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

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.

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.

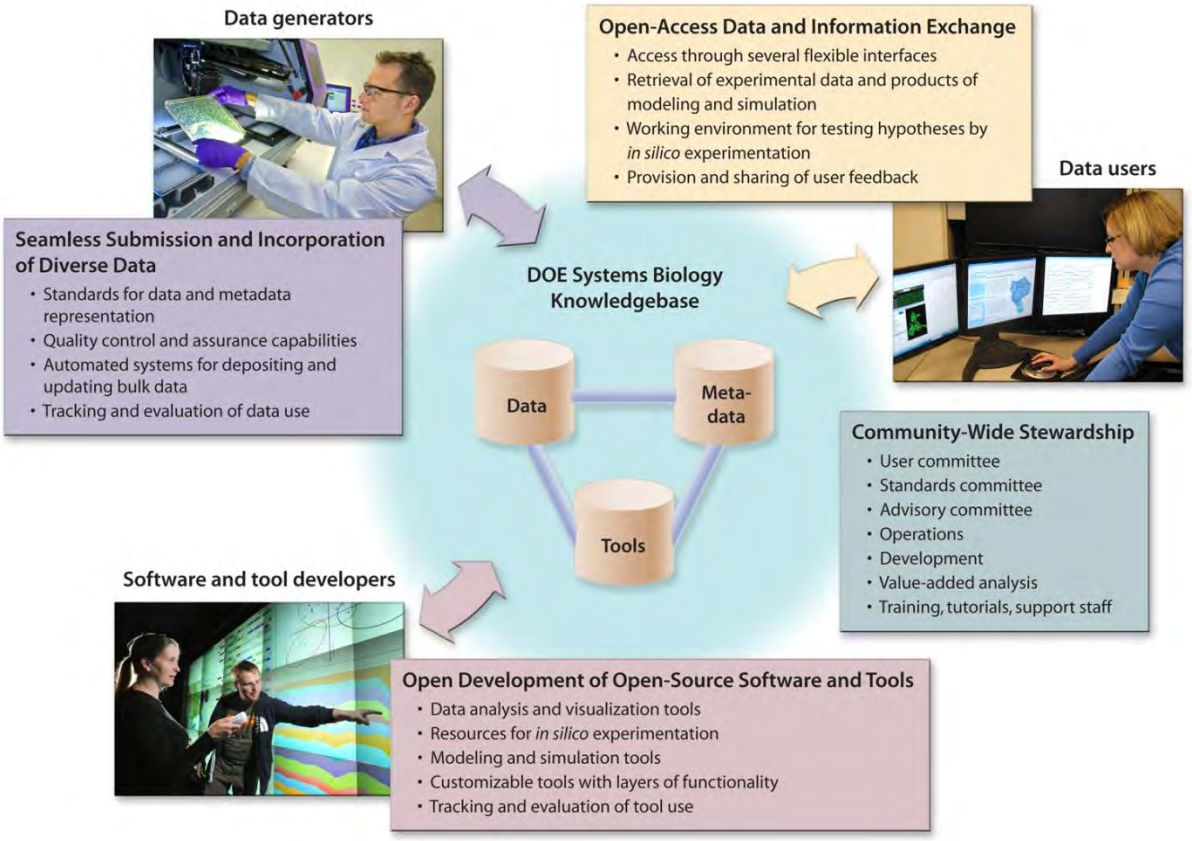


Fig. 1.1. DOE Systems Biology Knowledgebase: Establishing a Systems Biology Framework. The desired attributes and communities needed for a successful Systems Biology Knowledgebase are shown.

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).

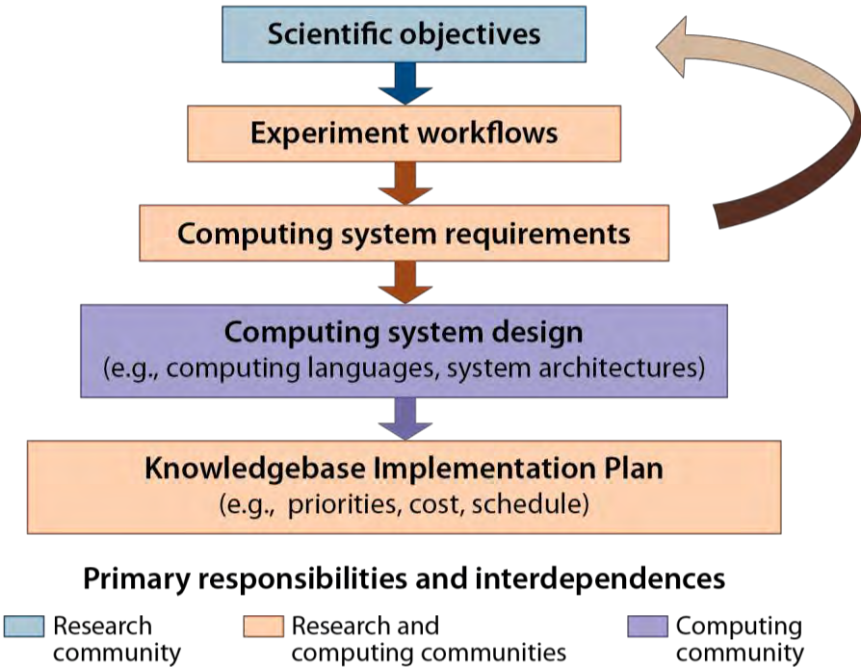


Fig. 1.2. Knowledgebase R&D Project: Scientific Objectives and Collaborations Critical to a Successful Knowledgebase Implementation Plan. The final product of this Knowledgebase R&D Project, the Knowledgebase Implementation Plan, specifies the components and functionality necessary for the systems biology 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.

