownas, Jennifer (ORNL)

Brettin, Tom (ORNL)

Bristow, Jim (LBNL, JGI)

Broughton, Jeff (LBNL)

Canon, Shane (LBNL)

Chen, Amy (LBNL, JGI)

Chivian, Dylan (LBNL, JBEI)

Collart, Frank (ANL)

Cottingham, Bob (ORNL)

Davison, Brian (ORNL)

D'haeseleer, Patrik (LLNL)

Drell, Daniel (DOE BER)

Foust, Cheri (ORNL)

Gorton, Ian (PNNL)

Gregurick, Susan (DOE BER)

Grossman, Arthur (Stanford University)

Hallam, Steven (University of British Columbia)

Heidelberg, Karla (University of Southern California)

Hess, Matthias (JGI)

Hugenholtz, Phil (JGI)

Jansson, Janet (LBNL)

Karp, Peter (SRI International)

Kleese van Dam, Kerstin (PNNL)

Kodner, Robin (University of Washington)

Konstantinos, Mavrommatis (JGI)

Kosky, Anthony (JGI)

Kuske, Cheryl (LANL)

Kyrpides, Nikos (JGI)

Land, Miriam (ORNL)

Landick, Robert (GLBRC, University of Wisconsin-Madison)

Liolios, Konstantinos (JGI)

Lykidis, Thauos (LBNL, JGI)

Mansfield, Betty (ORNL)

Markowitz, Victor (LBNL)

McCue, Lee Ann (PNNL)

Mead, David (Lucigen Corporation)

Meyer, Folker (ANL)

Muyzer, Gerard (Delft University of Technology)

Palaniappan, Krishna (LBNL)

Pletcher, David (LBNL)

Raymond, Jason (University of California, Merced)

Richmond, Kathryn (GLBRC, University of Wisconsin)

Sczyrba, Alexander (JGI)

Slater, Steven (University of Wisconsin)

Stepanauskas, Ramunas (Bigelow Laboratory for Ocean Sciences)

Swingley, Wesley (University of California, Merced)

Szeto, Ernest (LBNL)

Tatusova, Tatiana (NIH)

Thomas, Brian (University of California, Berkeley)

Tringe, Susannah (JGI)

Uberbacher, Edward (ORNL)

Wang, Zhong (LBNL)

Woyke, Tanja (JGI)

Acronyms

ANL	Argonne National Laboratory
DOE	U.S. Department of Energy
GLBRC	Great Lakes Bioenergy Research Center
JBEI	Joint BioEnergy Institute
JGI	Joint Genome Institute
LBNL	Lawrence Berkeley National Laboratory
NIH	National Institutes of Health
NREL	National Renewable Energy Laboratory
ORNL	Oak Ridge National Laboratory
PNNL	Pacific Northwest National Laboratory

DOE Knowledgebase System Development Workshop Report

June 1–3, 2010, Crystal City, Virginia

Workshop Organizers: Susan Gregurick (DOE) and Bob Cottingham and Brian Davison (Oak Ridge National Laboratory)

Table of Contents

1. Introduction

2. Background

3. Pre-Workshop Activities

- 3a. Conference Calls
- 3b: Templates
- 3c: Google Docs
- 3d: Goal to Establish Scientific Objectives and Requirements

4. Topics Discussed at Workshop

- 4a. Microbial Science Objectives
- 4b. Metagenomics/Meta-Communities Science Objectives
- 4c. Plant Science Objectives
- 4d. Computational Area Breakouts: System Architecture, Implementation Plan, and Governance

5. Post-Workshop Plan

Appendix 1: Agenda

Appendix 2: Participants and Observers

Appendix 3: Scientific Objectives Template

Appendix 4: Requirements Template

1. Introduction

This report reviews discussion and material associated with the Department of Energy (DOE) Knowledgebase System Development workshop held June 1 - 3, 2010. The goal of this workshop was to establish initial actionable plans to create the Knowledgebase. The first day focused on the prioritization of clear scientific objectives and specific requirements for the Knowledgebase derived from these objectives. The second day focused on the development of an implementation plan, system architecture, and governance for the initial system. The last day focused on finishing writing assignments leading to the Final Report, which will be the plan for creating the Knowledgebase.

First, a background summary is given below describing the purpose of the DOE Systems Biology Knowledgebase (Kbase) planning project. Next is a summary of pre-workshop activities, topics presented and discussed during the workshop, and post-workshop activities.

Since the goal of the Knowledgebase planning project is to develop an initial prioritized plan for a useful systems biology knowledgebase, there is a continued consensus that these initial efforts cannot be all things for all users. It is better to show strong success in a few areas than minimal progress in many areas. There was also continued consensus on the principles from past workshops on which Kbase is being founded that (1) science drives Knowledgebase development, (2) the project be a community effort, (3) that it be open access and open contribution, and (4) that it be distributed.

2. Background

The Department of Energy Genomic Science program, within the Office of Biological and Environmental Research (BER), supports science that seeks to achieve a predictive understanding of biological systems. By revealing the genetic blueprint and fundamental principles that control plant and microbial systems relevant to DOE missions, the Genomic Science program (genomicscience.energy.gov/) is providing the foundational knowledge that underlies biological approaches to producing biofuels, sequestering carbon in terrestrial ecosystems, and cleaning up contaminated environments.

Knowledgebase Vision and Background

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. To manage and effectively use this ever-increasing volume and diversity of data, the Genomic Science program is developing the DOE Systems Biology Knowledgebase—an open, community-driven cyberinfrastructure for sharing and integrating data, analytical software, and computational modeling tools. Historically, most bioinformatics efforts have been developed in isolation by people working on individual projects, resulting in isolated databases and methods. An integrated, community-oriented informatics resource, such as the Knowledgebase, would provide a broader and more powerful tool for conducting systems biology research relevant to BER's complex, multidisciplinary challenges in energy and environment. It also would be easily and widely applicable to all systems biology research.

In general, a knowledgebase is an organized collection of data, organizational methods, standards, analysis tools, and interfaces representing a body of knowledge. For the DOE Systems Biology Knowledgebase, these interoperable components would be contributed and integrated into the system over time, resulting in an increasingly advanced and comprehensive resource. Other elements of the Knowledgebase vision are defined in a March 2009 report (genomicscience.energy.gov/compbio/) based on a DOE workshop that brought together researchers with many different areas of expertise, ranging from environmental science to bioenergy. The report highlights several roles the Knowledgebase will need to serve.

Workshop Background

To develop a successful open informatics endeavor for systems biology (the Kbase effort), a series of workshops have been held to include key stakeholders (plant and microbial genomic researchers, bioinformaticians, computer scientists, database developers, and software engineers) and to elicit their goals, challenges, and expectations for the development and management of the Kbase. This final workshop was a culmination of these previous workshops to provide clear prioritization and tasks to allow the final design and implementation of the Kbase to be developed. The workshop was held June 1–3 and involved 80 participants representing university, national laboratory, and international researchers. In addition, the workshop had representation from DOE's Joint Genome Institute; DOE's Bioenergy Research Centers; NSF's iPLANT; and NIH's NCI and NCBI. The goal of the workshop is to develop a robust design and implementation plan for the Systems Biology Knowledgebase. The participants were charged with developing and prioritizing 3 to 5 scientific objectives in each of the areas of microbial, meta-communities, and plant research. From these scientific objectives, two days were spent developing scientific requirements, time frames, and effort for each of the scientific objectives. An additional half day was spent discussing detailed plans for architecture, implementation, and governance. Extensive pre-meeting conference calls helped to lay the groundwork of the science objectives. Participants were not charged to define funding or contractual structures, and they are continuing to finalize requirements based on the discussed objectives and transfer these into an implementation plan.

Outlined below are the prioritized scientific objectives and rough time frames for the implementation of these objectives. The details of the requirements of each objective as well as the architecture and governance plans will be developed over the summer, culminating in a final implementation plan report by September 30, 2010.

3. Pre-Workshop Activities

3a. Conference Calls

A series of conference calls were scheduled in May before the workshop. The first of these were to organize the three science area breakout leads. Each science area (Plant, Microbial, and Meta-Communities) had two leads for Scientific Objectives and two leads for Requirements. Then calls were held with all members of each breakout. The first call was to introduce the workshop and define what is meant by a Scientific Objective, and the second call focused on reviewing the Scientific Objective template and beginning discussion on what would be the recommended Strawman List of Scientific Objectives for each breakout. At the workshop, the Strawman List would be reviewed and finalized on the first day, and participants would establish the consensus priority (High, Medium, Low) and feasibility (Near: 1-3 years; Mid: 3-5 years; Long 5-10 years). Based on this, the top 3 to 5 Scientific Objectives would be the focus of the initial Kbase.

A third call introduced the template for Requirements and how these would be derived from a Scientific Objective. The most detailed and complete Requirements write-ups are needed for the top Scientific Objectives, with decreasing detail needed for mid- and long-term Objectives.

3b. Templates

The Scientific Objectives and Requirements templates—along with filled-out examples that were given to the participants—are included in Appendix 3 and 4, respectively. These provide focused guidance toward establishing the most important objectives and detailed requirements that guide Kbase development.

3c. Google Docs

In order to begin rapid development of the Scientific Objectives, a Google Docs folder was established for each breakout group. Writing teams then formed around each proposed Scientific Objective, and a significant amount of preliminary writing was accomplished in advance of the meeting. During the conference calls, multiple participants would edit the draft documents as they were being discussed. Initially, these areas were accessible only by members of the breakout. At the workshop all areas were made accessible to all participants.

3d. Goal to Establish Scientific Objectives and Requirements

For most attendees, this was a different kind of workshop. Its primary focus was on establishing the best Scientific Objectives and Requirements for the DOE Systems Biology Knowledgebase, especially the high-priority requirements for the first 1-3 years. The Requirements are the most important result of the workshop, as these define what the initial Kbase will be. The Science Area breakouts first focused on the Scientific Objectives, and then in the second half of the first day, the Breakout Leads switched and the focus was on Requirements with the same Breakout group. This process allows an easier transition from objectives to requirements and encourages feedback so the objectives are tractable.

4. Topics Discussed at Workshop

4a. Microbial Science Objectives

1. Integrated Description of Genomic Features

Summary: This objective will create the ability to represent and update experimental data and inferred knowledge about genes and genomes so that the experimental and computational results drive progressively richer and more accurate gene models and predictions. This ability would allow users to access existing genomic sequence information, upload new experimental data in order to define and refine models, and test consistency between the two. This objective was given high priority, as many other objectives require this ability to build on. This objective requires integration with JGI/IMG and NCBI and will require some standards development for data and access to large-scale computing resources. This objective will take 1-3 years.

