What is the purpose of this website?
This interactive wiki is a forum for the research community to help define requirements for the U.S. Department of Energy (DOE) Systems Biology Knowledgebase (KBase). Users are encouraged to click on the "How to Participate" link to join the conversation, post comments, and help develop the Knowledgebase. Click on the "Conceptual Design Report" drafting page to see the current status and contribute.

What is the Systems Biology Knowledgebase (KBase)?
The DOE Office of Biological and Environmental Research (BER) is supporting the development of the Systems Biology Knowledgebase, a cyberinfrastructure to facilitate a new level of scientific inquiry by serving as a central component for the integration of modeling, simulation, experimentation, and bioinformatic approaches. The Knowledgebase would focus on DOE science-application areas and meeting the needs of the DOE Genomic Science program, yet it also would be widely and easily applicable to all systems biology research. In addition to supporting data storage, retrieval, and management capabilities, KBase also would enable new knowledge acquisition and management, through free and open access to data, analysis tools, and information for the scientific research community. KBase, therefore, must serve multiple roles, including:

- A 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.

See the Project Plan for more information about the Knowledgebase and the principles guiding its development.

How will the Knowledgebase be developed?
This project will document the requirements necessary for the subsequent development of the Systems Biology Knowledgebase. A series of workshops, pilot projects, and activities exploring options for hardware infrastructure will be used to determine the scope, cost, and schedule for the Knowledgebase.

1. Workshops. To develop a successful open informatics endeavor for DOE systems biology, this project will identify all user groups (e.g., plant and microbial genomic researchers, bioinformaticians, computer scientists, database developers, software engineers, etc.) and elicit their expectations in regards to the opportunities and requirements for developing and managing KBase. The workshops will also address the cultural transition the informatics community will need to make from individual project based efforts toward research community based informatics.

2. Pilot Projects. In order to explore some possible Systems Biology Knowledgebase implementations and architectures, as well as demonstrate and characterize the range of computational challenges facing Genomic Science projects, part of this project will support the development of pilot projects to address three goals:

   ○ Develop benchmarks for existing computational biology and bioinformatics programs on existing architectures,
   ○ Develop prototypic computational biology and bioinformatics programs on new architectures including cloud architectures, and
   ○ Develop novel and integrative web platforms as possible solutions to bioinformatics problems in anticipation of and to inform a future Systems Biology Knowledgebase.

3. Hardware Infrastructure. Recognizing that there are a range of biocomputing problems and that different computing architectures may be required, an objective of this effort will be to provide access to a range of computing architectures that can be accomplished with available or off-the-shelf products.

What will be the final product of this project?
The final product of this project is the Final Implementation Plan that will document the scope, cost, and schedule of the complete KBase. This report will be completed and posted on September 30, 2010.

We want Your participation
Lastly, your input to this wiki site (i.e., your comments and suggestions) is needed in order for this project to succeed. If you are interested in helping to shape the future Knowledgebase, please join in. One of the guiding principles for Knowledgebase is open development – meaning anyone can contribute. So please sign up, start contributing, and become part of the first online wiki content systems biology community of its kind.
Acknowledgement

The Knowledgebase R&D project is sponsored by the Office of Biological and Environmental Research in the DOE Office of Science with American Recovery and Reinvestment Act 2009 funding and performed at Oak Ridge National Laboratory (ORNL). ORNL is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.
Open development is one of the founding principles of the DOE Systems Biology Knowledgebase. Anyone can contribute. Sign up below to join the community and participate in drafting the conceptual design report specifying the requirements for the Knowledgebase.

Please click on this link and provide your name, organization, phone number, and e-mail address. You will be contacted when you have been granted editorial privileges.

Become part of the **first** online wiki content systems biology community of its kind.
Project Description

This project will create the conceptual design and implementation planning necessary for subsequent development of the Systems Biology Knowledgebase (KBase). A fully functional Systems Biology Knowledgebase is envisioned to be a cyberinfrastructure for systems biology information and data that not only includes data storage, retrieval, and management, but also enables new knowledge acquisition and management, through free and open access to data, analysis tools, and information for the scientific research community. Knowledgebase capabilities are envisioned to include:

- Curation, not just data, but models and representations of scientific concepts
- Analysis including the ability to compare methods and inventory of results
- Simulation including the ability to modify and improve models
- Prediction based on simulation and analysis to form new hypotheses
- Experimental design and comparison between prediction and results

The Knowledgebase would drive two classes of work: (1) experimental design and (2) modeling and simulation. Integrating data derived from computational predictions and modeling, as envisioned in the Knowledgebase project, would increase data completeness, fidelitly, and accuracy. These advancements in turn would greatly improve modeling and simulation, leading to new experimentation, analyses, and mechanistic insight. Scientists' ever-increasing exploitation of the dynamic linkages among data integration, experimentation, and modeling and simulation - aided by KBase - will advance efforts to achieve a predictive understanding of the functions of biological systems.

