DOE Knowledgebase System Development Workshop Report

June 1–3, 2010, Crystal City, Virginia

Workshop Organizers: Susan Gregurick (DOE) and Bob Cottingham and Brian Davison (Oak Ridge National Laboratory)

Table of Contents

1. Introduction

This report reviews discussion and material associated with the Department of Energy (DOE) Knowledgebase System Development workshop held June 1 - 3, 2010. The goal of this workshop was to establish initial actionable plans to create the Knowledgebase. The first day focused on the prioritization of clear scientific objectives and specific requirements for the Knowledgebase derived from these objectives. The second day focused on the development of an implementation plan, system architecture, and governance for the initial system. The last day focused on finishing writing assignments leading to the Final Report, which will be the plan for creating the Knowledgebase.

First, a background summary is given below describing the purpose of the DOE Systems Biology Knowledgebase (Kbase) planning project. Next is a summary of pre-workshop activities, topics presented and discussed during the workshop, and post-workshop activities.

Since the goal of the Knowledgebase planning project is to develop an initial prioritized plan for a useful systems biology knowledgebase, there is a continued consensus that these initial efforts cannot be all things for all users. It is better to show strong success in a few areas than minimal progress in many areas. There was also continued consensus on the principles from past workshops on which Kbase is being founded that (1) science drives Knowledgebase development, (2) the project be a community effort, (3) that it be open access and open contribution, and (4) that it be distributed.

2. Background

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

Knowledgebase Vision and Background

The emergence of systems biology as a research paradigm and approach for DOE missions has resulted in dramatic increases in data flow from a new generation of genomics-based technologies. To manage and effectively use this ever-increasing volume and diversity of data, the Genomic Science program is developing the DOE Systems Biology Knowledgebase—an open, community-driven cyberinfrastructure for sharing and integrating data, analytical software, and computational modeling tools. Historically, most bioinformatics efforts have been developed in isolation by people working on individual projects, resulting in isolated databases and methods. An integrated, community-oriented informatics resource, such as the Knowledgebase, would provide a broader and more powerful tool for conducting systems biology research relevant to BER's complex, multidisciplinary challenges in energy and environment. It also would be easily and widely applicable to all systems biology research.

In general, a knowledgebase is an organized collection of data, organizational methods, standards, analysis tools, and interfaces representing a body of knowledge. For the DOE Systems Biology Knowledgebase, these interoperable components would be contributed and integrated into the system over time, resulting in an increasingly advanced and comprehensive resource. Other elements of the Knowledgebase vision are defined in a March 2009 report (genomicscience.energy.gov/compbio/) based on a DOE workshop that brought together researchers with many different areas of expertise, ranging from environmental science to bioenergy. The report highlights several roles the Knowledgebase will need to serve.

Workshop Background

To develop a successful open informatics endeavor for systems biology (the Kbase effort), a series of workshops have been held to include key stakeholders (plant and microbial genomic researchers, bioinformaticians, computer scientists, database developers, and software engineers) and to elicit their goals, challenges, and expectations for the development and management of the Kbase. This final workshop was a culmination of these previous workshops to provide clear prioritization and tasks to allow the final design and implementation of the Kbase to be developed. The workshop was held June 1–3 and involved 80 participants representing university, national laboratory, and international researchers. In addition, the workshop had representation from DOE's Joint Genome Institute; DOE's Bioenergy Research Centers; NSF's iPLANT; and NIH's NCI and NCBI. The goal of the workshop is to develop a robust design and implementation plan for the Systems Biology Knowledgebase. The participants were charged with developing and prioritizing 3 to 5 scientific objectives in each of the areas of microbial, meta-communities, and plant research. From these scientific objectives, two days were spent developing scientific requirements, time frames, and effort for each of the scientific objectives. An additional half day was spent discussing detailed plans for architecture, implementation, and governance. Extensive pre-meeting conference calls helped to lay the groundwork of the science objectives. Participants were not charged to define funding or contractual structures, and they are continuing to finalize requirements based on the discussed objectives and transfer these into an implementation plan.

Outlined below are the prioritized scientific objectives and rough time frames for the implementation of these objectives. The details of the requirements of each objective as well as the architecture and governance plans will be developed over the summer, culminating in a final implementation plan report by September 30, 2010.

3. Pre-Workshop Activities

3a. Conference Calls

A series of conference calls were scheduled in May before the workshop. The first of these were to organize the three science area breakout leads. Each science area (Plant, Microbial, and Meta-Communities) had two leads for Scientific Objectives and two leads for Requirements. Then calls were held with all members of each breakout. The first call was to introduce the workshop and define what is meant by a Scientific Objective, and the second call focused on reviewing the Scientific Objective template and beginning discussion on what would be the recommended Strawman List of Scientific Objectives for each breakout. At the workshop, the Strawman List would be reviewed and finalized on the first day, and participants would establish the consensus priority (High, Medium, Low) and feasibility (Near: 1-3 years; Mid: 3-5 years; Long 5-10 years). Based on this, the top 3 to 5 Scientific Objectives would be the focus of the initial Kbase.

A third call introduced the template for Requirements and how these would be derived from a Scientific Objective. The most detailed and complete Requirements write-ups are needed for the top Scientific Objectives, with decreasing detail needed for mid- and long-term Objectives.

3b. Templates

The Scientific Objectives and Requirements templates—along with filled-out examples that were given to the participants—are included in Appendix 3 and 4, respectively. These provide focused guidance toward establishing the most important objectives and detailed requirements that guide Kbase development.

3c. Google Docs

In order to begin rapid development of the Scientific Objectives, a Google Docs folder was established for each breakout group. Writing teams then formed around each proposed Scientific Objective, and a significant amount of preliminary writing was accomplished in advance of the meeting. During the conference calls, multiple participants would edit the draft documents as they were being discussed. Initially, these areas were accessible only by members of the breakout. At the workshop all areas were made accessible to all participants.

3d. Goal to Establish Scientific Objectives and Requirements

For most attendees, this was a different kind of workshop. Its primary focus was on establishing the best Scientific Objectives and Requirements for the DOE Systems Biology Knowledgebase, especially the high-priority requirements for the first 1-3 years. The Requirements are the most important result of the workshop, as these define what the initial Kbase will be. The Science Area breakouts first focused on the Scientific Objectives, and then in the second half of the first day, the Breakout Leads switched and the focus was on Requirements with the same Breakout group. This process allows an easier transition from objectives to requirements and encourages feedback so the objectives are tractable.

4. Topics Discussed at Workshop

4a. Microbial Science Objectives

1. Integrated Description of Genomic Features

Summary: This objective will create the ability to represent and update experimental data and inferred knowledge about genes and genomes so that the experimental and computational results drive progressively richer and more accurate gene models and predictions. This ability would allow users to access existing genomic sequence information, upload new experimental data in order to define and refine models, and test consistency between the two. This objective was given high priority, as many other objectives require this ability to build on. This objective requires integration with JGI/IMG and NCBI and will require some standards development for data and access to large-scale computing resources. This objective will take 1-3 years.