2. Reconstruction, Prediction, and Manipulation of Metabolic Networks

Summary: The scientific objective is to provide a method to evaluate the metabolic potential of an organism, predict the phenotypic outcome of specific metabolic or environmental interventions and perturbations, and establish metabolic kinetics capabilities and fluxes for short-term dynamic responses. This knowledge will lead to the informed

modification of one or more specific enzymes or the introduction of entirely new enzymes and/or pathways for metabolic engineering purposes. This objective would allow the community to better determine strategies for carbon flow manipulation and for understanding microbial communities. This objective requires integrating new experimental data with known data and models on metabolic pathways, as well as developing methods to automatically create new metabolic reconstructions from newly sequence organisms. This objective requires linking together known metabolic models with databases such as chEBI, UniPROT, KEGG, and GO with experimental data. This objective is given medium priority (3-5 years), and it is suggested to apply this objective to a selected set of organisms relevant to DOE's research efforts.

3. Microbial Gene Expression Regulatory Networks

Summary: The scientific objectives can be broadly divided into two components. The first is to enable automated inference of gene expression regulatory networks relying principally on expression profiling data. The second is to extend these inferred networks to include additional data types, both to refine the network predictions and to test them. The availability and evolution of genome-scale expression data and the rapid extension into new data types makes the definition of microbial gene expression regulatory networks an attractive goal of the Kbase project. In the short-term, inference of regulatory networks from just genome sequence and expression profiles under varied cellular conditions is possible and could be of general utility to researchers in constructing and understanding of carbon and nitrogen processes. Interconnection of the regulatory networks with metabolic reconstructions and multidimensional annotations (two other high-priority objectives identified by the Kbase microbial group) would greatly facilitate development of microbial systems biology. This objective could work synergistically with NIH pathway tools, EcoCYC, and DOE efforts such as MicrobesOnline and JGI. Much of the experimental work would come from the Bioenergy Research Centers and the larger DOE science-focused work on microbial systems. This project was given high priority. This objective can achieve some near-term goals but may take 2-10 years to complete. It was suggested to work on DOE-related organisms and in coordination with the second scientific objective.

4b. Metagenomics/Meta-Communities Science Objectives

1. Metabolic Modeling of Microbial Communities

Summary: This objective focuses specifically on modeling the metabolic processes within a microbial community, since this topic most directly ties into developing metagenomics workflows and the single microbial organisms systems biology tools (above). This predictive understanding of communities will progress in three stages: **(1) Understanding**: Descriptive models that provide insight into the metabolic role of the members within the community and their interactions. **(2) Prediction**: Predictive models that allow us to simulate the metabolic processes in the community and the response of community activity or composition to environmental conditions. **(3) Manipulation**: Eventually, these models will allow us to not only predict, but actively drive changes in the community into desired directions (e.g., to accelerate environmental processes such as bioremediation, cellulose

degradation, or carbon sequestration). This objective outlined as a first step the Knowledgebase need to develop workflows for analyzing metagenomes of a microbial community and to leverage existing data to create community metabolic models. This objective was seen as a medium priority (3-5 years) and would require leveraging existing tools (IMG, MG-RAST, CAMERA) and databases (BioCyc, KEGG) as well as developing analysis tools.

2. Expand Our Understanding of Poorly Studied Genes

Summary: Data generated in large-scale metagenomics projects can provide the information necessary to better understand the function of poorly characterized genes. This scientific objective is to develop approaches for (1) mining the data in order to identify previously unknown genes and (2) leveraging the wealth of metadata associated with metagenomic datasets, as well as gene/organism co-occurrence information in order to identify testable hypothesis about the function of newly identified or poorly characterized genes. This objective was given high priority and could leverage all of the metagenomic sequencing efforts from DOE and NIH.

3. Analysis of Understudied Microbial Phyla

Summary: The goal of this objective is to understand the role of unclassifiable members of a microbial community in terms of genetic and phenotypic comparison. To achieve this scientific objective, a specific requirement will be linking physiologic and metabolic datasets to metagenome annotations in order to provide context and evidence. This will create a product that is more informative and flexible. The specific datasets that will be utilized are the genomes and accompanying physiologic and metabolic data of understudied microbial phyla. Questions that this objective would address are: (1) where are members of a novel phylum found, (2) how do we facilitate phylogenetic binning to preclude assignment as orphan genes, and (3) what are the emerging concepts of their metabolomes? This objective was given medium priority (3-5 years) and requires the development of infrastructure and tools to accomplish the goals. This will likely be merged into Objective 2.

4. Metagenomic Interpretation to Identify Conditions Required for Growth by Key Microbial Communities Relevant to DOE Missions.

Summary: Using a partial single microbial genome found within microbial communities, can we predict how to cultivate (and isolate) this target species? Put another way, can we predict culture conditions from genomic information? This will require metagenomic sequence, assembly into species genomes, and pathway analysis of these partially assembled genomes. While workflows do exist to perform some of these tasks, they will need to be developed much further and altered to make use of supercomputing facilities to handle gap-finding exercises. It is not clear if relevant tools exist, and this was given medium priority, as it will take years to develop (5-10 years).

4c. Plant Science Objectives

1. Integration of Phenotypic and Experimental Metadata to Enable Prediction of Biomass Properties based on Genotype

Summary: Improvements in computational infrastructure are required to support and contextualize experimental plant phenotype data to an extent that will enable one to predict the changes in the physical properties of biomass properties that occur as a result of environmental changes and genetic diversity or manipulation. Achievement of this ambitious goal depends on the creation of robust semantic infrastructure for collection, annotation, and storage of diverse phenotypic and environmental datasets. These data include measurements such as photographic images and analytical spectra that capture visible phenotypes and chemotypes that are fundamentally related to yield and physiological performance and sustainability. Specifically, this infrastructure will be used as a basis for software applications that extract, quantify, and catalogue phenotypic features from the data for the purpose of data mining and further analysis. This involves association of the data with relevant metadata to enable querying, modeling, clustering, and comparison of the data from diverse datasets generated by different platforms. Attainment of the scientific objective requires appropriate vocabulary standards for wide variety of data and metadata that describe phenotypes, chemotypes, genotypes, and the experiments designed to collect this data. Although several such standards and ontologies exist, they require additional expressiveness to achieve the objective. In order to share the relevant experimental data and ensure its completeness (in terms of associated metadata, etc), a community approved standard for the Minimum Information for A Plant Phenotyping Experiment (MIAPPHE) would be helpful. However, such a standard does not currently exist. The development of all of these standards demands a long-term, committed collaboration between computer scientists and plant scientists. This objective was seen as high priority and could be carried out in 3 to 5 years. This would require a community of scientists to agree to standards of data to describe phenotypes and needs to be coordinated with iPLANT.

2. Assemble Regulatory Omics Data in Common Platforms to Enable Annotation, Comparisons, and Modeling

Summary: This objective will integrate several key types of regulatory omic data and associated quality and metadata for six target plant species: *Brachypodium*, *Chlamydomonas*, poplar, sorghum, switchgrass, and *Miscanthus*. This information will support the other objectives, including annotation, comparison, and modeling. RNA levels as measured by expression arrays or RNA-Seq are no longer sufficient to evaluate mechanisms and networks that regulate plant transcriptomes. The Kbase must also include available small RNA and target RNA information, differential RNA processing and decay information, and epigenetic marks such as DNA methylation and histone modifications. This information is important for data integration and to fill in important missing links in gene regulatory networks within a species and facilitate their comparison across two or more species. In the short term (1-3 years), classical transcriptome data (microarrays and mRNA

seq) as well as small RNA and basic proteomic data will be assembled. Epigenetics data, small RNA target data/RNA degradome data, other types of RNA processing data, and additional proteomic data will be assembled after year one, with the most developed genomes such as *Brachypodium* beginning first. The data will be made publicly accessible with user-friendly web interfaces and downloadable for power users. The acquired data will include sequences, quality information (e.g., Q values) and associated metadata. Sources will include NCBI (GenBank, GEO, SRA), the DOE JGI, ArrayExpress, and PLEXdb. This was given high priority and could be accomplished in 1-3 years. This requires collaboration with iPANT and USDA for selection of relevant species.

3. Improvement and Availability of Plant Genome Annotation Datasets

Summary: Currently, plant genomes are typically annotated in isolation and with varying methods. Even more problematic is that the annotation is rarely, if ever, updated. As a consequence, annotation across genomes is not comparable, becomes stale rapidly, and frequently is of undocumented quality. Without confidence in the gene model annotations, biological interpretations will be greatly hampered, if not erroneous. The research goal is to generate high-quality, documented, uniform, and integrated annotation for plant genomes. Six target genomes have been identified (*Brachypodium*, *Chlamydomonas*, sorghum, poplar, switchgrass, and *Miscanthus*). The goal is to develop a platform that results in higher-quality annotation than what has been provided to date rather than to annotate more genomes. In the initial phase, only two genomes that are phylogentically diverse will be annotated in years 1-2. Subsequently, in years 2-3—with refinement of the platform—another two genomes will be annotated, and the platform will be further refined. In years 3-10, all genomes will be iteratively annotated to capture newly available empirical data and algorithmic improvements. This scientific objective would need to be coordinated with the 'omics data integration objectives and with DOE JGI, NCBI, iPLANT, and the plant communities. This was given high priority and would be accomplished in 1-3 years.

4. Modeling, Simulation, and Validation

Summary: Enable semi-automated inference, construction, simulation, validation, and query of complex multilevel (gene, protein, metabolite, small RNA, organelle, cell, and tissue) models of plant life, with a focus on models useful for integration and exploration of experimental data types collected during study of biomass recalcitrance, the carbon cycle, and bioremediation. Four sub-objectives proposed herein are automation and streamlining of model construction; development of a semi-automated model validation process; development of advanced semantic querying capability targeted to biological models and representations; and phylogenetic inference of functional networks (itself a model construction exercise). Model construction and validation are very closely aligned with Kbase objectives. Exploratory model construction is completely dependent on a conceptual framework, together with multiple datasets (annotated genome, proteomic, metabolomic, transcriptomics) to populate instances of this framework. Validation depends on well-structured and -annotated experimental data. At the same time, the dependencies are modular, which facilitates separate development of software for specific or more

generalized tasks. Semantic query will enable scientists to more rapidly and precisely develop hypotheses and conclusions from the complex metabolic and regulatory models that arise from genome-scale studies. This science objective requires interfacing with existing plant genomic databases as well as KEGG, GO, Metacyc, PMN. This was given high priority but was also noted to take up to 10 years in stages.

4d. Computational Area Breakouts: System Architecture, Implementation Plan, and Governance

On the second day of the workshop, a follow-up set of breakouts was held to address the major topics associated with constructing the Knowledgebase computation system. The System Architecture group is working to establish the technical principles and basis for recommending specific System Architecture, considering specific architectural attributes and their relative priorities. The Implementation Plan group is evaluating each Scientific Objective and associated Requirements to assess the major tasks and recommended plan for implementation. The Governance group is considering and will recommend a governance model and principles that will guide the development, management, and operation of the Knowledgebase for the research community. Based on the Kbase vision, principles, and scientific objectives, each of these groups is working toward writing up recommendations for the associated sections of the Final Report.