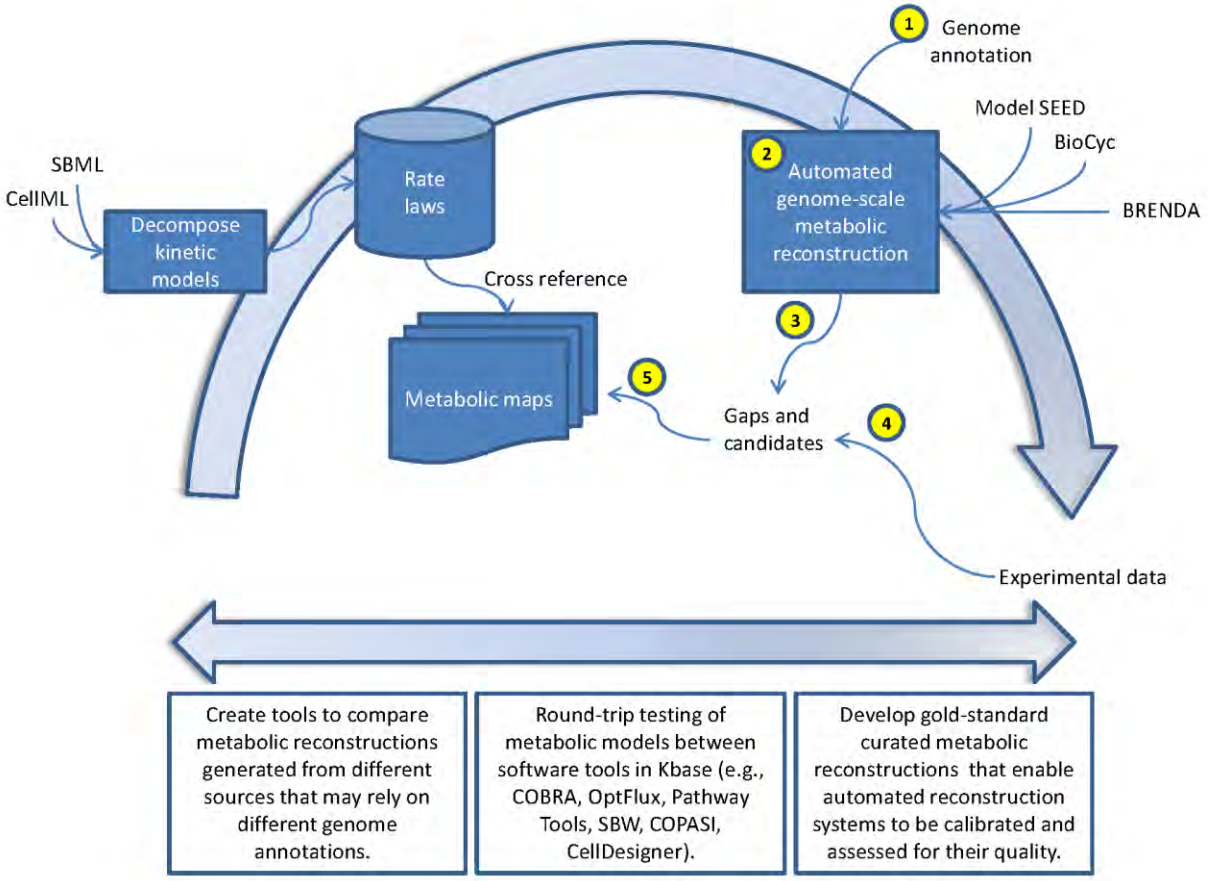


Fig. 2.1. Workflow for Reconstructing and Predicting Metabolic Networks. Based on a microbial genome annotation (1), generate a metabolic reconstruction (2) and an associated list of gaps and inconsistencies (3) that are then checked experimentally (4) and verified and corrected (5) to improve the database.

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.

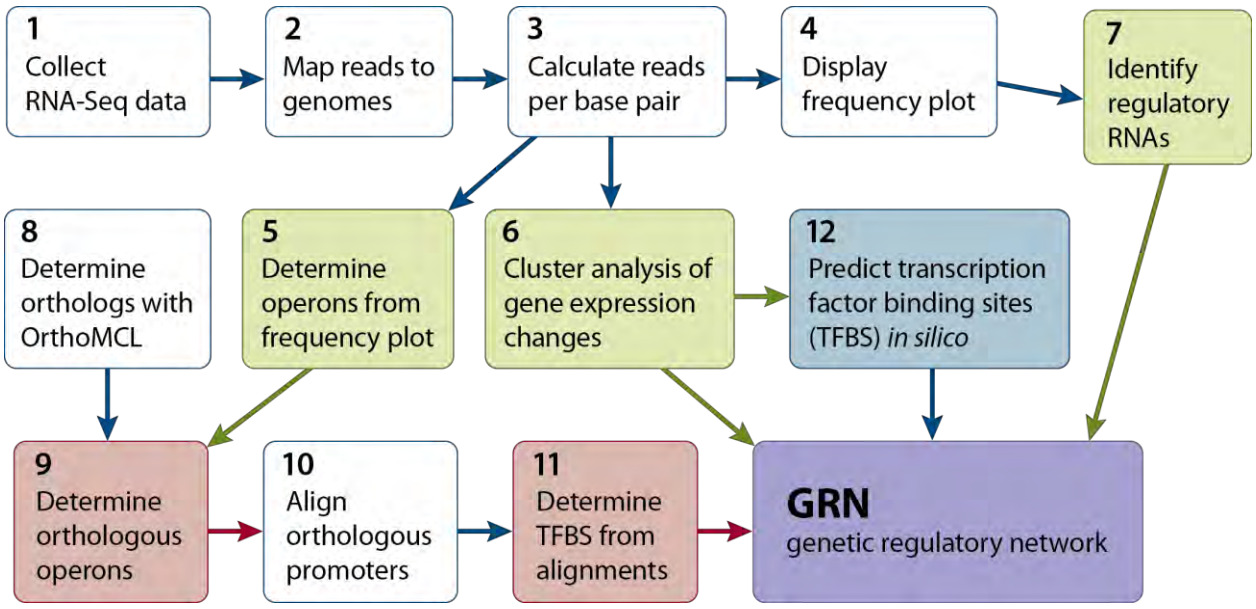


Fig. 2.2. Transcriptome Analysis Pipeline for Gene Regulatory Network Prediction. White boxes are procedures we already know how to do. Green boxes are procedures that have not been determined but are expected to be fairly easy to construct (year 1). Red boxes are procedures that will be more difficult 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.