2. Reconstruction, Prediction, and Manipulation of Metabolic Networks

Summary: The scientific objective is to provide a method to evaluate the metabolic potential of an organism, predict the phenotypic outcome of specific metabolic or environmental interventions and perturbations, and establish metabolic kinetics capabilities and fluxes for short-term dynamic responses. This knowledge will lead to the informed

modification of one or more specific enzymes or the introduction of entirely new enzymes and/or pathways for metabolic engineering purposes. This objective would allow the community to better determine strategies for carbon flow manipulation and for understanding microbial communities. This objective requires integrating new experimental data with known data and models on metabolic pathways, as well as developing methods to automatically create new metabolic reconstructions from newly sequence organisms. This objective requires linking together known metabolic models with databases such as chEBI, UniPROT, KEGG, and GO with experimental data. This objective is given medium priority (3-5 years), and it is suggested to apply this objective to a selected set of organisms relevant to DOE's research efforts.

3. Microbial Gene Expression Regulatory Networks

Summary: The scientific objectives can be broadly divided into two components. The first is to enable automated inference of gene expression regulatory networks relying principally on expression profiling data. The second is to extend these inferred networks to include additional data types, both to refine the network predictions and to test them. The availability and evolution of genome-scale expression data and the rapid extension into new data types makes the definition of microbial gene expression regulatory networks an attractive goal of the Kbase project. In the short-term, inference of regulatory networks from just genome sequence and expression profiles under varied cellular conditions is possible and could be of general utility to researchers in constructing and understanding of carbon and nitrogen processes. Interconnection of the regulatory networks with metabolic reconstructions and multidimensional annotations (two other high-priority objectives identified by the Kbase microbial group) would greatly facilitate development of microbial systems biology. This objective could work synergistically with NIH pathway tools, EcoCYC, and DOE efforts such as MicrobesOnline and JGI. Much of the experimental work would come from the Bioenergy Research Centers and the larger DOE science-focused work on microbial systems. This project was given high priority. This objective can achieve some near-term goals but may take 2-10 years to complete. It was suggested to work on DOE-related organisms and in coordination with the second scientific objective.

4b. Metagenomics/Meta-Communities Science Objectives

1. Metabolic Modeling of Microbial Communities

Summary: This objective focuses specifically on modeling the metabolic processes within a microbial community, since this topic most directly ties into developing metagenomics workflows and the single microbial organisms systems biology tools (above). This predictive understanding of communities will progress in three stages: **(1) Understanding**: Descriptive models that provide insight into the metabolic role of the members within the community and their interactions. **(2) Prediction**: Predictive models that allow us to simulate the metabolic processes in the community and the response of community activity or composition to environmental conditions. **(3) Manipulation**: Eventually, these models will allow us to not only predict, but actively drive changes in the community into desired directions (e.g., to accelerate environmental processes such as bioremediation, cellulose

degradation, or carbon sequestration). This objective outlined as a first step the Knowledgebase need to develop workflows for analyzing metagenomes of a microbial community and to leverage existing data to create community metabolic models. This objective was seen as a medium priority (3-5 years) and would require leveraging existing tools (IMG, MG-RAST, CAMERA) and databases (BioCyc, KEGG) as well as developing analysis tools.

2. Expand Our Understanding of Poorly Studied Genes

Summary: Data generated in large-scale metagenomics projects can provide the information necessary to better understand the function of poorly characterized genes. This scientific objective is to develop approaches for (1) mining the data in order to identify previously unknown genes and (2) leveraging the wealth of metadata associated with metagenomic datasets, as well as gene/organism co-occurrence information in order to identify testable hypothesis about the function of newly identified or poorly characterized genes. This objective was given high priority and could leverage all of the metagenomic sequencing efforts from DOE and NIH.

3. Analysis of Understudied Microbial Phyla

Summary: The goal of this objective is to understand the role of unclassifiable members of a microbial community in terms of genetic and phenotypic comparison. To achieve this scientific objective, a specific requirement will be linking physiologic and metabolic datasets to metagenome annotations in order to provide context and evidence. This will create a product that is more informative and flexible. The specific datasets that will be utilized are the genomes and accompanying physiologic and metabolic data of understudied microbial phyla. Questions that this objective would address are: (1) where are members of a novel phylum found, (2) how do we facilitate phylogenetic binning to preclude assignment as orphan genes, and (3) what are the emerging concepts of their metabolomes? This objective was given medium priority (3-5 years) and requires the development of infrastructure and tools to accomplish the goals. This will likely be merged into Objective 2.

4. Metagenomic Interpretation to Identify Conditions Required for Growth by Key Microbial Communities Relevant to DOE Missions.

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

4c. Plant Science Objectives

1. Integration of Phenotypic and Experimental Metadata to Enable Prediction of Biomass Properties based on Genotype

Summary: Improvements in computational infrastructure are required to support and contextualize experimental plant phenotype data to an extent that will enable one to predict the changes in the physical properties of biomass properties that occur as a result of environmental changes and genetic diversity or manipulation. Achievement of this ambitious goal depends on the creation of robust semantic infrastructure for collection, annotation, and storage of diverse phenotypic and environmental datasets. These data include measurements such as photographic images and analytical spectra that capture visible phenotypes and chemotypes that are fundamentally related to yield and physiological performance and sustainability. Specifically, this infrastructure will be used as a basis for software applications that extract, quantify, and catalogue phenotypic features from the data for the purpose of data mining and further analysis. This involves association of the data with relevant metadata to enable querying, modeling, clustering, and comparison of the data from diverse datasets generated by different platforms. Attainment of the scientific objective requires appropriate vocabulary standards for wide variety of data and metadata that describe phenotypes, chemotypes, genotypes, and the experiments designed to collect this data. Although several such standards and ontologies exist, they require additional expressiveness to achieve the objective. In order to share the relevant experimental data and ensure its completeness (in terms of associated metadata, etc), a community approved standard for the Minimum Information for A Plant Phenotyping Experiment (MIAPPHE) would be helpful. However, such a standard does not currently exist. The development of all of these standards demands a long-term, committed collaboration between computer scientists and plant scientists. This objective was seen as high priority and could be carried out in 3 to 5 years. This would require a community of scientists to agree to standards of data to describe phenotypes and needs to be coordinated with iPLANT.

2. Assemble Regulatory Omics Data in Common Platforms to Enable Annotation, Comparisons, and Modeling

Summary: This objective will integrate several key types of regulatory omic data and associated quality and metadata for six target plant species: *Brachypodium*, *Chlamydomonas*, poplar, sorghum, switchgrass, and *Miscanthus*. This information will support the other objectives, including annotation, comparison, and modeling. RNA levels as measured by expression arrays or RNA-Seq are no longer sufficient to evaluate mechanisms and networks that regulate plant transcriptomes. The Kbase must also include available small RNA and target RNA information, differential RNA processing and decay information, and epigenetic marks such as DNA methylation and histone modifications. This information is important for data integration and to fill in important missing links in gene regulatory networks within a species and facilitate their comparison across two or more species. In the short term (1-3 years), classical transcriptome data (microarrays and mRNA

seq) as well as small RNA and basic proteomic data will be assembled. Epigenetics data, small RNA target data/RNA degradome data, other types of RNA processing data, and additional proteomic data will be assembled after year one, with the most developed genomes such as *Brachypodium* beginning first. The data will be made publicly accessible with user-friendly web interfaces and downloadable for power users. The acquired data will include sequences, quality information (e.g., Q values) and associated metadata. Sources will include NCBI (GenBank, GEO, SRA), the DOE JGI, ArrayExpress, and PLEXdb. This was given high priority and could be accomplished in 1-3 years. This requires collaboration with iPANT and USDA for selection of relevant species.