5. Post-Workshop Plan

Each breakout topic group is finalizing its write-ups with a June 30 deadline. The focus has been on completing the Scientific Objectives and Requirements and then on integrating these into a Science Area report that will become part of the Final Report. In parallel, work is under way to create the Implementation Plan section for each of the Scientific Objectives that typically focuses on the 3-4 major required tasks and then the associated effort and expertise recommended to accomplish the tasks.

A follow on writing meeting will be held in July that will focus on finalizing the Implementation Plan for each Scientific Objective.

Appendix 1: Agenda

DOE Knowledgebase System Development Workshop

Crystal City, Virginia
Tuesday, June 1, 2010

June 1, 2010

9:00 a.m. – 9:10 a.m.	Welcome, Susan Gregurick
9:10 a.m. – 10:00 a.m.	Workshop Objectives and Expectations, Bob Cottingham
10:00 a.m. – 12:00 p.m.	Divide into Three Breakout Groups
	Microbial Communities — Scientific Objectives Breakout Leaders — Jim Liao and Wim Vermaas
	Plant Communities — Scientific Objectives Breakout Leaders — Maureen McCann and Pam Green
	Meta-Communities — Scientific Objectives Breakout Leaders — Jack Gilbert and Jared Leadbetter
10:30 a.m. – 10:45 a.m.	Break
12:00 p.m. – 12:30 p.m.	Working Lunch
12:30 p.m. – 2:00 p.m.	Breakout Groups report back on Scientific Objectives and Priorities
2:00 p.m. – 4:00 p.m.	Divide into Three Breakout Groups
	Microbial Communities — Requirements Breakout Leaders — Bernhard Palsson and Bob Landick
	Plant Communities — Requirements Breakout Leaders — Robin Buell and Will York
	Meta-Communities — Requirements Breakout Leaders — Steve Slater and Jeff Grethe
3:00 p.m. – 3:15 p.m.	Break
4:00 p.m. – 5:30 p.m.	Breakout Groups report back on Requirements and Priorities
5:30 p.m. – 5:45 p.m.	Conclusions and Adjourn, Bob Cottingham
6:30 p.m.	Working Dinner for Chairs and Breakout Leaders

June 2, 2010

8:00 a.m. – 8:15 a.m.	Recap of June 1, Bob Cottingham
8:15 a.m. – 10:00 a.m.	Divide into Three Breakout Groups

	Microbial Communities — Requirements Breakout Leaders — Bernhard Palsson and Bob Landick
	Plant Communities — Requirements Breakout Leaders — Robin Buell and Will York
	Meta-Communities — Requirements Breakout Leaders — Steve Slater and Jeff Grethe
10:00 a.m. – 10:15 a.m.	Break
10:15 a.m. – 12:00 a.m.	Breakout Groups report back on Final Requirements
12:00 p.m. – 12:30 p.m.	Working Lunch
12:30 a.m. – 4:00 p.m.	Divide into Three Breakout Groups
	System Architecture Breakout Leaders — Ian Gorton and Dan Stanzione
	Implementation Plan Breakout Leaders — Peter Karp and Ed Uberbacher
	Governance Breakout Leaders — Miron Livny and Steve Goff
3:00 p.m. – 3:15 p.m.	Break
4:00 p.m. – 5:30 p.m.	Breakout Groups report back on System Architecture, Implementation Plan, and Governance
5:30 p.m. – 5:45 p.m.	Conclusions and Adjourn, Bob Cottingham

June 3, 2010

9:00 a.m. – 9:40 a.m.	Recap of June 1st and 2nd, Bob Cottingham
9:40 a.m. – 10:30 a.m.	Writing Assignments
10:30 a.m. – 11:00 a.m.	Break
11:00 a.m. – 12:30 p.m.	Group Recap, Bob Cottingham • Where we are? • Missing pieces • Assignments
12:30 p.m. – 1:00 p.m.	Working Lunch
1:00 p.m. – 4:45 p.m.	Continue work
3:00 p.m. – 3:30 p.m.	Break
4:45 p.m. – 5:00 p.m.	Conclusions and Adjourn, Bob Cottingham

Knowledgebase Wiki: sites.google.com/a/systemsbiologyknowledgebase.org/kbase/

Appendix 2: Participants and Observers

Participants

Baliga, Nitin (Institute for Systems Biology)
Beliaev, Alex (PNNL)
Benton, David (GLBRC)
Blum, Paul (University of Nebraska, Lincoln)
Bowen, Ben
Brettin, Tom (ORNL)
Buell, Robin (Michigan State University)
Cannon, Bill (PNNL)
Canon, Shane (LBL)
Chang, Christopher (NREL)
Chivian, Dylan (JBEI/LBL)
Collart, Frank (ANL)
Cottingham, Bob (ORNL)
Desai, Narayan (ANL)
D'haeseleer, Patrik (LLNL)
Gilbert, Jack (Plymouth Marine Laboratory)
Gilna, Paul (BESC/ORNL)
Godzik, Adam (Sanford-Burnham Medical Res. Inst.)
Goff, Steve (iPLANT)
Gorton, Ian (PNNL)
Green, Pam (University of Delaware)
Grethe, Jeff (University of California, San Diego)
Haft, Daniel (J. Craig Venter Institute)
Jackson, Keith (LBNL)
Jenkins, Jerry (Hudson Alpha Inst. for Biotechnology)
Kalluri, Udaya (BESC/ORNL)
Karp, Peter (SRI International)
Kelly, Bob (University of North Carolina)
Kleese van Dam, Kerstin (PNNL)
Landick, Bob (University of Wisconsin)
Lansing, Carina (PNNL)
Leadbetter, Jared (California Inst. of Technology)
Liao, Jim (University of California, Los Angeles)
Liu, Jenny Yan (PNNL)
Livny, Miron (University of Wisconsin)
Mahadevan, Krishna (University of Toronto)
Markowitz, Victor (LBNL/JGI)
Maslov, Sergei (BNL)
McCann, Maureen (Purdue University)
McCue, Lee Ann (PNNL)
Methe, Barbara (J. Craig Venter Institute)
Meyer, Folker (ANL)
Mockler, Todd (Oregon State University)
Osterman, Andrei (Burnham)
Palsson, Bernhard (University of California, San Diego)
Pop, Mihai (University of Maryland)
Reed, Jenny (University of Wisconsin)
Romine, Margie (PNNL)
Samatove, Nagiza (North Carolina State University)
Sayler, Gary (University of Tennessee, Knoxville)
Setubal, Joao (Virginia Bioinformatics Institute)
Slater, Steve (GLBRC)
Stanzione, Dan (University of Texas)
Stevens, Rick (ANL)
Tatusova, Tatiana (NIH)
Tobias, Chris (USDA)
Uberbacher, Edward (ORNL)
Vermaas, Wim (Arizona State University)
White, Owen (University of Maryland)
Wu, Cathy (University of Delaware)
Yan, Koon-Kiu (Yale University)
York, Will (University of Georgia)
Zengler, Karsten (University of California, San Diego)

Observers

Bayer, Paul (DOE BER)

Bownas, Jennifer (ORNL)

Christen, Kris (University of Tennessee)

Foust, Cheri (ORNL)

Graber, Joe (DOE BER)

Gregurick, Susan (DOE BER)

Haun, Holly (University of Tennessee)

Houghton, John (DOE BER)

Katz, Arthur (DOE BER)

Mansfield, Betty (ORNL)

Nagahara, Larry (National Cancer Institute)

Ronning, Cathy (DOE BER)

Schexnayder, Susan (University of Tenn., Knoxville)

Weatherwax, Sharlene (DOE BER)

Yousef, Shireen (DOE BER)

Acronyms

ANL	Argonne National Laboratory
BESC	BioEnergy Science Center
BNL	Brookhaven National Laboratory
DOE	U.S. Department of Energy
BER	Biological and Environmental Research
GLBRC	Great Lakes Bioenergy Research Center
JBEI	Joint BioEnergy Institute
JGI	Joint Genome Institute
LBNL	Lawrence Berkeley National Laboratory
LLNL	Lawrence Livermore National Laboratory
NIH	National Institutes of Health
NREL	National Renewable Energy Laboratory
ORNL	Oak Ridge National Laboratory
PNNL	Pacific Northwest National Laboratory
USDA	U.S. Department of Agriculture

Appendix 3: Scientific Objectives Template

Scientific Objective: <title – Note: 1 Objective per each filled in template>

Breakout Group: <group>

Contributing Authors: <authors>

Date: <date>

1. Scientific Objective

Brief statement of Scientific objective

[What is the scientific or research goal? What is written here will usually be derived after filling in the remainder of the template. Responses to sections below will help to refine this statement. Sometimes it is easier to think of an objective in terms of a problem that exists that needs to be solved.]

Background information

[Include ongoing experiments, future planned experiments, historical results, literature references, relevant past impediments to research progress, etc.]

2. Prioritization

[This is meant to help prioritize this scientific objective in the context of other scientific objectives. There are several axes of consideration. One is the need or benefit to the research community. Another is the level of difficulty or feasibility.]

PRIORITY (check one): ___ HIGH ___ MEDIUM ___ LOW

Potential Benefits

[Why is this objective important? What is its level of impact? What would the benefit be? Who would benefit?]

Feasibility of success Near, Mid and Long term

[What is the level of difficulty? How likely is it that this objective can be achieved in a 1-3 year time frame? What would be the measure of success? Consider and rate feasibility in the near term (1-3 years, midterm (3-5 years) and long term (5-10 years). The most important objectives, those that are high priority and most feasible in the near term must have the most detail. Mid and long term can be provided in decreasing levels of detail.]

TERM (check one): _ NEAR (1-3 years) ___ MID (3-5 years) ___ LONG (5-10 years)

Relevance to DOE systems biology knowledgebase project

[The DOE Genomic Science program's ultimate goal of achieving a predictive understanding of biological systems is a daunting challenge and will require the integration of immense amounts of diverse information. The DOE Systems Biology Knowledgebase is envisioned as an open cyber-infrastructure to integrate systems biology data, analytical software, and computational modeling tools that will be freely available to the scientific community. Briefly explain how the proposed objective is relevant to what is envisioned for Kbase.]

Synergies/Leverage: Potential overlap with other projects or funding agencies

[Are there existing systems that relate to this objective such as NCBI, GenBank, BioCyc, iPlant, etc. Is there a potential for synergy that would benefit both efforts? Is there a potential overlatp that needs to be resolved?]

3. Specificity

[This section pertains to finding the right level of objective, especially avoiding objectives that are specified at too high a level. Start with a high level objective and refine. What is the specific science question to be answered?]

4. Details

[The intent here is to begin to capture elements that form the basis for continuity between the science objective and the software requirements that are derived from this objective. We start to articulate high level requirements here that are further refined in the requirements document.]