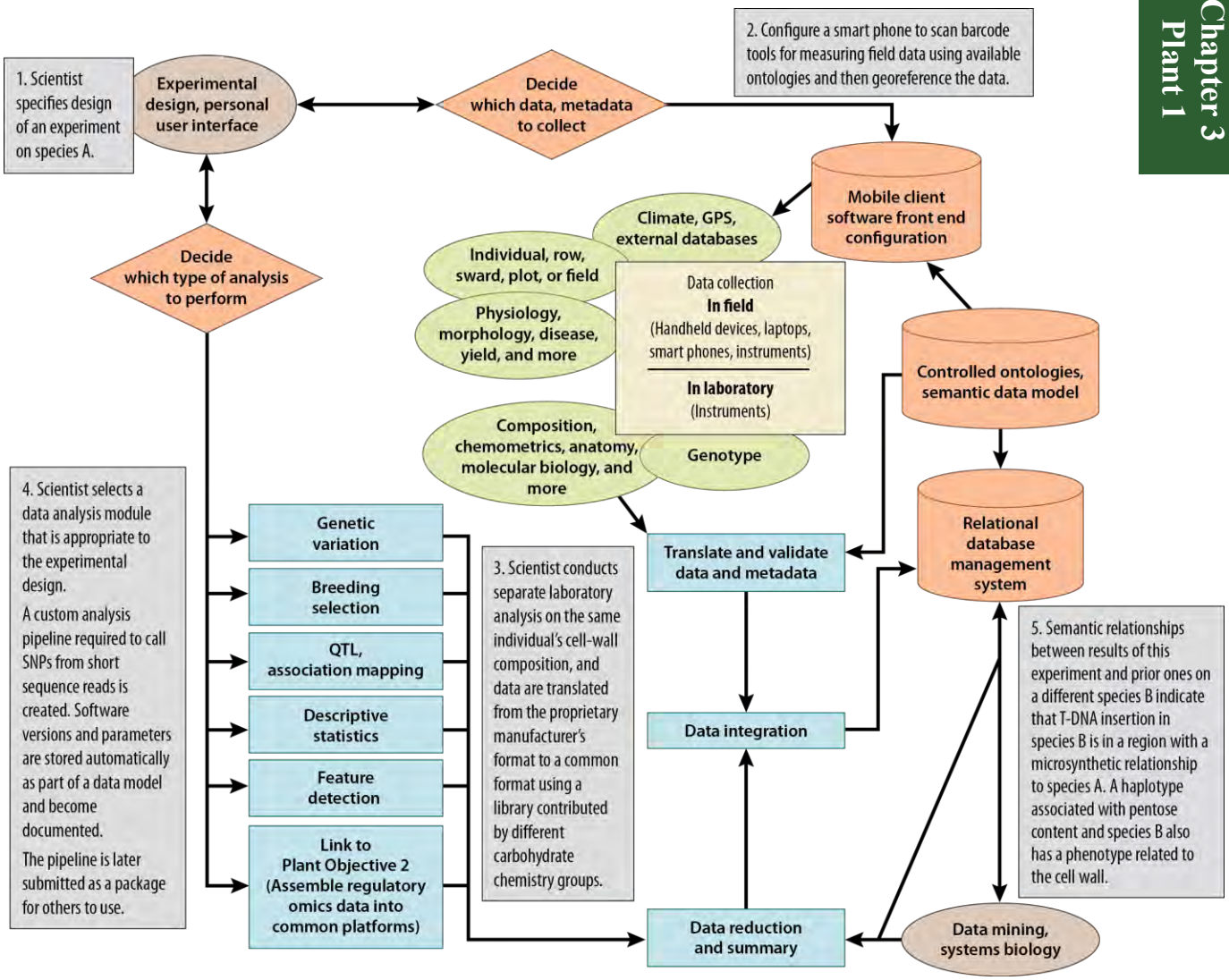


Fig. 3.1. Example Workflow for Integrating Phenotypic and Experimental Data and Metadata to Predict Biomass Properties from Genotype.

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.