The Knowledgebase will transform the way the biology community views data and information by providing a contextual and interactive environment to drive and transform technology revolutions and ensure the Department of Energy (DOE) meets mission critical challenges in energy and the environment. Specifically, the Knowledgebase will accelerate research by integrating together data and information on microbial and plant systems to enable a clearer understanding of the processes involved in bioenergy production and environmental stabilization.
This project will define the scope, cost, schedule, and infrastructural needs for KBase by gathering community input from a series of workshops and supporting pilot projects. A partnership between the DOE Office of Biological and Environmental Research and the DOE Office of Advanced Scientific Computing Research will support activities focused on KBase computational infrastructure.

**Purpose**

As biological research advances and enables the pursuit of more ambitious objectives, projects are becoming larger and more complex, encompassing more diverse and specialized capabilities. This describes the situation both experimentally and computationally. In order for these projects to be successful, there is a need for increasing cooperation and standardization. In addition, advances in technological capabilities associated with large-scale biological research projects are producing exponentially more data. All of these trends are leading toward requiring a new kind of computational infrastructure in order for the overall scientific effort to be successful.

Ultimately, to do large-scale science in the future, it will be necessary to have an associated large-scale, open-community computational capability for data management and analysis. The vision for an integrated experimental framework for accessing, comparing, analyzing, modeling, and testing systems biology data was described in the DOE report, Systems Biology Knowledgebase for a New Era in Biology.

**Background**

Historically, individual research groups were largely independent and funded as such. Not surprisingly this resulted in associated computational systems being independently developed and largely disconnected. Because of a lack of standards, both computational and experimental, these systems are not readily integrated today. As research using new technologies becomes more productive and collaborative, it is necessary for the computational systems supporting this research to reflect this change. This is the transition from individual research projects to large-scale community efforts – in many ways a cultural change.

As described in a recent article in *Science* (Gordon Bell et al., 6 March 2009 323: 1297-1298), 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.

The Knowledgebase will be a substantial software engineering effort unlike anything done before in this community. As such, it demands a clear description of the stakeholders, what they require, and what they would define as success. Furthermore, success for this project will be as much about scientific accomplishment as technological achievement.

**Objectives**
The Knowledgebase project objectives are as follows:

- Engage the computer science research community in the challenges of biocomputing for the future.
- Transition the BER informatics community from individual project based efforts toward research community based informatics.
- Provide pilot examples of software.
- Provide infrastructure examples of hardware and software.
- Develop strategies for a successful open informatics science endeavor.
- Establish scope, specifications, and requirements for Knowledgebase implementation.

**Guiding Principles**

The guiding principles for this project include:

- Open contribution— Meaning data and methods will be available for anyone to use.
- Open source – Source code will be freely available to access, modify, and redistribute under the same terms such as the GNU General Public License.
- Open development – Meaning anyone can contribute to the development following the organization guidelines. This would be analogous to submitting a publication. A review process by an authoritative group would determine if the contribution meets established criteria.
  - Allow for user accounts such that data and code can be held private and analysis can be conducted in a private environment.
  - Allow for tracking history, versions, and provenance so that analysis done today can be usefully compared with analysis done previously.

**Plan for Success**

In order to be successful, the Knowledgebase will focus on scientific goals and strong community involvement while emphasizing the cultural change from individual to community science. This plan for success will include significant efforts toward:

- Assessing quality of experimental data
- Establishing experimental protocol and standards
- Tracking provenance of data, including analytical processes

**Final Report - The Final Implementation Plan**

The DOE Systems Biology Knowledgebase project is a significant software engineering and operations effort. Successfully building such a system depends on sufficiently detailed specifications and requirements. The Final Implementation Plan will be the conceptual design document establishing the scope, cost, and schedule of the Knowledgebase effort.
Final Implementation Plan

The DOE Systems Biology Knowledgebase project is a significant software engineering and development effort to provide a computational environment for researchers to contribute data and analysis methods to model dynamic cellular systems of plants and microbes. Successfully building such a system depends on sufficiently detailed specifications and requirements. The Final Implementation Plan document is the conceptual design and initial plan and is attached below. Download the entire document, or download by section.

Download document:
Final Implementation Plan

Download document by section:
Executive Summary
Body
Appendix
Pilot Projects

The goal of the pilot project is to identify computational problems and solutions in the context of the Knowledgebase that inform design approaches for the final report. Five pilot projects were selected and are listed below along with a summary of their findings.

**Argonne National Laboratory, Folker Meyer, Principal Investigator**

*An Experience Report: Porting the Existing MG-RAST Multi-User Web Application to the Cloud*, Jared Wilkening, Andreas Wilke, Elizabeth M. Glass, Narayan L. Desai, and Folker Meyer

**Lawrence Berkeley National Laboratory, Adam Arkin, Principal Investigator**

*Design Requirements and Prototypes of Workflows in the SBKB for Support of Engineering of Metabolic Pathways*, Dylan Chivian, John Bates, Paramvir Dehal, Marcin Joachimiak, Morgan Price, Vinay Satish Kumar, and Adam Arkin

**Lawrence Berkeley National Laboratory, Victor Markowitz, Principal Investigator**

*Database Management Systems Technologies for Computational Biology & Bioinformatics Applications*, Victor Markowitz