3. Improvement and Availability of Plant Genome Annotation Datasets

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

4. Modeling, Simulation, and Validation

Summary: Enable semi-automated inference, construction, simulation, validation, and query of complex multilevel (gene, protein, metabolite, small RNA, organelle, cell, and tissue) models of plant life, with a focus on models useful for integration and exploration of experimental data types collected during study of biomass recalcitrance, the carbon cycle, and bioremediation. Four sub-objectives proposed herein are automation and streamlining of model construction; development of a semi-automated model validation process; development of advanced semantic querying capability targeted to biological models and representations; and phylogenetic inference of functional networks (itself a model construction exercise). Model construction and validation are very closely aligned with Kbase objectives. Exploratory model construction is completely dependent on a conceptual framework, together with multiple datasets (annotated genome, proteomic, metabolomic, transcriptomics) to populate instances of this framework. Validation depends on well-structured and -annotated experimental data. At the same time, the dependencies are modular, which facilitates separate development of software for specific or more

generalized tasks. Semantic query will enable scientists to more rapidly and precisely develop hypotheses and conclusions from the complex metabolic and regulatory models that arise from genome-scale studies. This science objective requires interfacing with existing plant genomic databases as well as KEGG, GO, Metacyc, PMN. This was given high priority but was also noted to take up to 10 years in stages.

4d. Computational Area Breakouts: System Architecture, Implementation Plan, and Governance

On the second day of the workshop, a follow-up set of breakouts was held to address the major topics associated with constructing the Knowledgebase computation system. The System Architecture group is working to establish the technical principles and basis for recommending specific System Architecture, considering specific architectural attributes and their relative priorities. The Implementation Plan group is evaluating each Scientific Objective and associated Requirements to assess the major tasks and recommended plan for implementation. The Governance group is considering and will recommend a governance model and principles that will guide the development, management, and operation of the Knowledgebase for the research community. Based on the Kbase vision, principles, and scientific objectives, each of these groups is working toward writing up recommendations for the associated sections of the Final Report.

5. Post-Workshop Plan

Each breakout topic group is finalizing its write-ups with a June 30 deadline. The focus has been on completing the Scientific Objectives and Requirements and then on integrating these into a Science Area report that will become part of the Final Report. In parallel, work is under way to create the Implementation Plan section for each of the Scientific Objectives that typically focuses on the 3-4 major required tasks and then the associated effort and expertise recommended to accomplish the tasks.

A follow on writing meeting will be held in July that will focus on finalizing the Implementation Plan for each Scientific Objective.

Appendix 1: Agenda

DOE Knowledgebase System Development Workshop

Crystal City, Virginia
Tuesday, June 1, 2010

June 1, 2010

9:00 a.m. – 9:10 a.m.	Welcome, Susan Gregurick
9:10 a.m. – 10:00 a.m.	Workshop Objectives and Expectations, Bob Cottingham
10:00 a.m. – 12:00 p.m.	Divide into Three Breakout Groups
	Microbial Communities — Scientific Objectives Breakout Leaders — Jim Liao and Wim Vermaas
	Plant Communities — Scientific Objectives Breakout Leaders — Maureen McCann and Pam Green
	Meta-Communities — Scientific Objectives Breakout Leaders — Jack Gilbert and Jared Leadbetter
10:30 a.m. – 10:45 a.m.	Break
12:00 p.m. – 12:30 p.m.	Working Lunch
12:30 p.m. – 2:00 p.m.	Breakout Groups report back on Scientific Objectives and Priorities
2:00 p.m. – 4:00 p.m.	Divide into Three Breakout Groups
	Microbial Communities — Requirements Breakout Leaders — Bernhard Palsson and Bob Landick
	Plant Communities — Requirements Breakout Leaders — Robin Buell and Will York
	Meta-Communities — Requirements Breakout Leaders — Steve Slater and Jeff Grethe
3:00 p.m. – 3:15 p.m.	Break
4:00 p.m. – 5:30 p.m.	Breakout Groups report back on Requirements and Priorities
5:30 p.m. – 5:45 p.m.	Conclusions and Adjourn, Bob Cottingham
6:30 p.m.	Working Dinner for Chairs and Breakout Leaders

June 2, 2010

8:00 a.m. – 8:15 a.m.	Recap of June 1, Bob Cottingham
8:15 a.m. – 10:00 a.m.	Divide into Three Breakout Groups

	Microbial Communities — Requirements Breakout Leaders — Bernhard Palsson and Bob Landick
	Plant Communities — Requirements Breakout Leaders — Robin Buell and Will York
	Meta-Communities — Requirements Breakout Leaders — Steve Slater and Jeff Grethe
10:00 a.m. – 10:15 a.m.	Break
10:15 a.m. – 12:00 a.m.	Breakout Groups report back on Final Requirements
12:00 p.m. – 12:30 p.m.	Working Lunch
12:30 a.m. – 4:00 p.m.	Divide into Three Breakout Groups
	System Architecture Breakout Leaders — Ian Gorton and Dan Stanzione
	Implementation Plan Breakout Leaders — Peter Karp and Ed Uberbacher
	Governance Breakout Leaders — Miron Livny and Steve Goff
3:00 p.m. – 3:15 p.m.	Break
4:00 p.m. – 5:30 p.m.	Breakout Groups report back on System Architecture, Implementation Plan, and Governance
5:30 p.m. – 5:45 p.m.	Conclusions and Adjourn, Bob Cottingham

June 3, 2010

9:00 a.m. – 9:40 a.m.	Recap of June 1st and 2nd, Bob Cottingham
9:40 a.m. – 10:30 a.m.	Writing Assignments
10:30 a.m. – 11:00 a.m.	Break
11:00 a.m. – 12:30 p.m.	Group Recap, Bob Cottingham • Where we are? • Missing pieces • Assignments
12:30 p.m. – 1:00 p.m.	Working Lunch
1:00 p.m. – 4:45 p.m.	Continue work
3:00 p.m. – 3:30 p.m.	Break
4:45 p.m. – 5:00 p.m.	Conclusions and Adjourn, Bob Cottingham