Scientific discovery process (workflows)

[Have workflows already been developed or can they be derived from existing work?]

Inputs

[What datasets would be required? Are there data standards? Are there available data sources or examples? Are there publications that use or describe an associated analysis process?]

Outputs

[What would the results be?]

Tools

[Existing or new analysis software]

REFERENCES

[Use as needed]

APPENDICES

[Use as needed]

- **Figures**
- **Tables**

EXAMPLE Scientific Objective: Improve Prediction of Microbial Gene Regulatory Networks

Breakout Group: Microbial

Date: May 12, 2010

5. Scientific Objective

Brief statement of Scientific objective

[What is the scientific or research goal? What is written here will usually be derived after filling in the remainder of the template. Responses to sections below will help to refine this statement. Sometimes it is easier to think of an objective in terms of a problem that exists that needs to be solved.]

Informative Example: Next generation sequencing technology will provide high quality RNA-Seq data at low cost. This presents an opportunity to substantially improve the quality of predicted gene regulatory networks compared with what has been possible with expression microarrays. This data together with transcription factor binding site predictions or determinations will provide the necessary data to built genetic regulatory networks for microbial genomes. High quality genetic regulatory networks of experimentally tractable organisms would increase the efficiency of experimental designs and genetic engineering. In the long term, having a collection of transcript profiles collected in a high quality, standardized manner across DOE relevant organisms such that genetic regulatory networks could be automatically determined in the context of the Kbase would provide an extremely valuable resource to advance microbial research.

Background information

[Include ongoing experiments, future planned experiments, historical results, literature references, relevant past impediments to research progress, etc.]

Informative Example: Next generation sequencing technology provides high quality RNA-Seq data at low cost. When acquired in sufficient quantity RNA-Seq data has dramatically better dynamic range and sensitivity than gene expression arrays and will probably replace them in 3-5 years. Transcriptome data can be used to define operons including transcription initiation and termination sites. Cluster analysis over multiple conditions will identify co-regulated operons and therefore defines co-regulated promoters. This data together with transcription factor binding site predictions or determinations will provide the necessary data to built genetic regulatory networks for microbial genomes.

6. Prioritization

[This is meant to help prioritize this scientific objective in the context of other scientific objectives. There are several axes of consideration. One is the need or benefit to the research community. Another is the level of difficulty or feasibility.]

Informative Example: Since genetic regulatory networks will facilitate efficient genetic engineering and other experimental designs (Cho et al., 2009), the priority of this objective should be high. (This statement of priority can be written at the workshop based on discussion.)

PRIORITY (check one): ___ HIGH ___ MEDIUM ___ LOW

Potential Benefits

[Why is this objective important? What is its level of impact? What would the benefit be? Who would benefit?]

Informative Example: Genetic regulatory networks of experimentally tractable organisms would increase the efficiency of experimental designs and genetic engineering. Microbes will be increasingly more important in manipulating a variety of organic molecules for biofuels, alternative plastics, other biochemical feedstocks, carbon sequestration and environmental remediation. Having the ability to efficiently manipulate and engineer these organisms will be absolutely crucial for cost effective design and large scale production of useful biochemicals.

Feasibility of success Near, Mid and Long term

[What is the level of difficulty? How likely is it that this objective can be achieved in a 1-3 year time frame? What would be the measure of success? Consider and rate feasibility in the near term (1-3 years, midterm (3-5 years) and long term (5-10 years). The most important objectives, those that are high priority and most feasible in the near term must have the most detail. Mid and long term can be provided in decreasing levels of detail.]

TERM (check one): _X_ NEAR (1-3 years) ___ MID (3-5 years) ___ LONG (5-10 years)

Informative Example: Collecting RNA-Seq data is already feasible and will only become more cost effective as third generation sequencing technologies are available in the next year. The community is already engaged in the development of analytical tools capable of integrating genomic DNA sequence and RNA-Seq data. The methods for predicting operons and their structure, cluster analysis of transcriptomic data to predict co-regulation of operons, predicting transcription factor binding sites and regulatory elements are already available but need to be streamlined and integrated. Implementing these kinds of analytical capabilities within the Kbase would be feasible in the first 1-2 years. RNA-Seq data is expected

to be widely available in the midterm 3-5 years and will need to be standardized to avoid some of the problems seen with GEO. Producing a functional genetic regulatory network for one or more bacterial organisms important to the Bioenergy centers appears to be achievable in 2-3 years if sufficient resources are applied.

Relevance to DOE systems biology knowledgebase project

[The DOE Genomic Science program's ultimate goal of achieving a predictive understanding of biological systems is a daunting challenge and will require the integration of immense amounts of diverse information. The DOE Systems Biology Knowledgebase is envisioned as an open cyber-infrastructure to integrate systems biology data, analytical software, and computational modeling tools that will be freely available to the scientific community. Briefly explain how the proposed objective is relevant to what is envisioned for Kbase.]

Informative Example: Predicting genetic regulatory networks requires integration of standardized sets of data and associated analysis methods along with the ability to test and improve the methods as envisioned for the Kbase. In the long term, having a collection of transcript profiles collected in a high quality, standardized manner across DOE relevant organisms such that genetic regulatory networks could be automatically determined in the context of the Kbase would provide an extremely valuable resource to advance microbial research.

Synergies/Leverage: Potential overlap with other projects or funding agencies

[Are there existing systems that relate to this objective such as NCBI, GenBank, BioCyc, iPlant, etc. Is there a potential for synergy that would benefit both efforts? Is there a potential overlatp that needs to be resolved?]

Informative Example: Generating the necessary RNA-Seq data would leverage JGI's production sequencing capabilities and could be synchronized with the genomic sequencing, while developing the analysis pipeline could be accomplished by ORNL's annotation group and incorporated into its' annotation pipeline. Individual PIs and smaller projects already pursue such analysis based on microarray data and the decreasing cost of RNA-Seq will eventually make RNA-Seq transcriptomics routine. Having a community of data integrated based on standards will provide a powerful resource. A natural byproduct will be better gene models and operon structures. This information will augment what is available in GenBank. An ancillary objective would be to update the associated annotation in GenBank.

7. Specificity

[This section pertains to finding the right level of objective, especially avoiding objectives that are specified at too high a level. Start with a high level objective and refine. What is the specific science question to be answered?]

Informative Example: Integration of 'omics data especially in complex systems such as plant microbe interfaces is an ambitious challenge that is too high level for the purposes of establishing version 1 of the Kbase, and not feasible in the near term (1-3 years) although it would be appropriate for the long term with a suitable scientific focus. However this high level aim could be made more tractable by simplifying in several ways. First, focus in on a simpler system such as a specific microbe. Second, instead of integrating all 'omics, take just two types of 'omics data, say genomic and transcriptomic as in this example.

In this example we started by considering 'omics integration and the large number of possible scientific objectives might derive from that such as a substantial model of major subsystems of a cell which would clearly be overly ambitious. By considering various combinations of 'omics data the level can be refined. In this example we recognized that by integrating just two kinds of 'omics data, genomics and transcriptomic using RNA-Seq, we would be able to have a science objective of improved prediction of gene regulatory networks that would be something tractable to accomplish in the relative near term within the Kbase and something useful to the microbial research community.

8. Details

[The intent here is to begin to capture elements that form the basis for continuity between the science objective and the software requirements that are derived from this objective. We start to articulate high level requirements here that are further refined in the requirements document.]

Scientific discovery process (workflows)

[Have workflows already been developed or can they be derived from existing work?]

Informative Example: Genetic regulatory networks have been created for *E. coli* (Cho et al., 2009) and *Halobacteria salinarum* NRC-1 (Bonneau et al., 2007). These papers describe workflows.

Inputs

[What datasets would be required? Are there data standards? Are there available data sources or examples? Are there publications that use or describe an associated analysis process?]

Informative Example: For a particular microbe of interest it would be expected that a finished genome sequence is available and for a few phylogenetically related organisms. In addition it would be expected that RNA-Seq of multiple growth states would have been obtained to a high level of coverage.

Outputs

[What would the results be?]

Informative Example: The results would be genetic regulatory network predictions for all microbes studied.

Tools

[Existing or new analysis software]

Informative Example: Numerous independent tools that have been developed. It will be necessary to develop analytical pipelines based on agreed workflows that integrate the RNA-Seq data, genomic sequence data, gene expression array data (if available), transcription factor binding site predictions and experimental verification (if available) in order to generate the genetic regulatory network predictions for a particular microbe.

REFERENCES

[Use as needed]

Bonneau, R., Facciotti, M.T., Reiss, D.J., Schmid, A.K., Pan, M., Kaur, A., Thorsson, V., Shannon, P., Johnson, M.H., Bare, J.C.*, et al.* (2007). A predictive model for transcriptional control of physiology in a free living cell. Cell *131*, 1354-1365.

Cho, B.K., Zengler, K., Qiu, Y., Park, Y.S., Knight, E.M., Barrett, C.L., Gao, Y., and Palsson, B.O. (2009). The transcription unit architecture of the Escherichia coli genome. Nat Biotechnol *27*, 1043-1049.

APPENDICES

[Use as needed]

- **Figures**
- **Tables**

Appendix 4: Requirements Template

Software Requirements for Scientific Objective: *Improve Prediction of Microbial Gene Regulatory Networks*

Breakout Group: Microbial

Reference Scientific Objective Number in Group: ____

Date: May 18, 2010

1 Scientific objective

[Describe the scientific objective that this software system requirements document supports. Description can be derived from the Scientific Objective template]

2 Resulting Requirements

[In the following sections list the requirements resulting directly from the identified scientific objective. Provide information for each requirement stating whether there are technologies available today to fulfill all or part of it that you are aware of, or if you expect that new development would be required. All requirements should indicate whether they are near, medium or long term requirements. The following Impact Factor is your group's assessment of the impact that addressing these requirements would have toward improving research productivity.]

IMPACT FACTOR (check one): ___ HIGH ___ MEDIUM ___ LOW

2.1 Process of the science (incl. workflow)

[Describe the process by which scientists use or want to use the data, software, and instruments for knowledge discover such as a scientific workflow. Identify both required and optional components. Indicate the state of the art of the different parts of the workflow.]

2.2 Instruments to support the achievement of the science objectives.

[List or describe instruments that generate relevant data connected to the scientific workflow above.]

2.3 User interfaces

[Describe generally who the users will be, and the user interfaces that play a role in achieving the scientific objective in the context of the workflows outlined above. Not all user interfaces will be directly involved in a workflow, and these if they exist should be captured here as well.]

2.4 Programmatic interfaces

[Describe the interfaces that will provide programmatic access to data or functionality that allow for automated data access, analyses and workflows in the context of the workflows outlined above. Not all programmatic interfaces will be directly involved in a workflow, and if these exist, capture them here as well.]