**Pacific Northwest National Laboratory, Ian Gorton, Principal Investigator**
Exploring Architecture Options for Workflows in a Federated, Cloud-based Systems Biology Knowledgebase, Ian Gorton, Yan Liu, Jian Yin, Leeann McCue, Bill Cannon, and Gordon Anderson

Pacific Northwest National Laboratory, Kerstin Kleese van Dam, Principal Investigator

Final Evaluation Report for the Semantic Driven Knowledge Discovery and Integration in the Systems Biology Knowledgebase Project, Kerstin Kleese van Dam, Cliff Joslyn, Lee Ann McCue, Bill Cannon, Carina Lansing, Zoe Guillen, Margaret Romine, Gordon Anderson, and Abigail Corrigan
Porting the existing MG-RAST multi-user web application to the cloud

Knowledgebase R&D Pilot Project

Jared Wilkening, Andreas Wilke, Elizabeth M. Glass, Narayan L. Desai, Folker Meyer

Argonne National Laboratory

The Metagenomics RAST server (MG-RAST) allows public upload and analysis of data via its web portal and has been enabling over 2000 data submitting users since 2007. To continue providing this community resource to groups in over 30 countries the team supporting the server had to identify a means of scaling up analytical capacity. The task seemed daunting as for the analysis of 1 GB of metagenomic data, MG-RAST requires approx. 2,000 CPU hours on a recent Intel Nehalem machine with moderate memory (8), the computational cost dominated by computing sequence similarity searches. For this project we investigated computations in the MG-RAST computational pipeline to determine their suitability to the cloud computational paradigm.

The MG-RAST system provides a web portal and a computational pipeline for the analysis of metagenomics datasets. Users submit sequence data sets via the MG-RAST portal, a web interface implemented as a series of Perl cgi scripts. This data is loaded into a Postgres database, along with information about which analytics should be performed and metadata about the dataset. This data is used, in turn, to submit jobs into the computational pipeline, AWE, which manages work execution.

Due to a combination of increasing demand for analysis via MG-RAST, as well as the quickly increasing data set sizes, it has become clear that dedicated computing resources will not provide sufficient capacity for even the short term, with considerably larger shortfalls in the medium term. Our first approach to tackling this problem was to scale out the MG-RAST backend to existing shared resources at Argonne, followed by an approach using new cloud resources. This change posed some technical challenges, due to differences in the general infrastructure provided on shared systems, as well as the switch to using distributed resources.

The similarity analysis stage of the MG-RAST system was our initial target. This stage is implemented using NCBI BLAST. It is a good candidate for distribution, both because of its large overall resource consumption, and the computation uses a large fixed database that changes infrequently. The only input to this stage is a small query sequence, easily transmitted from AWE. AWE consists of a centralized set of python daemons that can communicate with clients via a RESTful interface. This approach is widely portable, as clients need only be able to perform HTTP requests to the server. The work queue, as well as results and statistics, are stored in a Postgres database. We have used Facebook’s Tornado framework to build a lightweight and efficient set of front tier web servers.

After becoming accustomed to dealing with performance issues, we were able to retune our database, modify our work management reliability scheme, and scale a single backend up to 500 compute nodes without issue. Further optimizations can be applied when they are needed; the architecture is designed to scale among multiple backend machines. However, since our current approach has been sufficiently scalable for the compute resources we have access to, we have not pursued this issue.
The use of microbial hosts for the production of chemicals has been one of the great success stories of biotechnology. We envision a key use of the systems biology knowledgebase for the engineering of new metabolic function in microbes for the production of fuels, high value chemicals, and environmental activities such as bioremediation. The fundamental workflow we envision has a user selecting a host organism based on qualities important for the industrial process envisioned or on the ability to make molecules related to those desired by the user. The user must then determine if the organism is able to utilize the feedstock molecules and transform them into internal metabolites from which production of the target chemicals becomes possible. Finally, the user needs to determine the optimal modification of the host microbe to efficiently create the target molecule without severe detriment to host health either by the unbalancing of metabolism through over-draw of essential metabolites, unbalancing of co-factors/energy molecules, or by producing intermediates toxic to the host organism.

Determining the range of chemical transformations accessible to a natural microorganism requires accurate, preferably evidence-based, assignment of function to genes for that organism. Metabolic reconstruction still suffers from a number of challenges surrounding accurate annotation of proteins. These same problems plague retrosynthesis where it is essential that the genes from diverse organisms have as accurate and specific functional assignments as possible so that the algorithms can efficiently find the needed chemistries. Lastly, we identified a need for and began addressing the development of interfaces for navigating metabolic networks and experimental functional *omics data using a “Google-Like” Application for Metabolic Maps (GLAMM).

Retrosynthesis pathways and genes will be linked back to MicrobesOnline to permit further examination with its powerful comparative systems biology tools, a prototype for the SBKB, including phylogenetic gene trees, genome context and operon predictions, functional residue alignments, protein-protein interaction data, and basic structural models to permit developing a mutually consistent set of genes for introducing the viable candidate retrosynthetic pathways into the host microorganism.
The aim of this project was to examine new database management system technologies for supporting efficient analysis of very large genome and metagenome sequence datasets.