Knowledgebase Wiki: sites.google.com/a/systemsbiologyknowledgebase.org/kbase/

Appendix 2: Participants and Observers

Participants

Baliga, Nitin (Institute for Systems Biology)
Beliaev, Alex (PNNL)
Benton, David (GLBRC)
Blum, Paul (University of Nebraska, Lincoln)
Bowen, Ben
Brettin, Tom (ORNL)
Buell, Robin (Michigan State University)
Cannon, Bill (PNNL)
Canon, Shane (LBL)
Chang, Christopher (NREL)
Chivian, Dylan (JBEI/LBL)
Collart, Frank (ANL)
Cottingham, Bob (ORNL)
Desai, Narayan (ANL)
D'haeseleer, Patrik (LLNL)
Gilbert, Jack (Plymouth Marine Laboratory)
Gilna, Paul (BESC/ORNL)
Godzik, Adam (Sanford-Burnham Medical Res. Inst.)
Goff, Steve (iPLANT)
Gorton, Ian (PNNL)
Green, Pam (University of Delaware)
Grethe, Jeff (University of California, San Diego)
Haft, Daniel (J. Craig Venter Institute)
Jackson, Keith (LBNL)
Jenkins, Jerry (Hudson Alpha Inst. for Biotechnology)
Kalluri, Udaya (BESC/ORNL)
Karp, Peter (SRI International)
Kelly, Bob (University of North Carolina)
Kleese van Dam, Kerstin (PNNL)
Landick, Bob (University of Wisconsin)
Lansing, Carina (PNNL)
Leadbetter, Jared (California Inst. of Technology)
Liao, Jim (University of California, Los Angeles)
Liu, Jenny Yan (PNNL)
Livny, Miron (University of Wisconsin)
Mahadevan, Krishna (University of Toronto)
Markowitz, Victor (LBNL/JGI)
Maslov, Sergei (BNL)
McCann, Maureen (Purdue University)
McCue, Lee Ann (PNNL)
Methe, Barbara (J. Craig Venter Institute)
Meyer, Folker (ANL)
Mockler, Todd (Oregon State University)
Osterman, Andrei (Burnham)
Palsson, Bernhard (University of California, San Diego)
Pop, Mihai (University of Maryland)
Reed, Jenny (University of Wisconsin)
Romine, Margie (PNNL)
Samatove, Nagiza (North Carolina State University)
Sayler, Gary (University of Tennessee, Knoxville)
Setubal, Joao (Virginia Bioinformatics Institute)
Slater, Steve (GLBRC)
Stanzione, Dan (University of Texas)
Stevens, Rick (ANL)
Tatusova, Tatiana (NIH)
Tobias, Chris (USDA)
Uberbacher, Edward (ORNL)
Vermaas, Wim (Arizona State University)
White, Owen (University of Maryland)
Wu, Cathy (University of Delaware)
Yan, Koon-Kiu (Yale University)
York, Will (University of Georgia)
Zengler, Karsten (University of California, San Diego)

Observers

Bayer, Paul (DOE BER)

Bownas, Jennifer (ORNL)

Christen, Kris (University of Tennessee)

Foust, Cheri (ORNL)

Graber, Joe (DOE BER)

Gregurick, Susan (DOE BER)

Haun, Holly (University of Tennessee)

Houghton, John (DOE BER)

Katz, Arthur (DOE BER)

Mansfield, Betty (ORNL)

Nagahara, Larry (National Cancer Institute)

Ronning, Cathy (DOE BER)

Schexnayder, Susan (University of Tenn., Knoxville)

Weatherwax, Sharlene (DOE BER)

Yousef, Shireen (DOE BER)

Acronyms

ANL	Argonne National Laboratory
BESC	BioEnergy Science Center
BNL	Brookhaven National Laboratory
DOE	U.S. Department of Energy
BER	Biological and Environmental Research
GLBRC	Great Lakes Bioenergy Research Center
JBEI	Joint BioEnergy Institute
JGI	Joint Genome Institute
LBNL	Lawrence Berkeley National Laboratory
LLNL	Lawrence Livermore National Laboratory
NIH	National Institutes of Health
NREL	National Renewable Energy Laboratory
ORNL	Oak Ridge National Laboratory
PNNL	Pacific Northwest National Laboratory
USDA	U.S. Department of Agriculture

Appendix 3: Scientific Objectives Template

Scientific Objective: <title – Note: 1 Objective per each filled in template>

Breakout Group: <group>

Contributing Authors: <authors>

Date: <date>

1. Scientific Objective

Brief statement of Scientific objective

[What is the scientific or research goal? What is written here will usually be derived after filling in the remainder of the template. Responses to sections below will help to refine this statement. Sometimes it is easier to think of an objective in terms of a problem that exists that needs to be solved.]

Background information

[Include ongoing experiments, future planned experiments, historical results, literature references, relevant past impediments to research progress, etc.]

2. Prioritization

[This is meant to help prioritize this scientific objective in the context of other scientific objectives. There are several axes of consideration. One is the need or benefit to the research community. Another is the level of difficulty or feasibility.]

PRIORITY (check one): ___ HIGH ___ MEDIUM ___ LOW

Potential Benefits

[Why is this objective important? What is its level of impact? What would the benefit be? Who would benefit?]

Feasibility of success Near, Mid and Long term

[What is the level of difficulty? How likely is it that this objective can be achieved in a 1-3 year time frame? What would be the measure of success? Consider and rate feasibility in the near term (1-3 years, midterm (3-5 years) and long term (5-10 years). The most important objectives, those that are high priority and most feasible in the near term must have the most detail. Mid and long term can be provided in decreasing levels of detail.]

TERM (check one): _ NEAR (1-3 years) ___ MID (3-5 years) ___ LONG (5-10 years)

Relevance to DOE systems biology knowledgebase project

[The DOE Genomic Science program's ultimate goal of achieving a predictive understanding of biological systems is a daunting challenge and will require the integration of immense amounts of diverse information. The DOE Systems Biology Knowledgebase is envisioned as an open cyber-infrastructure to integrate systems biology data, analytical software, and computational modeling tools that will be freely available to the scientific community. Briefly explain how the proposed objective is relevant to what is envisioned for Kbase.]

Synergies/Leverage: Potential overlap with other projects or funding agencies

[Are there existing systems that relate to this objective such as NCBI, GenBank, BioCyc, iPlant, etc. Is there a potential for synergy that would benefit both efforts? Is there a potential overlatp that needs to be resolved?]

3. Specificity

[This section pertains to finding the right level of objective, especially avoiding objectives that are specified at too high a level. Start with a high level objective and refine. What is the specific science question to be answered?]

4. Details

[The intent here is to begin to capture elements that form the basis for continuity between the science objective and the software requirements that are derived from this objective. We start to articulate high level requirements here that are further refined in the requirements document.]

Scientific discovery process (workflows)

[Have workflows already been developed or can they be derived from existing work?]

Inputs

[What datasets would be required? Are there data standards? Are there available data sources or examples? Are there publications that use or describe an associated analysis process?]

Outputs

[What would the results be?]

Tools

[Existing or new analysis software]

REFERENCES

[Use as needed]

APPENDICES

[Use as needed]

- **Figures**
- **Tables**

EXAMPLE Scientific Objective: Improve Prediction of Microbial Gene Regulatory Networks

Breakout Group: Microbial

Date: May 12, 2010

5. Scientific Objective

Brief statement of Scientific objective

[What is the scientific or research goal? What is written here will usually be derived after filling in the remainder of the template. Responses to sections below will help to refine this statement. Sometimes it is easier to think of an objective in terms of a problem that exists that needs to be solved.]

Informative Example: Next generation sequencing technology will provide high quality RNA-Seq data at low cost. This presents an opportunity to substantially improve the quality of predicted gene regulatory networks compared with what has been possible with expression microarrays. This data together with transcription factor binding site predictions or determinations will provide the necessary data to built genetic regulatory networks for microbial genomes. High quality genetic regulatory networks of experimentally tractable organisms would increase the efficiency of experimental designs and genetic engineering. In the long term, having a collection of transcript profiles collected in a high quality, standardized manner across DOE relevant organisms such that genetic regulatory networks could be automatically determined in the context of the Kbase would provide an extremely valuable resource to advance microbial research.

Background information

[Include ongoing experiments, future planned experiments, historical results, literature references, relevant past impediments to research progress, etc.]