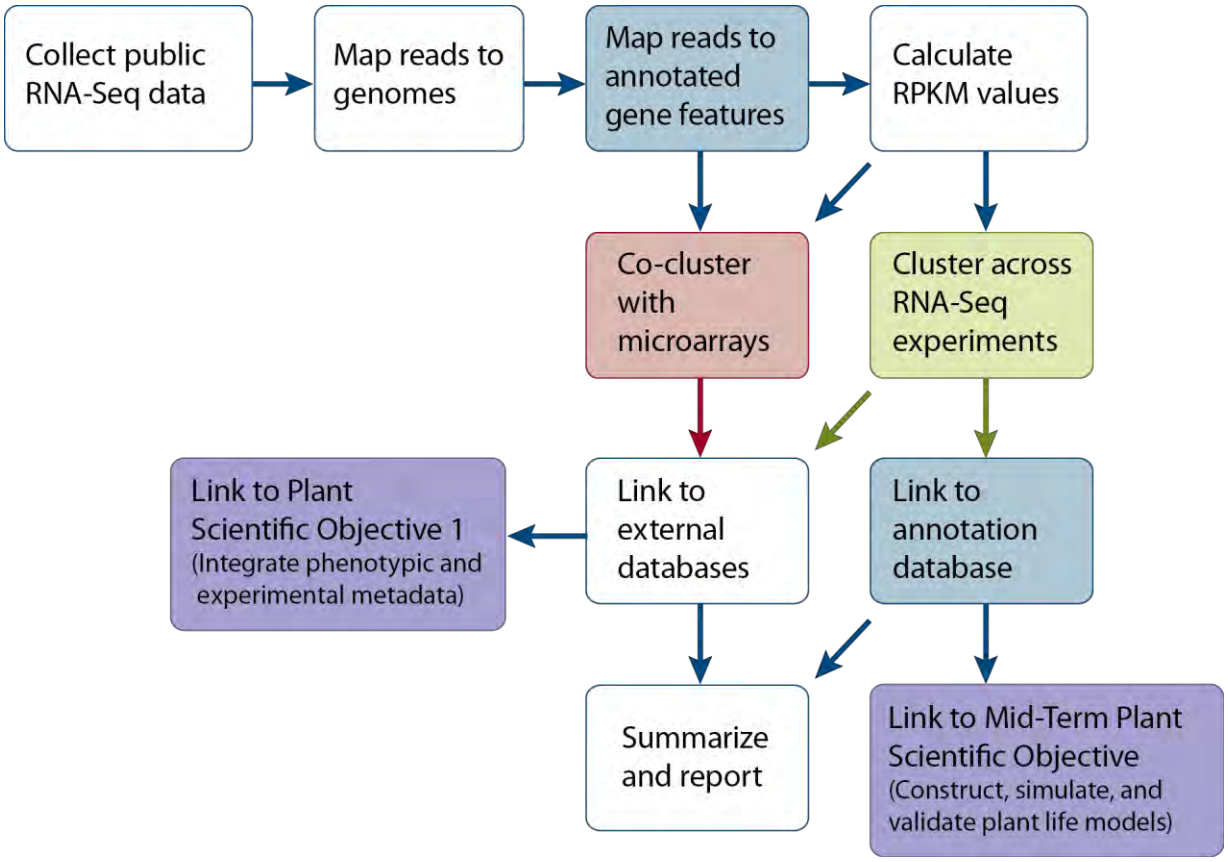


Fig. 3.2. Transcriptome Analysis Pipeline for RNA-Seq Data. White boxes are established procedures. The green box is a procedure that has not been developed but is expected to be fairly easy to construct. The red box is a procedure that will require research efforts. Blue boxes depict a linkage to existing 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.

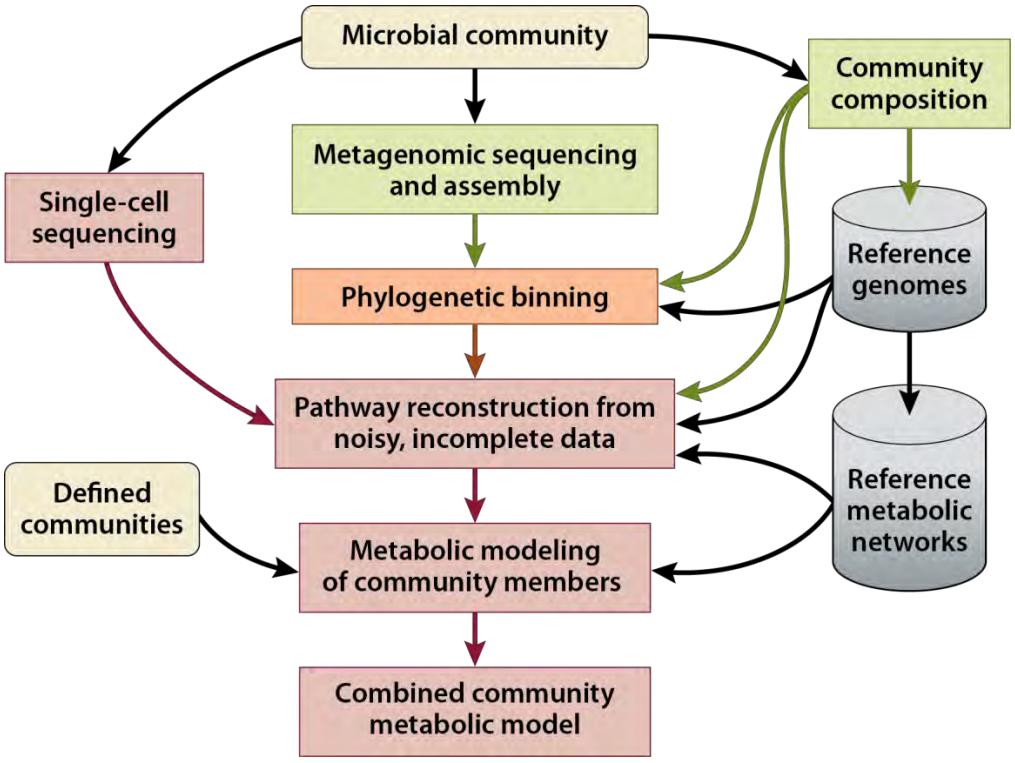


Fig. 4.1. Workflow for Reconstructing Metabolic Models from Metagenomic Data. The "bottom-up" approach involves data from individual species or members, and the "top-down" approach relies on whole-metacommunity analysis. In addition to available genomic sequencing technologies, Kbase must prepare to 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.)