Comparative analysis of genomic and metagenomic datasets is usually based on integrating these datasets in the context of databases implemented using relational commercial database management systems (DBMS) such as Oracle or open source DBMS such as MySQL. The rapid increase in the number and size of these datasets results in a decrease in performance of typical comparative analysis tools, such as examining putative operons across microbial genomes. A recent benchmark of relational DBMS\(^1\) indicates that new database management technologies are better suited for scientific data management applications. We set out to evaluate the usage of cloud based data management technologies for handling large genome and metagenome datasets, in particular Hadoop data management components for data storage and querying. Hbase\(^2\) is a distributed, column-oriented data store that supports real-time access to extremely large data.

Cloud based data management technologies can be potentially very useful for a wide variety of genome and metagenome data management applications. We used as a case study the Integrated Microbial Genomes (IMG) system. IMG currently stores the results of "all vs. all" pairwise gene comparisons in sequence similarity files. These files are tab-delimited files generated by NCBI's blastall program with the \(-m8\) option, containing the identifiers of pairs of matching genes, scores pertaining the strength of match such as alignment percentage identity, regions of matches, bit score, and an evaluation of statistical significance through the expectation value (E-value). Storing the results in flat files has several disadvantages in comparison to tabular storage. Modifying individual entries is challenging, and queries are significantly harder then would be the case in tabular storage.

Our testing of HBase shows that distributed tabular storage has significant long term potential for the GTL Knowledgebase, but that current HBase versions are not ready for large-scale production use today. Issues with both stability and performance will need to be addressed before HBase can be used in a production Knowledgebase application.

We encountered significant difficulties in running HBase in a stable fashion. We encountered frequent crashes and performance problems while attempting to bulk load data. Some of these problems were surely caused by our inexperience in running Hadoop/Hbase in a production environment, but others are likely the result of the relative immaturity of the software. Both Hadoop and HBase are undergoing rapid development currently and we anticipate that many of these stability problems will be addressed over the next year or two.

\(^2\) Hbase: http://hadoop.apache.org/hbase/
Exploring Architecture Options for Workflows in a Federated, Cloud-based Systems Biology Knowledgebase

Ian Gorton, Yan Liu, Jian Yin, Leeann McCue, Bill Cannon, Gordon Anderson
Pacific Northwest National Laboratory
Richland, WA

Systems biology is characterized by a large community of scientists who use a wide variety of fragmented and competing data sets and computational tools of all scales to support their research. In order to provide a more coherent computational environment for systems biology, we are working as part of the Department of Energy Systems Biology Knowledgebase (Kbase) project to define a federated cloud-based system architecture. The Kbase will eventually host massive amounts of biological data, provide high performance and scalable computational resources, and support a large user community with tools and services to enable them to utilize the Kbase resources. We investigated the design of a workflow infrastructure suitable for use in the Kbase. The approach utilizes standards-based workflow description and open source integration technologies, and incorporates a data aware workflow execution layer for exploiting data locality in the federated architecture.

An overview of our use case for the Kbase depicting data sets and computations is shown in the Figure. A biologist retrieves data relevant to the organism under study from GenBank (a public database) and this is input into a script to translate the data into a format needed for the next step in the workflow. An Hadoop-based computation, Polygraph, is then invoked with the translated GenBank data and proteomics data. This produces a list or outputs (peptides) that can be fed into a desktop-based visualization tool that takes inputs data from the analysis and produces the spectrum of the genome where both published annotations and orphan peptides can be discovered.

We have implemented this workflow utilizing a federated cloud-based architecture consistent with the KBase architecture. In this prototype, we explored the necessary software architectures and technologies to enable data-location driven workflow for the Kbase. We coordinate the workflow by means of routing the workflow tasks to distributed Web services on a Cloud using REST APIs. The actual computation to be executed is determined by MeDICi pipelines that are driven by the data location and computation demands. In our experience, separating the workflow definition from its coordination is crucial to enable the extensible integration of the workflow with federated resources that are distributed across Cloud-based platforms.

The jBPM workflow definition tool was used as it allows customized workflow coordination to be easily deployed to its workflow engine, and interacts with REST APIs. Other scientific workflow tools such as Taverna and myExperiment are also candidate technologies for use as workflow tools in the UAL of the KBase architecture. Taverna provides a workbench to launch workflows published and shared on the myExperiment site. Taverna is efficient in accessing Web services defined with the WSDL and SOAP protocol. For generic REST APIs using HTTP protocol, some wrappers using Taverna’s specific language need to be developed for Taverna to invoke REST APIs.