Informative Example: Next generation sequencing technology provides high quality RNA-Seq data at low cost. When acquired in sufficient quantity RNA-Seq data has dramatically better dynamic range and sensitivity than gene expression arrays and will probably replace them in 3-5 years. Transcriptome data can be used to define operons including transcription initiation and termination sites. Cluster analysis over multiple conditions will identify co-regulated operons and therefore defines co-regulated promoters. This data together with transcription factor binding site predictions or determinations will provide the necessary data to built genetic regulatory networks for microbial genomes.

6. Prioritization

[This is meant to help prioritize this scientific objective in the context of other scientific objectives. There are several axes of consideration. One is the need or benefit to the research community. Another is the level of difficulty or feasibility.]

Informative Example: Since genetic regulatory networks will facilitate efficient genetic engineering and other experimental designs (Cho et al., 2009), the priority of this objective should be high. (This statement of priority can be written at the workshop based on discussion.)

PRIORITY (check one): ___ HIGH ___ MEDIUM ___ LOW

Potential Benefits

[Why is this objective important? What is its level of impact? What would the benefit be? Who would benefit?]

Informative Example: Genetic regulatory networks of experimentally tractable organisms would increase the efficiency of experimental designs and genetic engineering. Microbes will be increasingly more important in manipulating a variety of organic molecules for biofuels, alternative plastics, other biochemical feedstocks, carbon sequestration and environmental remediation. Having the ability to efficiently manipulate and engineer these organisms will be absolutely crucial for cost effective design and large scale production of useful biochemicals.

Feasibility of success Near, Mid and Long term

[What is the level of difficulty? How likely is it that this objective can be achieved in a 1-3 year time frame? What would be the measure of success? Consider and rate feasibility in the near term (1-3 years, midterm (3-5 years) and long term (5-10 years). The most important objectives, those that are high priority and most feasible in the near term must have the most detail. Mid and long term can be provided in decreasing levels of detail.]

TERM (check one): _X_ NEAR (1-3 years) ___ MID (3-5 years) ___ LONG (5-10 years)

Informative Example: Collecting RNA-Seq data is already feasible and will only become more cost effective as third generation sequencing technologies are available in the next year. The community is already engaged in the development of analytical tools capable of integrating genomic DNA sequence and RNA-Seq data. The methods for predicting operons and their structure, cluster analysis of transcriptomic data to predict co-regulation of operons, predicting transcription factor binding sites and regulatory elements are already available but need to be streamlined and integrated. Implementing these kinds of analytical capabilities within the Kbase would be feasible in the first 1-2 years. RNA-Seq data is expected

to be widely available in the midterm 3-5 years and will need to be standardized to avoid some of the problems seen with GEO. Producing a functional genetic regulatory network for one or more bacterial organisms important to the Bioenergy centers appears to be achievable in 2-3 years if sufficient resources are applied.

Relevance to DOE systems biology knowledgebase project

[The DOE Genomic Science program's ultimate goal of achieving a predictive understanding of biological systems is a daunting challenge and will require the integration of immense amounts of diverse information. The DOE Systems Biology Knowledgebase is envisioned as an open cyber-infrastructure to integrate systems biology data, analytical software, and computational modeling tools that will be freely available to the scientific community. Briefly explain how the proposed objective is relevant to what is envisioned for Kbase.]

Informative Example: Predicting genetic regulatory networks requires integration of standardized sets of data and associated analysis methods along with the ability to test and improve the methods as envisioned for the Kbase. In the long term, having a collection of transcript profiles collected in a high quality, standardized manner across DOE relevant organisms such that genetic regulatory networks could be automatically determined in the context of the Kbase would provide an extremely valuable resource to advance microbial research.

Synergies/Leverage: Potential overlap with other projects or funding agencies

[Are there existing systems that relate to this objective such as NCBI, GenBank, BioCyc, iPlant, etc. Is there a potential for synergy that would benefit both efforts? Is there a potential overlatp that needs to be resolved?]

Informative Example: Generating the necessary RNA-Seq data would leverage JGI's production sequencing capabilities and could be synchronized with the genomic sequencing, while developing the analysis pipeline could be accomplished by ORNL's annotation group and incorporated into its' annotation pipeline. Individual PIs and smaller projects already pursue such analysis based on microarray data and the decreasing cost of RNA-Seq will eventually make RNA-Seq transcriptomics routine. Having a community of data integrated based on standards will provide a powerful resource. A natural byproduct will be better gene models and operon structures. This information will augment what is available in GenBank. An ancillary objective would be to update the associated annotation in GenBank.

7. Specificity

[This section pertains to finding the right level of objective, especially avoiding objectives that are specified at too high a level. Start with a high level objective and refine. What is the specific science question to be answered?]

Informative Example: Integration of 'omics data especially in complex systems such as plant microbe interfaces is an ambitious challenge that is too high level for the purposes of establishing version 1 of the Kbase, and not feasible in the near term (1-3 years) although it would be appropriate for the long term with a suitable scientific focus. However this high level aim could be made more tractable by simplifying in several ways. First, focus in on a simpler system such as a specific microbe. Second, instead of integrating all 'omics, take just two types of 'omics data, say genomic and transcriptomic as in this example.

In this example we started by considering 'omics integration and the large number of possible scientific objectives might derive from that such as a substantial model of major subsystems of a cell which would clearly be overly ambitious. By considering various combinations of 'omics data the level can be refined. In this example we recognized that by integrating just two kinds of 'omics data, genomics and transcriptomic using RNA-Seq, we would be able to have a science objective of improved prediction of gene regulatory networks that would be something tractable to accomplish in the relative near term within the Kbase and something useful to the microbial research community.

8. Details

[The intent here is to begin to capture elements that form the basis for continuity between the science objective and the software requirements that are derived from this objective. We start to articulate high level requirements here that are further refined in the requirements document.]

Scientific discovery process (workflows)

[Have workflows already been developed or can they be derived from existing work?]

Informative Example: Genetic regulatory networks have been created for *E. coli* (Cho et al., 2009) and *Halobacteria salinarum* NRC-1 (Bonneau et al., 2007). These papers describe workflows.

Inputs

[What datasets would be required? Are there data standards? Are there available data sources or examples? Are there publications that use or describe an associated analysis process?]

Informative Example: For a particular microbe of interest it would be expected that a finished genome sequence is available and for a few phylogenetically related organisms. In addition it would be expected that RNA-Seq of multiple growth states would have been obtained to a high level of coverage.

Outputs

[What would the results be?]

Informative Example: The results would be genetic regulatory network predictions for all microbes studied.

Tools

[Existing or new analysis software]

Informative Example: Numerous independent tools that have been developed. It will be necessary to develop analytical pipelines based on agreed workflows that integrate the RNA-Seq data, genomic sequence data, gene expression array data (if available), transcription factor binding site predictions and experimental verification (if available) in order to generate the genetic regulatory network predictions for a particular microbe.

REFERENCES

[Use as needed]

Bonneau, R., Facciotti, M.T., Reiss, D.J., Schmid, A.K., Pan, M., Kaur, A., Thorsson, V., Shannon, P., Johnson, M.H., Bare, J.C.*, et al.* (2007). A predictive model for transcriptional control of physiology in a free living cell. Cell *131*, 1354-1365.