2.5 Data

[Describe the data and data types required to meet the scientific objectives. Include publicly available data, reference data, and new experimentally derived data. Discuss how the data is obtained such as is the data to reside locally on Kbase or would it exist remotely, outside of Kbase. Data representations including semantic web technologies or references can be included here, as well as references to existing data standards or relational tables. If known, include computer hardware resource requirements – such as the size of the data collection, and type and size of compute resources (processors, memory, temporary storage) required to manage and process the data.]

2.6 Software

[Describe which software algorithms, services and packages will be needed, if they exist or not, to achieve the scientific objective, and what computer hardware or other resources and data these would utilize.]

Software purpose	Availability	Does it need improvement	Resource impact

2.7 Standards

[Specify requirements that are derived from existing standards and/or regulations. While we don't expect much in the form of regulation, we should list those existing standards that we will use and areas where new standards need to be developed.]

2.8 Governance

[Related governance issues (usage policy, data policy, overall governance structure, community engagement for usage and development) should be described here. Some governance issues map to components of the system and these mappings should be called out in the System Architecture. How can governance help the implementation of standards?]

2.9 Summary and prioritization of requirements

[Summarize and prioritize your requirements, which ones are essential and which one are nice to have or could wait. Which requirements are near term, midterm and long term?]

3 System Architecture Attributes

[The common attributes are performance, reliability, availability, security, portability, interoperability, and usability (usually speaks to the importance of user interfaces with which humans interact as compared to a fully automated system that users just depend on). Important attributes from the list above should be discussed in the context of the scientific objective. For example, does achieving the science objective require a system that runs 24/7 with a yearly downtime of less than 8 minutes (this reflects the system's availability attribute). Will it perform calculations that require thousands of cores in order to complete in a reasonable time. Prioritize the relative importance of each architecture attribute and provide explanations of why, for example, why would security be more or less or equal in importance to performance.]

4 Kbase Key Services

[Optional – do this if able: Provide a list and description of the major functions/services that the Kbase system will need to provide to meet the scientific objective(s). This could include a mapping of existing functions onto existing systems such as MicrobesOnline, IMG, etc., or new services such as a central resource for temporary storage of data from different sources to be jointly analyzed. Here we can get into the finer details of what the system will do in order to meet the scientific objectives. Each function should be called out as a sub heading in this section]

4.1.1
4.1.2
4.1.3

5 Risk Analysis and Mitigation strategies

[Compile the list of potential risks in meeting the requirements of the scientific objective.]

-
-
-

6 Acryonyms, definitions and abbreviations

7 References

EXAMPLE Software Requirements for Scientific Objective: *Improve Prediction of Microbial Gene Regulatory Networks*

Breakout Group: Microbial

Reference Scientific Objective Number in Group: _____

Date: May 18, 2010

8 Scientific objective

[Describe the scientific objective that this software system requirements document supports. Description can be derived from the Scientific Objective template]

Improve Prediction of Microbial Gene Regulatory Networks

Next generation sequencing technology will provide high quality RNA-Seq data at low cost. This presents an opportunity to substantially improve the quality of predicted gene regulatory networks compared with what has been possible with expression microarrays. This data together with transcription factor binding site predictions or determinations will provide the necessary data to built genetic regulatory networks for microbial genomes. High quality genetic regulatory networks of experimentally tractable organisms would increase the efficiency of experimental designs and genetic engineering. In the long term, having a collection of transcript profiles collected in a high quality, standardized manner across DOE relevant organisms such that genetic regulatory networks could be automatically determined in the context of the Kbase would provide an extremely valuable resource to advance microbial research.

When acquired in sufficient quantity RNA-Seq data has dramatically better dynamic range and sensitivity than gene expression arrays and will probably replace them in 3-5 years. Transcriptome data can be used to define operons including transcription initiation and termination sites. Cluster analysis over multiple conditions will identify co-regulated operons and therefore defines co-regulated promoters. This data together with transcription factor binding site predictions or determinations will provide the necessary data to built genetic regulatory networks for microbial genomes.

9 Resulting Requirements

[In the following sections list the requirements resulting directly from the identified scientific objective. Provide information for each requirement stating whether there are technologies available today to fulfill all or part of it that you are aware of, or if you expect that new development would be required. All requirements should indicate whether they are near, medium or long term requirements. The following Impact Factor is your group's assessment of the impact that addressing these requirements would have toward improving research productivity.]

IMPACT FACTOR (check one): ___ HIGH ___ MEDIUM ___ LOW

9.1 Process of the science (incl. workflow)

[Describe the process by which scientists use or want to use the data, software, and instruments for knowledge discover such as a scientific workflow. Identify both required and optional components. Indicate the state of the art of the different parts of the workflow.]

(NOTE: This is an example that has been intentionally simplified and therefore extensions such as validation with computational or experimental methods such as 5' RACE to identify additional transcription initiation sites, or transcription factor regulatory ligand determinations have been removed. Others are welcome to take this as a starting point and expand for a specific "real" Scientific Objective.)

Taken from Scientific Objective section 4.2 Inputs: For a particular microbe of interest it would be expected that a finished genome sequence is available and for a few phylogenetically related organisms. In addition it would be expected that RNA-Seq of multiple growth states would have been obtained to a high level of coverage.

For the organism of interest it is assumed that the genome has been completely sequenced, fully annotated, and that RNA-Seq data is available for a minimum of 10 growth curves with 6 time points and 3 biological replicates on biological conditions relevant to the functional network(s) of interest.

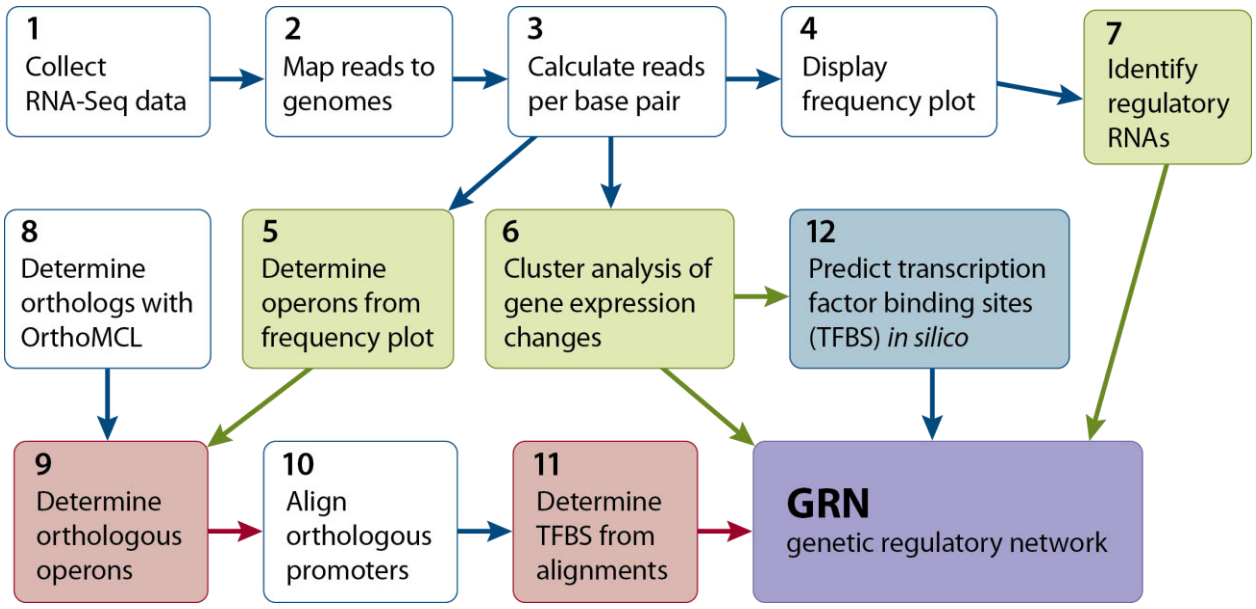


Figure 1. Transcriptome Analysis Pipeline for Gene Regulatory Network Prediction. White boxes are procedures we already know how to do. Green boxes are procedures that have not been determined but expected to be fairly easy to construct (year 1). Red boxes are procedures that will be more difficult to construct (year 2). Blue boxes are techniques that are optional, but would increase the accuracy of the analysis. The purple box is the final product (year 2).

1. Collect RNA-Seq data and the accompanying metadata for each growth curve. The metadata could include optical density, substrate consumption, metabolites, temperature, and stirring condition. Although some of this data could be manually collected the Kbase would need to have the ability to store it in conjunction with the RNA-Seq as an experimental project.
2. Map RNA-Seq data to genome.
3. Calculate reads/bp (normalize and calculate expression levels of each gene and/or operon).
4. Display frequency plot for visual inspection and rule development for algorithms to identify the operons and regulatory RNAs in steps 5 and 7.
5. Determine operons from mapped reads (generate a separate list for each growth curve). These should include all the genes, the transcription initiation sites (TIS) and terminations sites with accuracy of a few bp for each operon.
6. Perform cluster analysis on the calculated gene expression levels to determine co-regulated operons.
7. Identify regulatory RNAs (unknown riboswitches and small regulatory RNAs) based on analysis derived rules identified in step 4 with expert guidance.
8. Determine orthologs from multiple related genomes using OrthoMCL or some other software tool.

9. Determine orthologous promoters from multiple related genomes.
10. Align orthologous promoters using Muscle or ClustalW.
11. Determine Sigma Factor and other Transcription Factor Binding Sites (TFBS) from alignments.
12. Use in silico TFBS prediction tools together with co-regulated operons to predict additional TFBS (import known TFBS from a database such as RegTransBase).
13. Predict Genetic Regulatory Network.

Further work after the initial implementation (years 1-2) would include evaluation of additional technologies and experimental verification to improve the process (5' RACE to identify additional TISs, microfluidic TFBS determinations and transcription factor regulatory ligand determinations). As the quality of the gene regulatory network predictions improves and the models are validated the workflow will be increasingly automated (years 3-5).

9.2 Instruments to support the achievement of the science objectives.

[List or describe instruments that generate relevant data connected to the scientific workflow above.]

Kbase should support RNA-Seq data from Solexa, ABI Solid, and 454. For the future there may be additional machines that will need to be supported such as PacBio. These instruments produce data of particular types and sizes that will need to be stored and/or managed within the context of the Kbase and are further described in the Data section below.

There will be potential for use of automated or multi-well instruments for generating growth curve data. Metadata such as optical density may be recorded manually or in spreadsheets output from instruments and Kbase will need to have capabilities for manual input or upload of such electronic data that would then be integrated within the experimental project.

9.3 User interfaces

[Describe generally who the users will be, and the user interfaces that play a role in achieving the scientific objective in the context of the workflows outlined above. Not all user interfaces will be directly involved in a workflow, and these if they exist should be captured here as well.]

The anticipated users will include biologists who wish to analyze their data, bioinformaticists who want to analyze data, contribute or improve methods and use existing methods, and scientists requiring information and visual representations for scientific publications. It is

anticipated that users will come from the academic, government and industry communities. It is not anticipated that there will be users at a level below the university level.