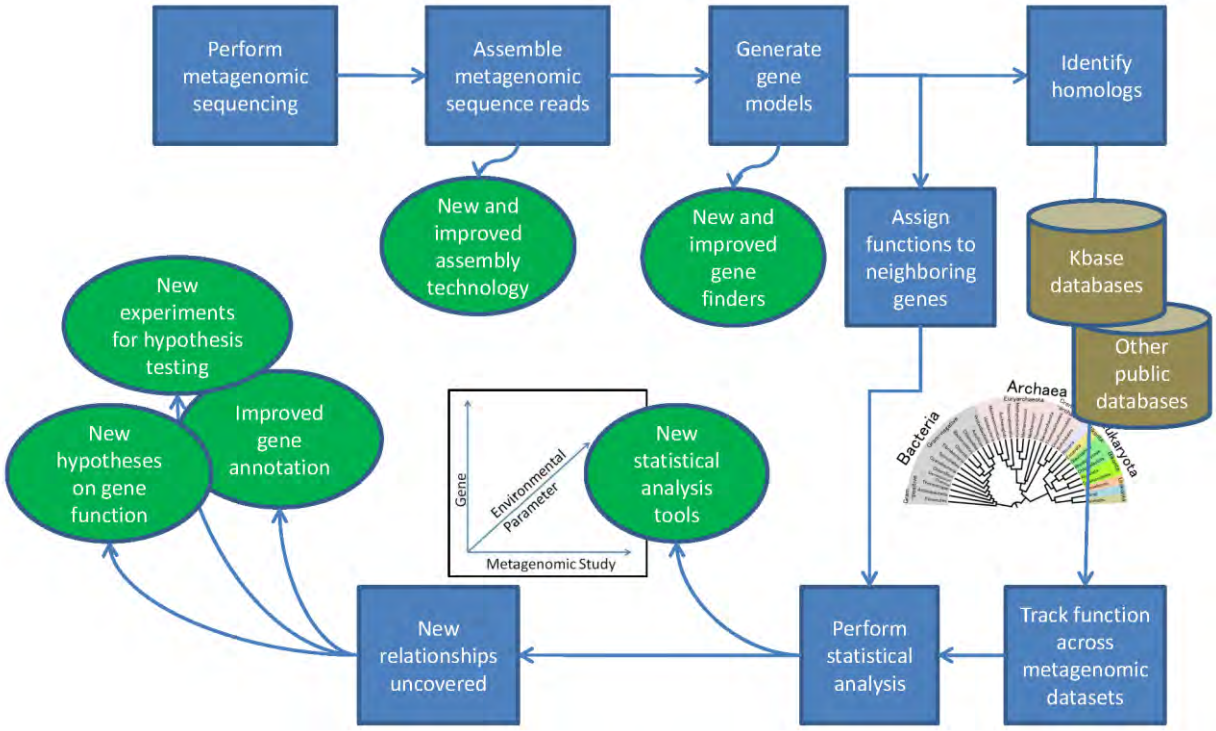


Fig. 4.2. Workflow for Mining Metagenomic Data to Identify Unknown Genes and Develop Testable Hypotheses about Their Function. Unknown genes from metagenomic studies may be assigned hypothetical functions based on their occurrence across a range of genomic and metagenomic datasets, the environmental 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).