Cho, B.K., Zengler, K., Qiu, Y., Park, Y.S., Knight, E.M., Barrett, C.L., Gao, Y., and Palsson, B.O. (2009). The transcription unit architecture of the Escherichia coli genome. Nat Biotechnol *27*, 1043-1049.

APPENDICES

[Use as needed]

- **Figures**
- **Tables**

Appendix 4: Requirements Template

Software Requirements for Scientific Objective: *Improve Prediction of Microbial Gene Regulatory Networks*

Breakout Group: Microbial

Reference Scientific Objective Number in Group: ____

Date: May 18, 2010

1 Scientific objective

[Describe the scientific objective that this software system requirements document supports. Description can be derived from the Scientific Objective template]

2 Resulting Requirements

[In the following sections list the requirements resulting directly from the identified scientific objective. Provide information for each requirement stating whether there are technologies available today to fulfill all or part of it that you are aware of, or if you expect that new development would be required. All requirements should indicate whether they are near, medium or long term requirements. The following Impact Factor is your group's assessment of the impact that addressing these requirements would have toward improving research productivity.]

IMPACT FACTOR (check one): ___ HIGH ___ MEDIUM ___ LOW

2.1 Process of the science (incl. workflow)

[Describe the process by which scientists use or want to use the data, software, and instruments for knowledge discover such as a scientific workflow. Identify both required and optional components. Indicate the state of the art of the different parts of the workflow.]

2.2 Instruments to support the achievement of the science objectives.

[List or describe instruments that generate relevant data connected to the scientific workflow above.]

2.3 User interfaces

[Describe generally who the users will be, and the user interfaces that play a role in achieving the scientific objective in the context of the workflows outlined above. Not all user interfaces will be directly involved in a workflow, and these if they exist should be captured here as well.]

2.4 Programmatic interfaces

[Describe the interfaces that will provide programmatic access to data or functionality that allow for automated data access, analyses and workflows in the context of the workflows outlined above. Not all programmatic interfaces will be directly involved in a workflow, and if these exist, capture them here as well.]

2.5 Data

[Describe the data and data types required to meet the scientific objectives. Include publicly available data, reference data, and new experimentally derived data. Discuss how the data is obtained such as is the data to reside locally on Kbase or would it exist remotely, outside of Kbase. Data representations including semantic web technologies or references can be included here, as well as references to existing data standards or relational tables. If known, include computer hardware resource requirements – such as the size of the data collection, and type and size of compute resources (processors, memory, temporary storage) required to manage and process the data.]

2.6 Software

[Describe which software algorithms, services and packages will be needed, if they exist or not, to achieve the scientific objective, and what computer hardware or other resources and data these would utilize.]

Software purpose	Availability	Does it need improvement	Resource impact

2.7 Standards

[Specify requirements that are derived from existing standards and/or regulations. While we don't expect much in the form of regulation, we should list those existing standards that we will use and areas where new standards need to be developed.]

2.8 Governance

[Related governance issues (usage policy, data policy, overall governance structure, community engagement for usage and development) should be described here. Some governance issues map to components of the system and these mappings should be called out in the System Architecture. How can governance help the implementation of standards?]

2.9 Summary and prioritization of requirements

[Summarize and prioritize your requirements, which ones are essential and which one are nice to have or could wait. Which requirements are near term, midterm and long term?]

3 System Architecture Attributes

[The common attributes are performance, reliability, availability, security, portability, interoperability, and usability (usually speaks to the importance of user interfaces with which humans interact as compared to a fully automated system that users just depend on). Important attributes from the list above should be discussed in the context of the scientific objective. For example, does achieving the science objective require a system that runs 24/7 with a yearly downtime of less than 8 minutes (this reflects the system's availability attribute). Will it perform calculations that require thousands of cores in order to complete in a reasonable time. Prioritize the relative importance of each architecture attribute and provide explanations of why, for example, why would security be more or less or equal in importance to performance.]

4 Kbase Key Services

[Optional – do this if able: Provide a list and description of the major functions/services that the Kbase system will need to provide to meet the scientific objective(s). This could include a mapping of existing functions onto existing systems such as MicrobesOnline, IMG, etc., or new services such as a central resource for temporary storage of data from different sources to be jointly analyzed. Here we can get into the finer details of what the system will do in order to meet the scientific objectives. Each function should be called out as a sub heading in this section]

4.1.1
4.1.2
4.1.3

5 Risk Analysis and Mitigation strategies

[Compile the list of potential risks in meeting the requirements of the scientific objective.]

-
-
-

6 Acryonyms, definitions and abbreviations

7 References

EXAMPLE Software Requirements for Scientific Objective: *Improve Prediction of Microbial Gene Regulatory Networks*

Breakout Group: Microbial

Reference Scientific Objective Number in Group: _____

Date: May 18, 2010

8 Scientific objective

[Describe the scientific objective that this software system requirements document supports. Description can be derived from the Scientific Objective template]

Improve Prediction of Microbial Gene Regulatory Networks

Next generation sequencing technology will provide high quality RNA-Seq data at low cost. This presents an opportunity to substantially improve the quality of predicted gene regulatory networks compared with what has been possible with expression microarrays. This data together with transcription factor binding site predictions or determinations will provide the necessary data to built genetic regulatory networks for microbial genomes. High quality genetic regulatory networks of experimentally tractable organisms would increase the efficiency of experimental designs and genetic engineering. In the long term, having a collection of transcript profiles collected in a high quality, standardized manner across DOE relevant organisms such that genetic regulatory networks could be automatically determined in the context of the Kbase would provide an extremely valuable resource to advance microbial research.

When acquired in sufficient quantity RNA-Seq data has dramatically better dynamic range and sensitivity than gene expression arrays and will probably replace them in 3-5 years. Transcriptome data can be used to define operons including transcription initiation and termination sites. Cluster analysis over multiple conditions will identify co-regulated operons and therefore defines co-regulated promoters. This data together with transcription factor binding site predictions or determinations will provide the necessary data to built genetic regulatory networks for microbial genomes.

9 Resulting Requirements

[In the following sections list the requirements resulting directly from the identified scientific objective. Provide information for each requirement stating whether there are technologies available today to fulfill all or part of it that you are aware of, or if you expect that new development would be required. All requirements should indicate whether they are near, medium or long term requirements. The following Impact Factor is your group's assessment of the impact that addressing these requirements would have toward improving research productivity.]

IMPACT FACTOR (check one): ___ HIGH ___ MEDIUM ___ LOW

9.1 Process of the science (incl. workflow)

[Describe the process by which scientists use or want to use the data, software, and instruments for knowledge discover such as a scientific workflow. Identify both required and optional components. Indicate the state of the art of the different parts of the workflow.]

(NOTE: This is an example that has been intentionally simplified and therefore extensions such as validation with computational or experimental methods such as 5' RACE to identify additional transcription initiation sites, or transcription factor regulatory ligand determinations have been removed. Others are welcome to take this as a starting point and expand for a specific "real" Scientific Objective.)

Taken from Scientific Objective section 4.2 Inputs: For a particular microbe of interest it would be expected that a finished genome sequence is available and for a few phylogenetically related organisms. In addition it would be expected that RNA-Seq of multiple growth states would have been obtained to a high level of coverage.