In our current work, we are investigating the necessary mechanisms and software tools required for integrating the infrastructure layer with UAL with an open specification. We are also extending our adaptive MeDICi prototype framework so that the abstract workflow definitions created by the users can be dynamically mapped to the underlying federated Cloud resources in a data-driven manner.
It is the goal of the DOE Systems Biology Knowledgebase to become a community driven infrastructure for sharing and integration of Data and Analysis tools. This new infrastructure should enable the science community to move towards a new era in Biology, where it is possible to gain a predictive understanding of biological systems to enable them to address core DOE Missions and societal needs. Hereby the community wide accessibility of biological data and the capability to integrate and analyze this data within its environmental context are seen as key technical functionalities the Bio-Knowledgebase has to enable. The ultimate success of the Bio-Knowledgebase will however not only rely on its technical ability to meet the communities fast changing information needs, but even more so on its ability to motivate the community to actively participate in its development. This project has demonstrated over the past 8 months that semantic technologies are not only mature enough to be considered, but will indeed be essential in achieving the goals of the Systems Biology Knowledgebase. It established that semantic technologies have the capability to:

- Deliver key technical functionalities in flexible data access and integration. Concept based (semantic) enterprise search and access services, federated and integrated across multiple heterogeneous life systems biology sources and ontology mapping helps to extend search across the boundaries of molecular biology. Ability to include data sources beyond the direct realm of systems biology such as those required for capturing environmental conditions.

- Support of flexible integration of data, analysis and workflows for experts and non-expert users of the Knowledgebase. Integration of data using concept mappings between different ontologies. User driven grouping of data, analysis and workflows into useful units, allowing easy reuse, sharing, change and annotation.

- Provide easy integration of existing DOE resources for data, application and workflows. Easy ‘LinkedData’ approach – using the web to publish and link resources that were not linked before - to publishing data, application and workflows in combination with biology centric semantic description via RDF. A simple application provides a new generic interface to existing resources, without any required changes to the underlying databases or data registries.

- Allow leveraging of community knowledge through utilization of the many existing domain ontologies via ontology mapping (through tools such as SOBOM or LOOM with up to 95% accuracy). Exploitation of significant overlaps in concepts between different ontologies as demonstrated on the NCBI BioPortal ontologies (Ghazvinian, 2009), allowing for easier integrated access to community resources.

- Ensure quick start up of the Knowledgebase and short term gains for its user community through integration and leveraging of existing core data collections and community developments.

The ‘Semantics Driven Knowledge Discovery and Integration in the Systems Biology Knowledgebase’ project has carried out extensive user requirement gathering through multiple avenues. The project members actively participated in the DOE Knowledgebase workshops, helping to define key scientific goals and resulting technical requirements for the meta-genomics, plant and microbe communities. The project considered the results of work of the European Life sciences Infrastructure for Biological Information project, and worked closely with the DOE funded Foundational Scientific Focus Area of Biological Systems Interactions (FSFA) at PNNL as well as PNNL Proteomics facilities.
The results of the requirement gathering were evaluated and used to define the design of suitable test scenarios for semantic services such as annotation, publication, search, access and integration for the Biology Knowledgebase to meet the community's scientific needs. In addition to scientific needs around direct data services, an equally strong requirement for collaborative user environments was established in which projects can share their knowledge internally as well as externally (publically) through interactions facilitated with the Biology Knowledgebase.

Based on findings of the requirement analysis the project developed a prototype test environment including:

- A collaborative, project-centric user environment which provides access to a range of semantic technologies regarding knowledge sharing, annotation, publication, search and access.
- A prototype data services infrastructure to support the user environment functionality in the Knowledgebase context and integrated its capabilities with other core Knowledgebase functionalities.

The suitability of the environment and the semantic technologies were assessed in terms of functionality and maturity for any Knowledgebase implementation.

The research and tests demonstrated that many semantic technologies had reached a sufficient level of maturity over the past few years and are indeed already used in production level environments for both research and commercial systems biology environments (e.g. GoPubMed Healthmash, Linked life data, Data.gov). Semantic technologies would allow the Knowledgebase to offer key enabling data services to its user community, on which many of the other higher level services such as analysis and workflows will strongly depend. Furthermore, semantic technologies would offer distinct advantages over other solutions in terms of their functionality, flexibility and adaptability to leverage existing resources and speed of deployment. In addition, the following observations were made during the development of the prototype test environment:

- A collaborative user environment that is ontology-driven (i.e., APIs written to a common vocabulary) is more extensible and can be implemented much more quickly. Without a common vocabulary, development of a user interface is error prone and extremely time consuming.
- A collaborative user environment that is ‘plugin’ based supports component reuse throughout the community. In addition, it allows users to customize their working environment to best meet their personal scientific needs.
- Developing a common vocabulary was critical to the success of the pilot project. Similarly, for the Knowledgebase to be successful, the biology community must engage and commit to developing a common vocabulary and mapping their data sources to it.
- Converting existing data to a common vocabulary was the most time intensive piece of the test environment, indicating that this will likely comprise a large portion of the initial Knowledgebase development.