Interfaces will be needed for specifying an experimental project and locating the relevant RNA-Seq and associated experimental metadata. Users will expect to have a login space where they describe their experiment that they will be able to save and return to at a later time.

Scientific data visualization is needed that renders genome annotation, gene expression information, operons, alternative transcriptional starts, and multiple sequence alignments. User interfaces for visualizing frequency plots that show depth of coverage (relative expression levels) for genes and operons will be needed. Additionally, an interface that allows users to visualize the resulting gene regulatory network model will be needed.

9.4 Programmatic interfaces

[Describe the interfaces that will provide programmatic access to data or functionality that allow for automated data access, analyses and workflows in the context of the workflows outlined above. Not all programmatic interfaces will be directly involved in a workflow, and if these exist, capture them here as well.]

Kbase will need to have programmatic interfaces to support specific queries such as to return a list of all experimental conditions that an organism has been exposed to for which there is gene expression data.

Also, software that determines expression levels, or predicts or refines operon predictions will need access to genome annotation. Therefore data interfaces to NCBI SRA (Sequence Read Archive), GEO (Gene Expression Onmibus) and GenBank (bacterial genomes) will be needed or perhaps application interfaces to IMG, RAST or JGI-ORNL annotation systems. Will also need to import known TFBS from a database such as RegTransBase

The results of the workflow to predict gene regulatory networks will also produce data that would be output to data interfaces such as to all of the systems mentioned above in order to supply new data, support publication submission or update annotation.

9.5 Data

[Describe the data and data types required to meet the scientific objectives. Include publicly available data, reference data, and new experimentally derived data. Discuss how the data is obtained such as is the data to reside locally on Kbase or would it exist remotely, outside of Kbase. Data representations including semantic web technologies or references can be included here, as well as references to existing data standards or relational tables. If known, include computer hardware resource requirements – such as the size of the data collection, and type and size of compute resources (processors, memory, temporary storage) required to manage and process the data.]

In the near term, we expect to see for a given experiment several hundred files from short read sequencing technology. These files, if based on Solexa technology, will range in size from 100Mbytes to 100GBytes for the next couple of years. Current size ceiling is at about 4GBytes compressed for one run. Total data storage required is based on coverage and number of replicates, conditions and time steps, and therefore would be a multiplicative factor of 4GB (180X minimum as proposed). For the first 1-3 years it is expected that there would be 30-100 datasets per year (each dataset corresponding to studies on one microbe), and then grow to 100-300 per year in the 3-5year time frame when this data will be coming from many laboratories.

Database and storage resources – terabyte to petabytes range storage are needed. Data reduction will play a role in keeping storage resources manageable. Online backup capabilities needed for disaster recovery and long term archival.

Data types that cover high-throughput technologies to interrogate the transcriptome, are required for this scientific objective.

Genome sequences and a full complement of annotation features are also required. The data representation model as characterized by a GenBank record is probably not sufficient. New data models that capture gene annotations and their relationships to other annotations will be required. Annotation can exist remotely as in the case of taxonomy information housed in the NCBI taxonomy database and other NCBI Entrez data for which stable access exists through NCBI web services.

The gene regulatory network from a data structure perspective is the collection of operons, transcription factor binding sites, sigma factor binding sites; and those parameters that affect kenetics. These would benefit from representation that is based in semantic web technology.

It is expected that relational database technology will play a limited role in so far as perhaps providing structured storage of ontologies and RDF tuples.

9.6 Software

[Describe which software algorithms, services and packages will be needed, if they exist or not, to achieve the scientific objective, and what computer hardware or other resources and data these would utilize.]

Software for performing transcriptome analysis will be needed as part of the workflow and for visualization. It will integrate existing available genome annotation and provide measures of confidence. Annotation quality will be accessed based on confidence. A specific module will focus directly on improved identification of transcription factors.

Improved annotation with confidence and evidence codes will be sent back to repositories if possible.

Clustering software will be needed to group genes and operons into clusters based on patterns of regulation. Whether a part of the clustering software or part of a different package, it is anticipated that software which focuses on the fine details of the operon such as alternative transcriptional starts and stops will be needed.

Clustering algorithms will be compute intensive. Other methods are manageable with mid-range servers.

Data visualization software that spans genome annotation, transcriptome analysis and clustering will also be needed.

Software purpose	Availability	Does it need improvement	Resource impact
Maps rna-Seq data to genome	Few	Probably not	Storage
Cluster analysis of gene expression changes	Many	Probably	Compute, Storage
Operon determination	Few	Yes	
In silico TFBS prediction	Many	Yes	Compute
Ortholog determination	Few	Probably not	
Orthologous operon determination	None		
Promoter alignment	Few	Yes	
Promoter prediction	Few	Yes	
Gene regulatory network prediction	Few	Yes	Compute, Storage

Table 1: Types of software required for this scientific objective. Column-Resource impact: Compute means requires significant processor resource (>100 cores), and Storage means requires significant storage resource (>1 TB).

9.7 Standards

[Specify requirements that are derived from existing standards and/or regulations. While we don't expect much in the form of regulation, we should list those existing standards that we will use and areas where new standards need to be developed.]

-Gene regulation ontology (GRO) for terms related to gene expression

-Gene ontology (GO) for terms related to biological processes, cellular location and gene function

-NCBI sequence read archive xml schemas for sequence read metadata

-GCDML xml schema for genome metadata

-MIAME regards gene expression arrays but may be relevant to RNA-Seq.

9.8 Governance

[Related governance issues (usage policy, data policy, overall governance structure, community engagement for usage and development) should be described here. Some governance issues map to components of the system and these mappings should be called out in the System Architecture. How can governance help the implementation of standards?]

A data release policy will need to be in place. This would most likely be the current DOE policy and it is assumed that the Kbase will enforce this. This implies a private login which maps to System Architecture.

9.9 Summary and prioritization of requirements

[Summarize and prioritize your requirements, which ones are essential and which one are nice to have or could wait. Which requirements are near term, midterm and long term?]

In silico prediction of TBFS can be postponed until other elements of the workflow are complete (midterm). Support for microarray data was considered but has not been included for the sake of simplicity. If it would part of the requirements it might be lower priority because we believe it is phasing out. Other requirements for possible inclusion would be various kinds of validation such as 5' RACE and TFBS verification (midterm).

10 System Architecture Attributes

[The common attributes are performance, reliability, availability, security, portability, interoperability, and usability (usually speaks to the importance of user interfaces with which humans interact as compared to a fully automated system that users just depend on). Important attributes from the list above should be discussed in the context of the scientific objective. For example, does achieving the science objective require a system that runs 24/7 with a yearly downtime of less than 8 minutes (this reflects the system's availability attribute). Will it perform calculations that require thousands of cores in order to complete in a

reasonable time. Prioritize the relative importance of each architecture attribute and provide explanations of why, for example, why would security be more or less or equal in importance to performance.]

Users will be expecting that the data they submit will be secure in accordance with the governance model. This would be the highest priority.

It is anticipated that there could be some performance issues resulting from the choice of clustering algorithms and the amount of input data. Performance and security are architecture issues considered of highest importance for this objective.

11 Kbase Key Services

[Optional – do this if able: Provide a list and description of the major functions/services that the Kbase system will need to provide to meet the scientific objective(s). This could include a mapping of existing functions onto existing systems such as MicrobesOnline, IMG, etc., or new services such as a central resource for temporary storage of data from different sources to be jointly analyzed. Here we can get into the finer details of what the system will do in order to meet the scientific objectives. Each function should be called out as a sub heading in this section]

11.1.1 Mapping RNA sequence reads to a genome
11.1.2 Identifying operons
11.1.3 Identifying alternative transcription starts and stops
11.1.4 Identifying transcription factor binding sites
11.1.5 Improvements to genome annotation based on services 4.1.1 – 4.1.4
11.1.6 Data structures for representing gene regulatory networks
11.1.7 Query services for retrieving gene regulatory network models
11.1.8 Query services for retrieving all experimental conditions that an organism has been exposed to for which there is gene expression data

12 Risk Analysis and Mitigation strategies

[Compile the list of potential risks in meeting the requirements of the scientific objective.]

- Unanticipated changes in technology (sequencing, microarray) that would significantly change the requirements or implementation plan. Mitigated by anticipating changes and adjusting requirements and implementation plan as soon as possible.
- Inadequate data or poor data quality that precludes a productive workflow as currently designed. Mitigate by testing typical datasets for adequacy and quality. Modify experimental protocol to correct and change minimum standards.
- Cluster analysis on these datasets requires more resources than currently anticipated. Mitigate by modifying algorithm accept some additional error in return for performance speed. Allow clustering on subsets to manually find the optimum with reduced error.

13 Acryonyms, definitions and abbreviations

14 References

APPENDIX E

References

Addo-Quaye, C., et al. 2008. "Endogenous siRNA and miRNA Targets Identified by Sequencing of the *Arabidopsis* Degradome," *Current Biology* **18**, 758–762.

Bell, G., et al. 2009. "Beyond the Data Deluge," *Science* **323** (5919), 1297–1298.

Bergmann, F. T., and B. G. Olivier. 2010. "SBML Level 3 Package Proposal: Flux Balance," *Nature Precedings*, doi:10.1038/npre.2010.4236.1

Chen, X. 2010. "Small RNAs: Secrets and Surprises of the Genome," *The Plant Journal* **61**, 941–958.

Coruzzi, G. M., and R. A. Gutiérrez (Eds.) 2009. *Plant Systems Biology*. Blackwell Publishing Ltd., Oxford, UK.

Dale, J. M., L. Popescu, and P. D. Karp. 2010. "Machine Learning Methods for Metabolic Pathway Prediction," *BMC Bioinformatics* **11**(15), doi:10.1186/1471-2105-11-15.

German, M. A., et al. 2008. "Global Identification of MicroRNA-Target RNA Pairs by Parallel Analysis of RNA Ends," *Nature Biotechnology* **26**(8), 941–946.

Green, M., and P.D. Karp. 2004. "A Bayesian Method for Identifying Missing Enzymes in Predicted Metabolic Pathway Databases," *BMC Bioinformatics* **5**(76).

Hu, P., et al. 2009. "Global Functional Atlas of *Escherichia coli* Encompassing Previously Uncharacterized Proteins," *PLoS Biology* **7**(4), 929–947.

Koide, T., W. L. Pang, and N. S. Baliga. 2009. "The Role of Predictive Modeling in Rationally Reengineering Biological Systems," *Nature Reviews Microbiology* **7**(4), 297–305.

Licatalosi, D. D., and R. B. Darnell. 2010. "RNA Processing and Its Regulation: Global Insights into Biological Networks," *Nature Reviews Genetics* **11**, 75–87.

Thiele, I., and B. O. Palsson. 2010. "A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction," *Nature Protocols* **5**(1), 93–121.