For the organism of interest it is assumed that the genome has been completely sequenced, fully annotated, and that RNA-Seq data is available for a minimum of 10 growth curves with 6 time points and 3 biological replicates on biological conditions relevant to the functional network(s) of interest.

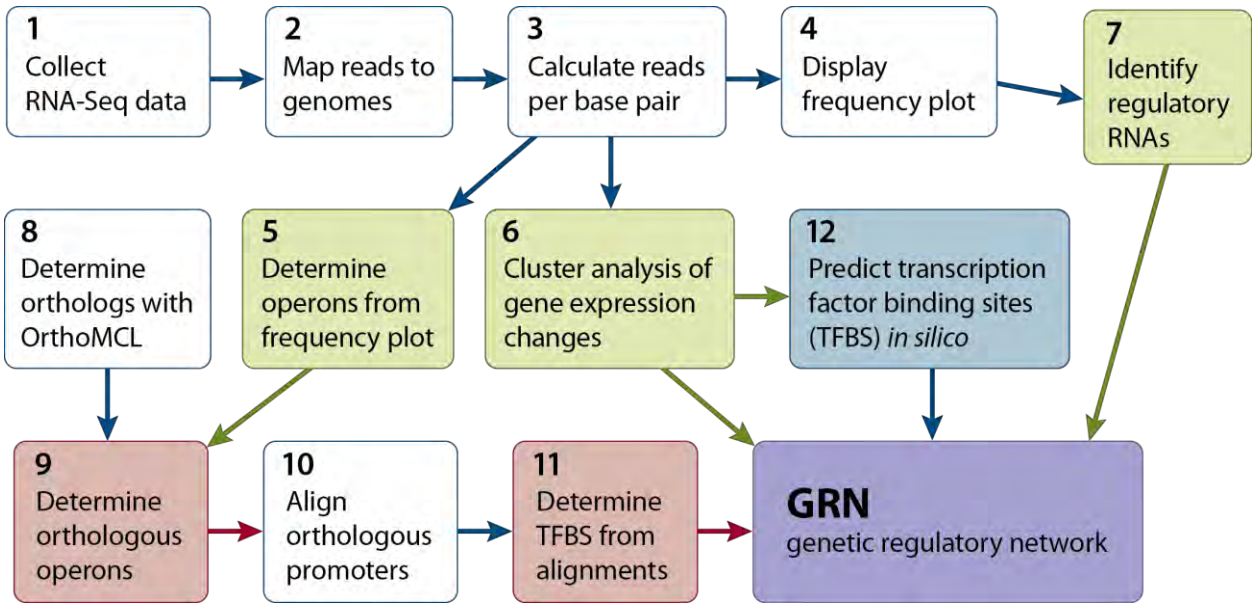

Figure 1. Transcriptome Analysis Pipeline for Gene Regulatory Network Prediction. White boxes are procedures we already know how to do. Green boxes are procedures that have not been determined but expected to be fairly easy to construct (year 1). Red boxes are procedures that will be more difficult to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

1. Collect RNA-Seq data and the accompanying metadata for each growth curve. The metadata could include optical density, substrate consumption, metabolites, temperature, and stirring condition. Although some of this data could be manually collected the Kbase would need to have the ability to store it in conjunction with the RNA-Seq as an experimental project.
2. Map RNA-Seq data to genome.
3. Calculate reads/bp (normalize and calculate expression levels of each gene and/or operon).
4. Display frequency plot for visual inspection and rule development for algorithms to identify the operons and regulatory RNAs in steps 5 and 7.
5. Determine operons from mapped reads (generate a separate list for each growth curve). These should include all the genes, the transcription initiation sites (TIS) and terminations sites with accuracy of a few bp for each operon.
6. Perform cluster analysis on the calculated gene expression levels to determine co-regulated operons.
7. Identify regulatory RNAs (unknown riboswitches and small regulatory RNAs) based on analysis derived rules identified in step 4 with expert guidance.
8. Determine orthologs from multiple related genomes using OrthoMCL or some other software tool.

9. Determine orthologous promoters from multiple related genomes.
10. Align orthologous promoters using Muscle or ClustalW.
11. Determine Sigma Factor and other Transcription Factor Binding Sites (TFBS) from alignments.
12. Use in silico TFBS prediction tools together with co-regulated operons to predict additional TFBS (import known TFBS from a database such as RegTransBase).
13. Predict Genetic Regulatory Network.

Further work after the initial implementation (years 1-2) would include evaluation of additional technologies and experimental verification to improve the process (5' RACE to identify additional TISs, microfluidic TFBS determinations and transcription factor regulatory ligand determinations). As the quality of the gene regulatory network predictions improves and the models are validated the workflow will be increasingly automated (years 3-5).

9.2 Instruments to support the achievement of the science objectives.

[List or describe instruments that generate relevant data connected to the scientific workflow above.]

Kbase should support RNA-Seq data from Solexa, ABI Solid, and 454. For the future there may be additional machines that will need to be supported such as PacBio. These instruments produce data of particular types and sizes that will need to be stored and/or managed within the context of the Kbase and are further described in the Data section below.

There will be potential for use of automated or multi-well instruments for generating growth curve data. Metadata such as optical density may be recorded manually or in spreadsheets output from instruments and Kbase will need to have capabilities for manual input or upload of such electronic data that would then be integrated within the experimental project.

9.3 User interfaces

[Describe generally who the users will be, and the user interfaces that play a role in achieving the scientific objective in the context of the workflows outlined above. Not all user interfaces will be directly involved in a workflow, and these if they exist should be captured here as well.]

The anticipated users will include biologists who wish to analyze their data, bioinformaticists who want to analyze data, contribute or improve methods and use existing methods, and scientists requiring information and visual representations for scientific publications. It is

anticipated that users will come from the academic, government and industry communities. It is not anticipated that there will be users at a level below the university level.

Interfaces will be needed for specifying an experimental project and locating the relevant RNA-Seq and associated experimental metadata. Users will expect to have a login space where they describe their experiment that they will be able to save and return to at a later time.

Scientific data visualization is needed that renders genome annotation, gene expression information, operons, alternative transcriptional starts, and multiple sequence alignments. User interfaces for visualizing frequency plots that show depth of coverage (relative expression levels) for genes and operons will be needed. Additionally, an interface that allows users to visualize the resulting gene regulatory network model will be needed.

9.4 Programmatic interfaces

[Describe the interfaces that will provide programmatic access to data or functionality that allow for automated data access, analyses and workflows in the context of the workflows outlined above. Not all programmatic interfaces will be directly involved in a workflow, and if these exist, capture them here as well.]

Kbase will need to have programmatic interfaces to support specific queries such as to return a list of all experimental conditions that an organism has been exposed to for which there is gene expression data.

Also, software that determines expression levels, or predicts or refines operon predictions will need access to genome annotation. Therefore data interfaces to NCBI SRA (Sequence Read Archive), GEO (Gene Expression Onmibus) and GenBank (bacterial genomes) will be needed or perhaps application interfaces to IMG, RAST or JGI-ORNL annotation systems. Will also need to import known TFBS from a database such as RegTransBase

The results of the workflow to predict gene regulatory networks will also produce data that would be output to data interfaces such as to all of the systems mentioned above in order to supply new data, support publication submission or update annotation.