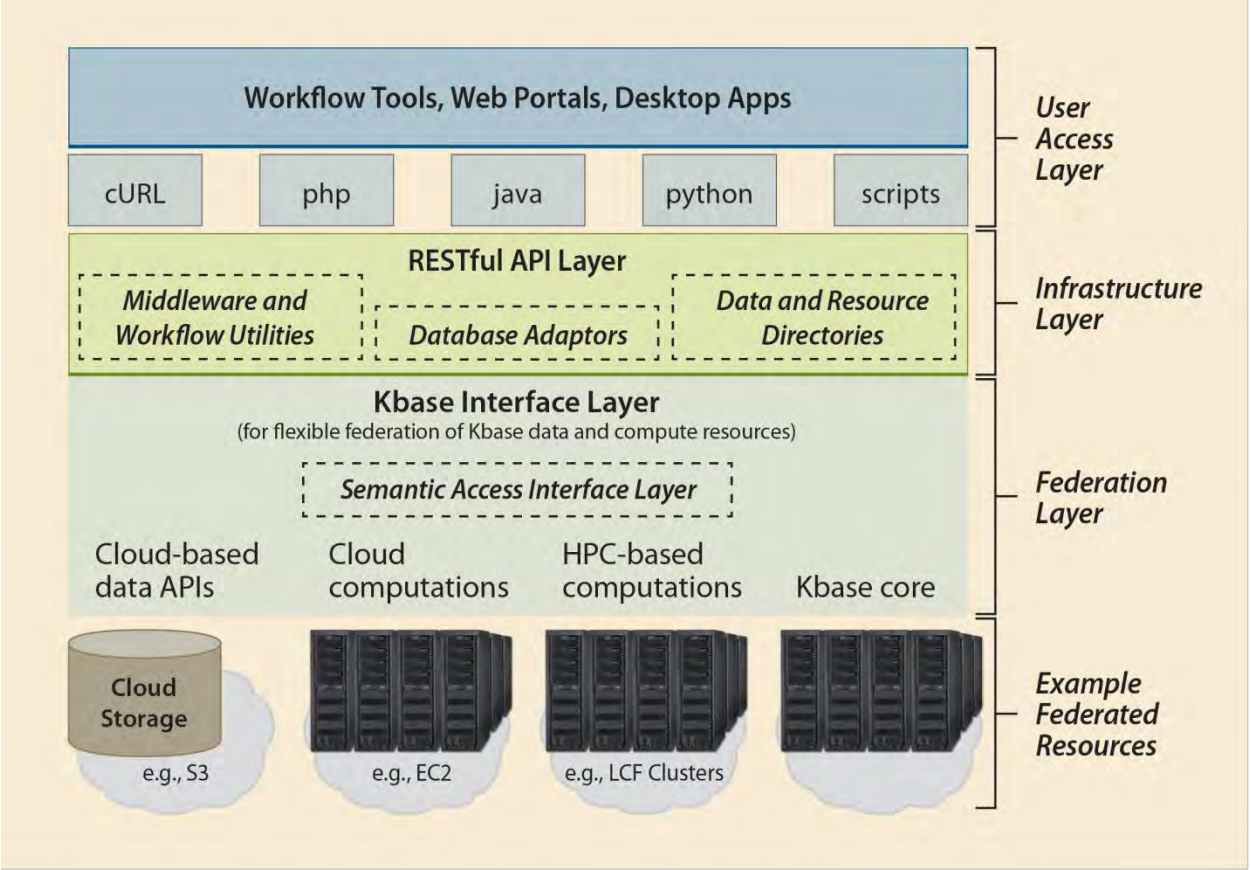


Fig. 7.1 Kbase Architecture Overview. The architecture comprises four layers: user access layer, infrastructure layer, federation layer, and federated resources. The purpose of each is described below.

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.

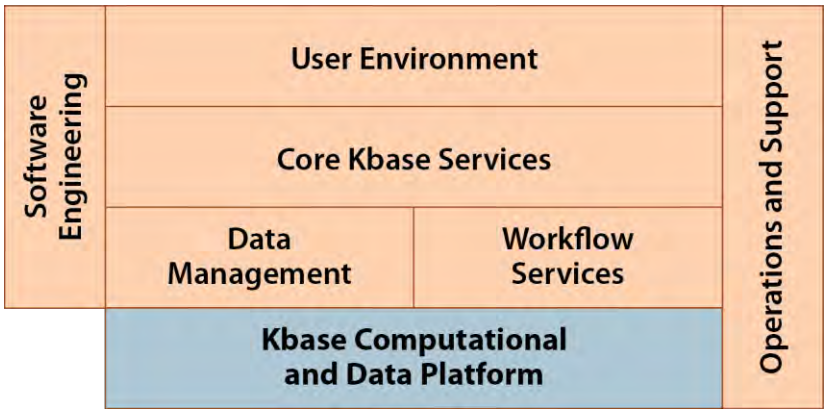


Fig. 8.1. Task Breakdown for Kbase Infrastructure Implementation.

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.

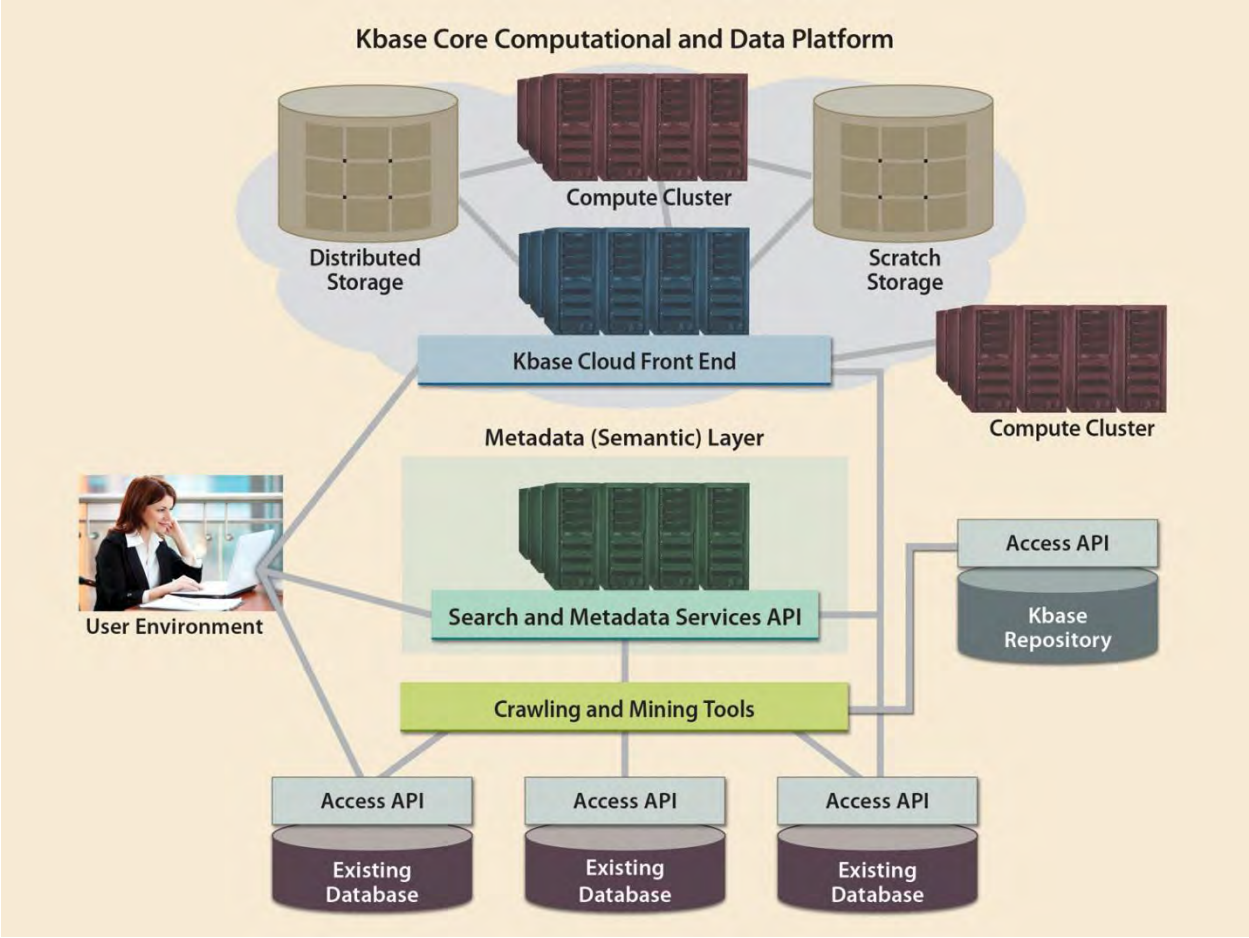


Fig. 8.2. Kbase Core Computational and Data Platform.