Wang, Z., M. Gerstein, and M. Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics," *Nature Reviews Genetics* **10**, 57–63.

Weber, M. M., et al. 2010. "A Previously Uncharacterized Gene, *yjfO* (*bsmA*), Influences *Escherichia coli* Biofilm Formation and Stress Response," *Microbiology* **156**, 139–147.

Yooseph, S., et al. 2007. "The *Sorcerer II* Global Ocean Sampling Expedition: Expanding the Universe of Protein Families," *PLoS Biology* **5**(3), 432–466.

Zhuang, K., et al. 2010. "Genome-Scale Dynamic Modeling of the Competition Between *Rhodoferax* and *Geobacter* in Anoxic Subsurface Environments," *The International Society for Microbial Ecology Journal*, advance online publication July 29, 2010; doi: 10.1038/ismej.2010.117

APPENDIX F

Acronyms

AFRI Agriculture and Food Initiative

AJAX Asynchronous Javascript and XML

ANI average nucleotide identity

ANOVA analysis of variance

API application programming interface

ARB/Silva database of aligned small and large RNA sequences

ASCR DOE Office of Advanced Scientific Computing Research

B biology (used in Staffing Resources tables)

BER DOE Office of Biological and Environmental Research

Bfx, **BFX** software or file extension regarding bioinformatics (used in Staffing Resources tables)

BioCyc collection of 673 pathway and genome databases

BioPAX Biological Pathway Exchange (a language for biological pathway data)

Bio2RDF atlas of post-genomic knowledge

BIRN Biomedical Informatics Research Network

BRC DOE Bioenergy Research Centers (includes the Joint BioEnergy Institute, Great Lakes Bioenergy Research Center, and BioEnergy Science Center)

BRENDA Braunschweig Enzyme Database

caBIG cancer Biomedical Informatics Grid

CAMERA Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis

cDNA complementary DNA

CellML XML-based open standard language for describing mathematical models

CGIAR Consultative Group on International Agricultural Research

ChEBI Chemical Entities of Biological Interest

CHIP-Seq used to analyze protein interaction with DNA

CLR Context Likelihood of Relatedness algorithm

COBRA constraint-based reconstruction and analysis toolbox

COG clusters of orthologous groups

COPASI Complex Pathway Simulator

CPU central processing unit

CS Computer science, computer scientist (used in Staffing Resources tables)

CSREES Cooperative State Research, Education, and Extension Service

DOE U.S. Department of Energy

DM data management

dhIP-Seq chip sequencing used to analyze protein interactions with DNA

dbEST Expressed Sequence Tag Database

dbGaP Database of Genotypes and Phenotypes

DOTUR software program for community diversity analysis

EBI European Bioinformatics Institute

EcoCyc bioinformatics database that describes the genome and biochemical machinery of *Escherichia coli*

EPA Environmental Protection Agency

ESnet Energy Sciences Network

EST expressed sequence tag

FBA flux balance analysis

FTP file transfer protocol

FIGfams protein families generated by the Fellowship for Interpretation of Genomes

FISH fluorescence *in situ* hybridization

FTE full-time equivalent

GB gigabyte

GC gas chromatography

GCDML Genomic Contextual Data Markup Language, xml schema for genome metadata

GDPC Genomic Diversity and Phenotype Connection

GDPDM Genomic Diversity and Phenotype Data Model

GenBank sequence database

GEBA Genomic Encyclopedia of Bacteria and Archaea

GEO Gene Expression Omnibus, database repository of high-throughput gene expression data and hybridization arrays, chips, and microarrays

GForge open-source system for software project management and collaboration

GMOD Generic Model Organisms Database

GO gene ontology

GOLD Genomes OnLine Database

GPS global positioning system

GPU graphics processing unit

GRN genetic regulatory network

GRO gene regulation ontology

GSC Genomic Standards Consortium

GWAS genome-wide association studies

HMP Human Microbiome Project

HPC high-performance computing

hPDL Hadoop Process Definition Language (an XML-based workflow definition language)

HTS high-throughput sequencing or high-throughput screening

IBP Integrated Breeding Platform

ICIS International Crop Information System

ICRISAT International Crop Research Institute for the Semi-Arrid Tropics

IMG Integrated Microbial Genomes

IMG/M Integrated Microbial Genomes with Microbiome samples (a data management and analysis system for metagenomes)

INSDC International Nucleotide Sequence Database Collaboration

I/O input/output

iPlant NSF-funded plant science research collaborative

IPGRI International Plant Genetic Resources Institute

IPPN International Plant Phenomics Networks

IT information technology

IUPAC/ASTM International Union of Pure and Applied Chemistry/ASTM International

JGI DOE Joint Genome Institute

JGIMP DOE Joint Genome Institute Metagenomics Program

JGIMGP DOE Joint Genome Institute Microbial Genomics Program

JGIPGP DOE Joint Genome Institute Plant Genomics Program

KiSAO Kinetic Simulation Algorithm Ontology

KPI key performance indicator

KEGG Kyoto Encyclopedia of Genes and Genomes

KO KEGG Orthology Database

LIMS laboratory information management system

M5 Metagenomics, Metadata and MetaAnalysis, Models, and MetaInfrastructure initiative funded by DOE

Matlab short for matrix laboratory, Matlab is programming language and interactive numerical computing environment

MetaCyc a database of metabolic pathways

MGA MetaGeneAnnotator

MG-RAST Metagenome Rapid Annotation using Subsystem Technology

MIAME Minimum Information About A Microarray Experiment

MIAPPHE Minimum Information About a Plant Phenotyping Experiment

MIENS Minimum Information about an Environmental Sequence

MIGS Minimum Information about a Genome Sequence

MIMS Minimum Information about a Metagenome Sequence

miRNA microRNA

Mothur open-source bioinformatics software package

MPD Mouse Phenome Database

MPHASYS Mouse Phenotype Analysis System

mRNA messenger RNA

mzXML open data format for storage and exchange of mass spectrometry data

nano-SIMS nano–secondary ion mass spectrometry

NOAA National Oceanic and Atmospheric Administration

NCBI National Center for Biotechnology Information

NCBO National Center for Biomedical Ontology

NIH National Institutes of Health

NIR near infrared

NISO National Information Standards Organization

NISO MIX NISO metadata for images in XML

NSF National Science Foundation

NINJA neighbor joining (algorithms)

NITRC Neuroimaging Informatics Tools and Resources Clearinghouse

OBI Ontology for Biomedical Investigations

OptFlux: an open-source software platform for *in silico* metabolic engineering

ORF open reading frame

OrthoMCL method for constructing orthologous groups across multiple eukaryotic taxa using a Markov Cluster algorithm

OTU operational taxonomic unit

OWL Web Ontology Language

PATO phenotype, attribute, and trait ontology

PB petabyte

PCA principal component analysis

PDB Protein Data Bank

PLEXdb Plant Expression Database

PMN Plant Metabolic Pathway Database, of which PlantCyc is a central feature

PO protein ontology

PySCes python simulator for cellular systems

Q values, quality information

QC quality control

qPCR quantitative polymerase chain reaction

QTL quantitative trait locus

RAST rapid annotation using subsystem technology

RAxML software to determine phylogenetic relationships through maximum likelihood

RAM random-access computer memory

RDBMS relational database management system

RDF Resource Description Framework

RDP Ribosomal Database Project

REST Representational State Transfer; software architectural style on which the web is built

RNA ribonucleic acid

rRNA ribosomal RNA

RNA-Seq high-throughput sequencing to determine a sample's RNA content, also called whole-transcriptome shotgun sequencing

RPKM reads per kilobase of transcript target per million mapped reads

SAMBA Statistical-Algorithmic Method for Bicluster Analysis

siRNA small interfering RNAs

S statistics (used in Staffing Resources tables)

SABIO-RK System for the Analysis of Biochemical Pathway–Reaction Kinetics

SBGN: Systems Biology Graphical Notation

SBML Systems Biology Markup Language

SBO Systems Biology Ontology

SBRML Systems Biology Results Markup Language

SBW Systems Biology Workbench

SDK software development kit

SE software engineering (used in Staffing Resources tables)

SED-ML Simulation Experiment Description Markup Language

SEED open-source tool for genome annotation

SO sequence ontology

SOAP Simple Object Access Protocol; lets applications exchange information over HTTP

SOLID next-generation sequencing platform that supports a wide range of applications

SRA Short Read Archive, an NCBI database

SQL Structured Query Language

SNP single-nucleotide polymorphism

SPARQL query language for Resource Description Framework

SUNDIALS SUite of Nonlinear and DIfferential/ALgebraic equation Solvers

SVD singular value decomposition

TB terabyte

TEDDY Terminology for the Description of Dynamics

TFBS transcription factor binding site

TF transcription factor

TIGRFAMS collection of protein families

TRN transcriptional regulatory network

TSS transcription start site

TTS transcription termination site

TU transcription unit

UAL user access layer (Kbase)

UCSB University of California, Santa Barbara

UDP User Datagram Protocol

UI user interface

UniFrac online tool for comparing microbial diversity in a phylogenetic context

UniProt Universal Protein Resource

USDA United States Department of Agriculture

VISTA comprehensive suite of programs and databases for comparative analysis of genomic sequences

XML extensible markup language

XPP X-Windows Phase Plane

5' RACE rapid amplification of 5' cDNA ends

APPENDIX G

Contributors and Observers[1]

Contributors

Paul Adams
Lawrence Berkeley National Laboratory

Eduard Akhunov
Kansas State University

Eric Allen
University of California, San Diego

Martin Allgaier
Lawrence Berkeley National Laboratory

Gordon Anderson
Pacific Northwest National Laboratory

Adam Arkin
Lawrence Berkeley National Laboratory

Steve Baenziger
University of Nebraska

Scott Baker
Pacific Northwest National Laboratory

Nitin Baliga
Institute for Systems Biology

Jill Banfield
University of California, Berkeley

Ali Barakat
Pennsylvania State University

William Barbazuk
University Florida

Chris Bare
Institute for Systems Biology

Eric Beers
Virginia Tech University

Alex Beliaev
Pacific Northwest National Laboratory

Jeffrey Bennetzen
University of Georgia

David Benton
University of Wisconsin

Rex Bernardo
University of Minnesota

William Berzonsky
South Dakota State University

Devaki Bhaya
Stanford University

Paul Blum
University of Nebraska, Lincoln

Ben Bowen
Lawrence Berkeley National Laboratory

Jim Bradeen
University of Minnesota

Mya Breitbart
University of South Florida

Tom Brettin
Oak Ridge National Laboratory

Jim Bristow
Lawrence Berkeley National Laboratory

Jeff Broughton
Lawrence Berkeley National Laboratory

Charles Brummer
University of Georgia

Marcia Buanafina
Pennsylvania State University

Robin Buell
Michigan State University

John Burke
University of Georgia

Victor Busov
Michigan Technological University

[1] This list does not include participants from the *Using Clouds for Parallel Computations in Systems Biology* workshop held at the 2009 Supercomputing meeting because it was a large open meeting without a formal participants list.