9.5 Data

[Describe the data and data types required to meet the scientific objectives. Include publicly available data, reference data, and new experimentally derived data. Discuss how the data is obtained such as is the data to reside locally on Kbase or would it exist remotely, outside of Kbase. Data representations including semantic web technologies or references can be included here, as well as references to existing data standards or relational tables. If known, include computer hardware resource requirements – such as the size of the data collection, and type and size of compute resources (processors, memory, temporary storage) required to manage and process the data.]

In the near term, we expect to see for a given experiment several hundred files from short read sequencing technology. These files, if based on Solexa technology, will range in size from 100Mbytes to 100GBytes for the next couple of years. Current size ceiling is at about 4GBytes compressed for one run. Total data storage required is based on coverage and number of replicates, conditions and time steps, and therefore would be a multiplicative factor of 4GB (180X minimum as proposed). For the first 1-3 years it is expected that there would be 30-100 datasets per year (each dataset corresponding to studies on one microbe), and then grow to 100-300 per year in the 3-5year time frame when this data will be coming from many laboratories.

Database and storage resources – terabyte to petabytes range storage are needed. Data reduction will play a role in keeping storage resources manageable. Online backup capabilities needed for disaster recovery and long term archival.

Data types that cover high-throughput technologies to interrogate the transcriptome, are required for this scientific objective.

Genome sequences and a full complement of annotation features are also required. The data representation model as characterized by a GenBank record is probably not sufficient. New data models that capture gene annotations and their relationships to other annotations will be required. Annotation can exist remotely as in the case of taxonomy information housed in the NCBI taxonomy database and other NCBI Entrez data for which stable access exists through NCBI web services.

The gene regulatory network from a data structure perspective is the collection of operons, transcription factor binding sites, sigma factor binding sites; and those parameters that affect kenetics. These would benefit from representation that is based in semantic web technology.

It is expected that relational database technology will play a limited role in so far as perhaps providing structured storage of ontologies and RDF tuples.

9.6 Software

[Describe which software algorithms, services and packages will be needed, if they exist or not, to achieve the scientific objective, and what computer hardware or other resources and data these would utilize.]

Software for performing transcriptome analysis will be needed as part of the workflow and for visualization. It will integrate existing available genome annotation and provide measures of confidence. Annotation quality will be accessed based on confidence. A specific module will focus directly on improved identification of transcription factors.

Improved annotation with confidence and evidence codes will be sent back to repositories if possible.

Clustering software will be needed to group genes and operons into clusters based on patterns of regulation. Whether a part of the clustering software or part of a different package, it is anticipated that software which focuses on the fine details of the operon such as alternative transcriptional starts and stops will be needed.

Clustering algorithms will be compute intensive. Other methods are manageable with mid-range servers.

Data visualization software that spans genome annotation, transcriptome analysis and clustering will also be needed.

Software purpose	Availability	Does it need improvement	Resource impact
Maps rna-Seq data to genome	Few	Probably not	Storage
Cluster analysis of gene expression changes	Many	Probably	Compute, Storage
Operon determination	Few	Yes	
In silico TFBS prediction	Many	Yes	Compute
Ortholog determination	Few	Probably not	
Orthologous operon determination	None		
Promoter alignment	Few	Yes	
Promoter prediction	Few	Yes	
Gene regulatory network prediction	Few	Yes	Compute, Storage

Table 1: Types of software required for this scientific objective. Column-Resource impact: Compute means requires significant processor resource (>100 cores), and Storage means requires significant storage resource (>1 TB).

9.7 Standards

[Specify requirements that are derived from existing standards and/or regulations. While we don't expect much in the form of regulation, we should list those existing standards that we will use and areas where new standards need to be developed.]

-Gene regulation ontology (GRO) for terms related to gene expression

-Gene ontology (GO) for terms related to biological processes, cellular location and gene function

-NCBI sequence read archive xml schemas for sequence read metadata

-GCDML xml schema for genome metadata

-MIAME regards gene expression arrays but may be relevant to RNA-Seq.

9.8 Governance

[Related governance issues (usage policy, data policy, overall governance structure, community engagement for usage and development) should be described here. Some governance issues map to components of the system and these mappings should be called out in the System Architecture. How can governance help the implementation of standards?]

A data release policy will need to be in place. This would most likely be the current DOE policy and it is assumed that the Kbase will enforce this. This implies a private login which maps to System Architecture.

9.9 Summary and prioritization of requirements

[Summarize and prioritize your requirements, which ones are essential and which one are nice to have or could wait. Which requirements are near term, midterm and long term?]

In silico prediction of TBFS can be postponed until other elements of the workflow are complete (midterm). Support for microarray data was considered but has not been included for the sake of simplicity. If it would part of the requirements it might be lower priority because we believe it is phasing out. Other requirements for possible inclusion would be various kinds of validation such as 5' RACE and TFBS verification (midterm).

10 System Architecture Attributes

[The common attributes are performance, reliability, availability, security, portability, interoperability, and usability (usually speaks to the importance of user interfaces with which humans interact as compared to a fully automated system that users just depend on). Important attributes from the list above should be discussed in the context of the scientific objective. For example, does achieving the science objective require a system that runs 24/7 with a yearly downtime of less than 8 minutes (this reflects the system's availability attribute). Will it perform calculations that require thousands of cores in order to complete in a

reasonable time. Prioritize the relative importance of each architecture attribute and provide explanations of why, for example, why would security be more or less or equal in importance to performance.]

Users will be expecting that the data they submit will be secure in accordance with the governance model. This would be the highest priority.

It is anticipated that there could be some performance issues resulting from the choice of clustering algorithms and the amount of input data. Performance and security are architecture issues considered of highest importance for this objective.

11 Kbase Key Services

[Optional – do this if able: Provide a list and description of the major functions/services that the Kbase system will need to provide to meet the scientific objective(s). This could include a mapping of existing functions onto existing systems such as MicrobesOnline, IMG, etc., or new services such as a central resource for temporary storage of data from different sources to be jointly analyzed. Here we can get into the finer details of what the system will do in order to meet the scientific objectives. Each function should be called out as a sub heading in this section]

11.1.1 Mapping RNA sequence reads to a genome
11.1.2 Identifying operons
11.1.3 Identifying alternative transcription starts and stops
11.1.4 Identifying transcription factor binding sites
11.1.5 Improvements to genome annotation based on services 4.1.1 – 4.1.4
11.1.6 Data structures for representing gene regulatory networks
11.1.7 Query services for retrieving gene regulatory network models
11.1.8 Query services for retrieving all experimental conditions that an organism has been exposed to for which there is gene expression data

12 Risk Analysis and Mitigation strategies

[Compile the list of potential risks in meeting the requirements of the scientific objective.]

- Unanticipated changes in technology (sequencing, microarray) that would significantly change the requirements or implementation plan. Mitigated by anticipating changes and adjusting requirements and implementation plan as soon as possible.
- Inadequate data or poor data quality that precludes a productive workflow as currently designed. Mitigate by testing typical datasets for adequacy and quality. Modify experimental protocol to correct and change minimum standards.
- Cluster analysis on these datasets requires more resources than currently anticipated. Mitigate by modifying algorithm accept some additional error in return for performance speed. Allow clustering on subsets to manually find the optimum with reduced error.

13 Acryonyms, definitions and abbreviations

14 References