Our test users have been impressed across the board by the functionality and ease of use of the components of the prototype test environment:

‘I think it is fantastic that we have access to these resources. I have desperately needed a way to share large datasets with collaborators and this development is turning out to be a great solution.’ Margaret Romine, PNNL

‘I found it to be a great way to collaborate within the project, access and share data.’ Margrethe (Gretta) Hauge Serres, Josephine Bay Paul Center, Marine Biological Laboratory

The projects felt that the infrastructure filled a key technology gap for their collaborative work. Subsequently the collaborative user environment has been adopted by a range of multi-site projects and is now actively in use. More projects are waiting to adopt the tools in the near future. All of them are looking forward to be able to utilize the expanded capabilities of the infrastructure once the Knowledgebase is available and are planning to actively participate in its usage and development.

The requirements evaluation and prioritization as well as extensive user tests of key components of the prototype influenced the overall Knowledgebase data infrastructure architecture design. The design and its integration with the overall architecture are documented in the DOE Bio Knowledgebase Implementation Plan. The project is closing its prototype development with an integrated system consisting of a collaborative user environment, semantic services and analysis workflows executed in commercial (Amazon) and non-commercial cloud (Magellan) environments (in collaboration with the ‘Architectures and Technologies for Knowledgebase Workflows’ project lead by Ian Gorton, PNNL).
The Systems Biology KBase pilot hardware named Kandinsky is a machine designed for optimal support of the Hadoop architecture/runtime. The original recommendation made by the science advisory board focused on examining the MapReduce programming paradigm and its applicability to bioinformatics applications. As a result, the hardware configuration includes over 0.5 petabytes of storage on local nodes under the direction of the Hadoop Distributed File System.

In addition to supporting Hadoop based applications, support for private cloud virtualization will be added via the Eucalyptus infrastructure software that enables establishment of private cloud computing environments. Eucalyptus is interface-compatible with the Amazon Web Services (AWS) cloud infrastructure, which means users can reuse existing AWS-compatible tools and scripts to manage their own private cloud, run Amazon Machine Images on their private cloud and cloud-burst to other public-clouds (also known as hybrid clouds --
a private on-premise cloud, in this case, a Eucalyptus cloud, working seamlessly with a public cloud).

General Information

- **Hardware specification.**

- **Software Stack:** Kandinsky is a 64 bit Linux system with CentOS Kernel version 5. The Hadoop distribution we use is Cloudera’s distribution of Hadoop version 3(CDH3), which comes with the following tool set:
  - HDFS.
  - HBase.
  - MapReduce.
  - Hive, Pig, Oozie, Sqoop, Flume, Hue and Zookeeper.

User Documentation

- **Getting Started:** First time users need to request a new account. Once the account is available the user can log into the system and access the local as well as Hadoop distributed file system (HDFS) which is part of the Hadoop cloud. To run any Hadoop based application the user will need to transfer their data from the local file system to distributed file system. The other popular data storage scheme on the Hadoop cloud is HBase which can be accessed on Kandinsky via a shell.

- **Available data:** We have the Sequence Read archive (SRA) data available HDFS and HBase. The compressed sequence files are available in HDFS whereas the metadata is available in a huge HBase single master table called SRADa ta.

- **Using Installed applications:** We currently have the following applications installed and ready to be used:
  - **Crossbow:** Crossbow is scalable, portable, and automatic Cloud Computing tool for finding SNPs in genomes from short read data.
  - **CloudBurst:** CloudBurst is a new parallel read-mapping algorithm optimized for mapping next-generation sequence data to the human genome and other reference genomes, for use in a variety of biological analyses including SNP discovery, genotyping, and personal genomics.

Developer Documentation

- **Code for creating your own private distributed SRA:** We have developed code libraries based on HDFS API and HBase API which allow individual centers to download their own copy of SRA
data and distribute it over their private Hadoop cloud. The software is available on Bitbucket and can be checked out via SRA_Hbase. We would be happy to discuss any development ideas regarding the code base.

**Related Resources**

- Apache Hadoop Project
- Hadoop Map Reduce
- Hadoop HBase.
- Cloudera CDH3.
- Eucalyptus.
Oak Ridge National Laboratory, with support from the DOE Office of Biological and Environmental Research, is organizing a series of highly focused, community-building workshops to address scientific and software-design requirements for the DOE Systems Biology Knowledgebase (KBase). The first four workshops will be held in conjunction with existing scientific meetings. Output from these workshops will contribute to the production of a design document for implementing the Knowledgebase.

Objectives for the Workshops

- Identify all user groups (e.g., plant and microbial genomic researchers, bioinformaticians, computer scientists, database developers, software engineers, etc.) and elicit their expectations in regards to the opportunities and requirements for developing and managing KBase.
- Document community input for the effective development of data capabilities for systems biology that could be applied specifically to plants and microbes (i.e., bacteria, archaea, fungi, and protists - unicellular eukaryotes such as microalgae) important to the fundamental science underlying DOE missions.
- Recognize the cultural transition the informatics community will need to make from individual project based efforts toward research community based informatics.
- Develop strategies for a successful open informatics science endeavor.

Workshop Summary Report

After each workshop, a Summary Report will be written. The Summary Report will document discussions and will be incorporated into the Conceptual Design Report.