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.

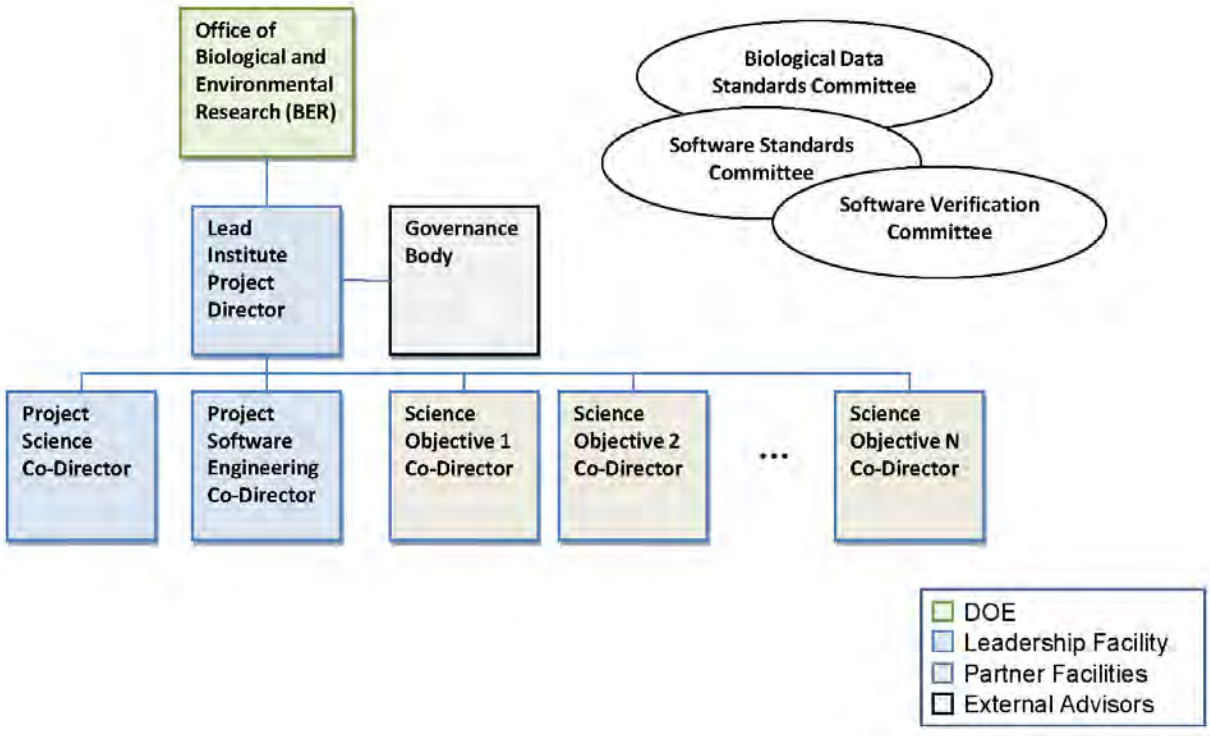


Fig. 10.1. Illustrative Management Structure.

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.

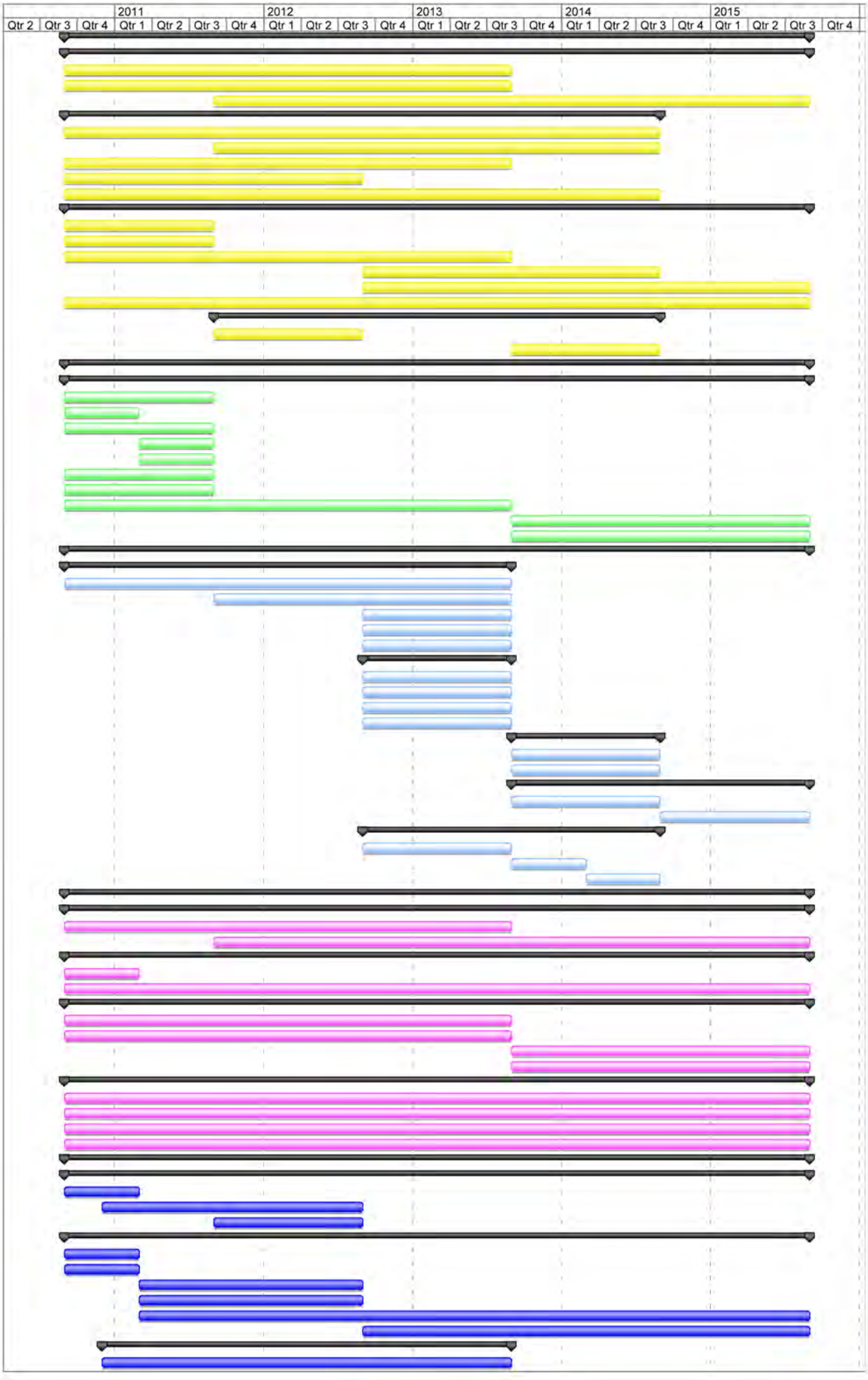
Table 10.1. Potential Risks and Mitigation Strategies.