Many people attended multiple workshops.

Patrick Byrne
Colorado State University

William Cannon
Pacific Northwest National Laboratory

Shane Canon
Lawrence Berkeley National Laboratory

Brian Cantwell
University of Tennessee, Knoxville

John Carlson
Pennsylvania State University

John-Marc Chandonia
Lawrence Berkeley National Laboratory

Christopher Chang
National Renewable Energy Laboratory

Amy Chen
Lawrence Berkeley National Laboratory

Dylan Chivian
Lawrence Berkeley National Laboratory

Tim Close
University of California, Riverside

James Cole
Michigan State University

Frank Collart
Argonne National Laboratory

Luca Comai
University of California, Davis

Robert Cottingham
Oak Ridge National Laboratory

Carlos Crisosto
University of California, Kearney

Richard Cronn
U.S. Department of Agriculture

Thomas Davis
University of New Hampshire

Brian Davison
Oak Ridge National Laboratory

Paramvir Dehal
Lawrence Berkeley National Laboratory

Narayan Desai
Argonne National Laboratory

Adam Deutschbauer
Lawrence Berkeley National Laboratory

Katrien Devos
University Georgia

Patrik D'haeseleer
Lawrence Livermore National Laboratory

Amit Dhingra
Washington State University

Mitch Doktycz
Oak Ridge National Laboratory

David Douches
Michigan State University

Andrew Doust
Oklahoma State University

Jorge Dubcovsky
University California, Davis

Ismail Dweikat
University of Nebraska

Ronan Fleming
University of Iceland

David Francis
Ohio State University

George Garrity
Michigan State University

Jack Gilbert
Plymouth Marine Laboratory

Bikram Gill
Kansas State University

Paul Gilna
Oak Ridge National Laboratory

Jim Giovannoni
Cornell University

Adam Godzik
Sanford-Burnham Medical Research Institute

Steve Goff
iPLANT

Jose Gonzalez
South Dakota State University

Ian Gorton
Pacific Northwest National Laboratory

Pam Green
University of Delaware

Jeff Grethe
University of California, San Diego

Arthur Grossman
Stanford University

Masood Hadi
Sandia National Laboratories

Daniel Haft
J. Craig Venter Institute

Steven Hallam
University of British Columbia

Maria Harrison
Cornell University

Caroline Harwood
University of Washington, Seattle

Loren Hauser
Oak Ridge National Laboratory

Patrick Hayes
Oregon State University

Sam Hazen
University Massachusetts

Terry Hazen
Lawrence Berkeley National Laboratory

Karla Heidelberg
University of Southern California

Alyssa Henning
Cornell University

Matthias Hess
Joint Genome Institute

Eva Huala
The Arabidopsis Information Resource

Phil Hugenholtz
Joint Genome Institute

Amy Iezzoni
Michigan State University

Eric Jackson
U.S. Department of Agriculture

Keith Jackson
Lawrence Berkeley National Laboratory

Scott Jackson
Purdue University

Janet Jansson
Lawrence Berkeley National Laboratory

Jerry Jenkins
Hudson Alpha Institute for Biotechnology

Nicholas Justice
University of California, Berkeley

Udaya Kalluri
Oak Ridge National Laboratory

Peter Karp
SRI International

Kimberly Keller
University of Missouri

Bob Kelly
University of North Carolina

James Kelly
Michigan State University

Robert Kelly
North Carolina State University

Joonhoon Kim
University of Wisconsin, Madison

Matias Kirst
University of Florida

Kerstin Kleese van Dam
Pacific Northwest National Laboratory

Steve Knapp
University of Georgia

Robin Kodner
University of Washington

Mavrommatis Konstantinos
Joint Genome Institute

Anthony Kosky
Joint Genome Institute

Julia Krushkal
University of Tennessee, Memphis

Cheryl Kuske
Argonne National Laboratory

Nikos Kyrpides
Joint Genome Institute

Miriam Land
Oak Ridge National Laboratory

Bob Landick
University of Wisconsin

Carina Lansing
Pacific Northwest National Laboratory

Jan Leach
Colorado State University

Jared Leadbetter
California Institute of Technology

Jim Liao
University of California, Los Angeles

Libbie Linton
Utah State University

Konstantinos Liolios
Joint Genome Institute

Jenny Yan Liu
Pacific Northwest National Laboratory

Miron Livny
University of Wisconsin

Thomas Lubberstedt
Iowa State University

Thauos Lykidis
Lawrence Berkeley National Laboratory

Yukari Maezato
University of Nebraska

Krishna Mahadevan
University of Toronto

Laura Marek
Iowa State University

Victor Markowitz
Lawrence Berkeley National Laboratory

Hector Garcia Martin
Lawrence Berkeley National Laboratory

Sergei Maslov
Brookhaven National Laboratory

Xavier Mayali
Lawrence Livermore National Laboratory

Michael Mazourek
Cornell University

Maureen McCann
Purdue University

Phil McClean
North Dakota State University

Susan McCouch
Cornell University

Lee Ann McCue
Pacific Northwest National Laboratory

David Mead
Lucigen Corporation

Barbara Methe
J. Craig Venter Institute

Folker Meyer
Argonne National Laboratory

Richard Michelmore
University of California, Davis

Jonathan Millen
University of Rochester

Amit Mitra
University of Nebraska

Todd Mockler
Oregon State University

Gary Muehlbauer
University of Minnesota

Lukas Mueller
Cornell University

Aindrila Mukhopadhyay
Lawrence Berkeley National Laboratory

Kristen Munch
National Renewable Energy Laboratory

Seth Murray
Texas A & M University

Gerard Muyzer
Delft University of Technology

Ambarish Nag
National Renewable Energy Laboratory

David Neale
University California, Davis

Joseph Onyilagha
University of Arkansas, Pine Bluff

Andrei Osterman
Burnham

Elizabeth Ottesen
Massachusetts Institute of Technology

Krishna Palaniappan
Lawrence Berkeley National Laboratory

Bernhard Palsson
University of California, San Diego

Jiwan Palta
University of Wisconsin

Chongle Pan
Oak Ridge National Laboratory

Nicolai Panikov
Northeastern University

Morey Parang
Oak Ridge National Laboratory

Charles Parker
Names for Life, LLC

Bahram Parvin
University of California

Cameron Peace
Washington State University

Zhaohua Peng
Mississippi State University

Andy Pereira
Virginia Tech University

Amanda Petrus
University of Connecticut

Madeleine Pincu
University of California, Irvine

David Pletcher
Lawrence Berkeley National Laboratory

Mihai Pop
University of Maryland

Iris Porat
Oak Ridge National Laboratory

Jason Raymond
University of California, Merced

Jenny Reed
University of Wisconsin

David Reiss
Institute for Systems Biology

Susanna Repo
University of California, Berkeley

Kathryn Richmond
University of Wisconsin

Errol Robinson
Pacific Northwest National Laboratory

Dmitry Rodionov
Burnham Institute

Dan Rokshar
Joint Genome Institute

Margie Romine
Pacific Northwest National Laboratory

Pam Ronald
University California, Davis

Jeffrey Ross-Ibarra
University of California, Davis

Steve Rounsley
University of Arizona

Nagiza Samatove
North Carolina State University

Herbert Sauro
University of Washington

Gary Sayler
University of Tennessee, Knoxville

John Warner Scott
University of Florida

Alexander Sczyrba
Joint Genome Institute

Joao Setubal
Virginia Bioinformatics Institute

Adrian Sharma
Massachusetts Institute of Technology

Judy Silber
Sound Vision Production

Blake Simmons
Lawrence Berkeley National Laboratory

Steve Singer
Lawrence Berkeley National Laboratory

Steve Slater
University of Wisconsin

Kevin Smith
University of Minnesota

Carol Soderlund
University of Arizona

David Spooner
University of Wisconsin

Dina St. Clair
University of California, Davis

Dan Stanzione
University of Texas

Ramunas Stepanauskas
Bigelow Laboratory for Ocean Sciences

Rick Stevens
Argonne National Laboratory

Steve Strauss
Oregon State University

Leonid Sukharnikov
University of Tennessee, Knoxville

Wesley Swingley
University of California, Merced

Ernest Szeto
Lawrence Berkeley National Laboratory

Tatiana Tatusova
National Institutes of Health

Ines Thiele
University of Iceland

Brian Thomas
University of California, Berkeley

Christian Tobias
U.S. Department of Agriculture

Susannah Tringe
Joint Genome Institute

Jerry Tuskan
Oak Ridge National Laboratory

Edward Uberbacher
Oak Ridge National Laboratory

Allen Van Deynze
University of California, Davis

Richard Veilleux
Virginia Tech University

Wim Vermaas
Arizona State University

Wilfred Vermerris
University of Florida

John Vogel
U.S. Department of Agriculture

Judy Wall
University of Missouri

Dong Wang
University of Nebraska

Zhong Wang
Lawrence Berkeley National Laboratory

Derrick White
University of Nebraska

Owen White
University of Maryland

Steven Wiley
Pacific Northwest National Laboratory

Tanja Woyke
Joint Genome Institute

Cathy Wu
University of Delaware

Shizhong Xu
University of California, Riverside

Koon-Kiu Yan
Yale University

Jian Yin
Pacific Northwest National Laboratory

Will York
University of Georgia

Janice Zale
University of Tennessee

Karsten Zengler
University of California, San Diego

Hongyan Zhu
University of Kentucky

Observers

Paul Bayer
Department of Energy

Jennifer Bownas
Oak Ridge National Laboratory

Peter Bretting
U.S. Department of Agriculture

Kris Christen
University of Tennessee

Daniel Drell
Department of Energy

Cheri Foust
Oak Ridge National Laboratory

Joe Graber
Department of Energy

Susan Gregurick
Department of Energy

Holly Haun
University of Tennessee

John Houghton
Department of Energy

Stephen Howell
National Science Foundation

Randy Johnson
U.S. Forest Service

Ed Kaleikau
U.S. Department of Agriculture

Arthur Katz
Department of Energy

William Klimke
National Center for Biotechnology Information

Shing Kwok
U.S. Department of Agriculture

Neocles Leontis
National Science Foundation

Liang-Shiou Lin
U.S. Department of Agriculture

Betty Mansfield
Oak Ridge National Laboratory

Gail McLean
U.S. Department of Energy

Larry Nagahara
National Cancer Institute

Jack Okamura
U.S. Department of Agriculture

Frank Olken
National Science Foundation

Cathy Ronning
Department of Energy

Susan Schexnayder
University of Tennessee, Knoxville

Jane Silverthorne
National Science Foundation

Marvin Stodolsky
U.S. Department of Energy

Sharlene Weatherwax
Department of Energy

Shireen Yousef
Department of Energy