<table>
<thead>
<tr>
<th>Workshop</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Plant and Animal Genome: Joint USDA-DOE Plant Genomics</td>
<td>Jan. 8, 2010</td>
</tr>
<tr>
<td>Knowledgebase Workshop</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>--</td>
</tr>
<tr>
<td><strong>3. DOE Genomic Science: Microbial Systems Biology Knowledgebase Workshop</strong></td>
<td>Feb. 9, 2010</td>
</tr>
<tr>
<td><strong>4. JGI User Meeting Knowledgebase Workshop</strong></td>
<td>March 23, 2010</td>
</tr>
<tr>
<td><strong>5. Knowledgebase Systems Development Workshop</strong></td>
<td>June 1, 2010</td>
</tr>
</tbody>
</table>

Supercomputing (Nov. 16, 2009) - kbase

Supercomputing 2009 (SC09) was in Portland, Oregon for its 21st annual conference. Recognized globally as the premier international conference on High Performance Computing (HPC), networking, storage and analysis, SC09 features interesting and innovative HPC scientific and technical applications from around the world.

Website: http://sc09.supercomputing.org/index.php

Using Clouds for Parallel Computations in Systems Biology

*Held at SC09 on Monday, November 16*

The aim of this workshop is to bring together researchers in the computing, systems biology, and computational biology fields. The workshop will focus particularly on applications of cloud computing. 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 GB of sequence data, or one-third of a genomics sequence space (current archives of complete genomic data comprise 51 GB). The situation is further complicated by the democratization of sequencing; many small centers can now independently create large sequence datasets. Moreover, the immense amount and variety of 'omics data that must be integrated together with genomics data in order 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 fall under the "embarrassingly parallel" header and should be ideally suited for cloud computing. However, challenging issues remain, including data transfer and local data availability on the cloud nodes. This workshop aims to bring together computer scientists, bioinformaticists, and computational biologists to discuss the feasibility of using cloud computing.

Website: http://www.mcs.anl.gov/events/workshops/sc09-sysbio/.
Workshop Organizers

1. Susan Gregurick, U.S. Department of Energy
2. Folker Meyer, Argonne National Laboratory
3. Bob Cottingham, Oak Ridge National Laboratory
4. Peg Folta, Lawrence Livermore National Laboratory
5. Elizabeth Glass, Argonne National Laboratory

Charge Questions
(To post, read, or comment on responses to these questions, see the Charge Questions subpage)

1. What are the characteristics of applications that would be appropriate for effective utilization of cloud architecture?
2. What are the hardware bottlenecks that prohibit cloud architectures from being easily adopted by high-throughput biological data analytics?
3. What are specific tools that need to be developed or enhanced in order to make cloud architectures easily adopted for biological data and bioinformatics algorithms?

Agenda and Location of Meeting Room
Click here to view the Agenda.
The workshop will be held in Room A104.

List of Abstracts
(To post, read, or comment on abstracts, see the Abstracts subpage)

1. Clouds in Systems Biochemistry, Christopher H. Chang
2. MapReduced Complete Composition Vectors for Genotyping, Marc Colosimo, Matthew Peterson, Scott Mardis, and Lynette Hirschman
3. A Distributed Terabyte-size Parallel Processing Systems Biology Platform Based on the Hadoop/MapReduce/HBase Framework, Ronald Taylor
5. Architectures for the Next Generation of Systems Biology, Thomas Brettin
6. Informatics, Thomas Brettin
   Commodity Computing in Genomics Research, Michael Schatz, Mihai Pop, Dan Sommer, and Ben Langmead
7. A Performance Comparison of Massively Parallel Sequence Matching Computations on Cloud Computing Platforms and HPC Clusters Using Hadoop, Shane Canon, Shreyas Cholia, John Shalf, Keith Jackson, Lavanya Ramakrishnan, and Victor Markowitz
8. Using MapReduce Technologies in Bioinformatics and Medical Informatics, Xiaohong Qiu, Jaliya Ekanayake, Thilina Gunarathe, Jong Youl Choi, Seung-Hee Bae, Scott Beason, and Geoffrey Fox
List of Presentations
(To read or comment on the presentations, see the Presentations subpage)

1. Architectures for the Next Generation of Systems Biology Informatics, Tom Brettin
2. AWE: Pipelines for Cloud Scale Bioinformatics, Narayan Desai, Folker Meyer, Jared Wilkening, James Yang, and Andreas Wilke
3. Challenges: How to Cope with an Explosion of Fascinating Data, Dawn Field
4. Cloud-Based Services for Large Scale Analysis of Sequence and Expression Data: Lessons from Cistrack, Robert Grossman
5. Cloud Computing with Nimbus, Kate Keahey
6. CloVR: A Virtual Appliance for Automated Analysis of Sequence Data, Sam Angiuoli
7. Commodity Computing in Genomics Research, Michael Schatz, Ben Langmead, Dan Sommer, and Mihai Pop
8. Future Directions for Cloud Computing in Systems and Computational Biology, Susan Gregurick and Bob Cottingham
9. Genomes in the Clouds: The UCSC Genomics Browsers and Distributed Biocomputation, David Haussler and Daniel Carlin
10. M5 and Multi OMICS, Eugene Kolker
11. Magellan at NERSC, Katherine Yelick
12. Magellan Cloud at ALCF, Pete Beckman
15. Towards a Consensus Annotation System, Owen White
17. Using MapReduce Technologies in Bioinformatics and Medical Informatics, Judy Qiu
18. ViPDAC, A Stand-Alone Proteomics Analysis Suite in the Cloud, Simon Twigger
19. Please click here to view The Role of Cloud Computing in Big Biology, Deepak Singh