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.

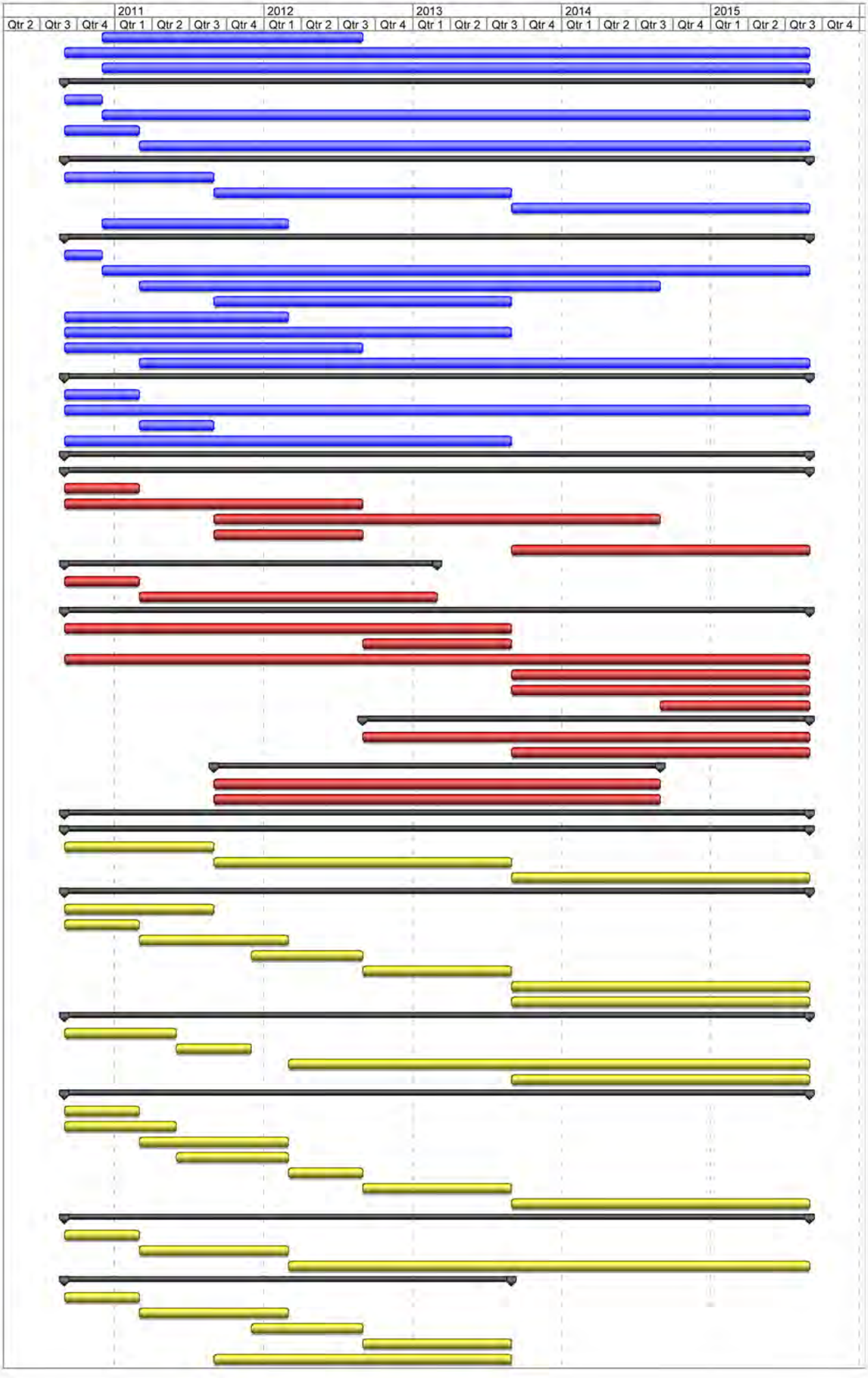
Table 10.1. Potential Risks and Mitigation Strategies.	
4. Cluster analysis on these datasets requires more resources than currently anticipated.	Modify algorithms by accepting some additional error in return for higher performance speed. Allow clustering on subsets to manually find the optimum with 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