Kbase Workshop Summary Report
Posted January 25, 2010
Joint USDA-DOE Plant Genomics Knowledgebase Workshop (Jan. 8, 2010)

The Plant and Animal Genome (PAG) Conference XVIII will be held January 9-13, 2010 in San Diego, California (website: http://www.intl-pag.org/). The Joint USDA-DOE Plant Genomics Knowledgebase Workshop is on January 8 from 7:30 a.m. to 6:00 p.m. This session will specifically address the requirements for effective development of data capabilities for systems biology as applied to plants, particularly relating to 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, it is important to integrate all types of data (including molecular, morphological, and “omics” data) for bioenergy-relevant plant species. Thus, the challenge will be to develop uniformity of data format and database architectures in order to effectively integrate diverse data types and enable user-friendly acquisition and analysis.

Workshop Organizers

1. Catherine Ronning, Department of Energy
2. Susan Gregurick, Department of Energy
3. Ed Kaleikau, U.S. Department of Agriculture
4. Gera Jochum, U.S. Department of Agriculture
5. Bob Cottingham, Oak Ridge National Laboratory

Charge Questions

**Question 1:** What types of experimental data are currently available, and of these which format(s) are most useful/valuable? Can data from various sources and of various types be standardized into this “ideal” format to be organized and integrated into one common, searchable application

- For example, a researcher studying cell wall biosynthesis in grasses may benefit from work being performed in poplar. How can we best facilitate cross-species comparisons? How
can we use these tools to leverage knowledge gained from the model species (*Arabidopsis*, rice, etc.) to crop plants?

**Question 2:** What are the challenges for plant bioinformatics in a 2-3 year time frame? Given the development of an integrated, uniform database (Question 1), what types of analyses do you foresee, and what types of analysis tools will maximize utility of the database?

- How do we best organize pathways, processes, etc., and how can we organize and distinguish common processes from taxon-specific processes? How can these informatics resources best be used to enhance plant breeding (i.e., “genotype to phenotype”)? Will these resources be effective in designing decision support tools for plant breeders in the field?

*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?*

**List of Responses to Charge Questions**

(Click here to read responses to the Charge Questions)

1. Response to Charge Questions, Robin Buel
2. Response to Charge Questions, David Douches
3. Response to Charge Questions, Andrew Doust
4. Response to Charge Questions, Jorge Dubcovsky
5. Response to Charge Questions, Maria Harrison
6. Response to Charge Questions, Thomas Lubberstedt
7. Response to Charge Questions, Seth C. Murray
8. Response to Charge Questions, David Neale
9. Response to Charge Questions, Zhaohua Peng
10. Response to Charge Questions, Steve Strauss
11. Response to Charge Questions, Janice Zale
12. Response to Charge Questions, Peijian Cao, Manfred Zorn, and Pamela Ronald
13. Response to Charge Questions, Jeffrey Ross-Ibarra
14. Response to Charge Questions, Dong Wang
15. Response to Charge Questions, E. Beers, A. Dickerman, and A. Brunner
16. Response to Charge Questions, Patrick Byrne
17. Response to Charge Questions, Tim Close
18. Response to Charge Questions, Sam Hazen
19. Response to Charge Questions, Eva Huala
20. Response to Charge Questions, Scott Jackson
21. Response to Charge Questions, Christian Tobias
22. Response to Charge Questions, Wilfred Vermerris
23. Response to Charge Questions, John Vogel
24. Response to Charge Questions, David Francis
25. Response to Charge Questions, Matias Kirst
26. Response to Charge Questions, Katrien M. Devos
27. Response to Charge Questions, Pam Green
28. Response to Charge Questions, Ismail Dweikat
29. Response to Charge Questions, Jan Leach
30. Response to Charge Questions, Luca Comai
31. Response to Charge Questions, Charles Brummer
32. Response to Charge Questions, Jerry Tuskan
33. Response to Charge Questions, Richard Veilleux
34. Response to Charge Questions, Steve Rounsley
35. Response to Charge Questions, Phil McClean
36. Response to Charge Questions, T. M. Davis
37. Response to Charge Questions, Victor Busov
38. Response to Charge Questions, Andy Pereira
39. Response to Charge Questions, Lukas Mueller
40. Response to Charge Questions, Jim Giovannoni

**Agenda and Location of Meeting Room**
Click here to view the Agenda. The Workshop will be held in the Pacific Salon 3 room.

**List of Abstracts**
(Click here to read Abstracts)

1. *Brachypodium distachyon*: A New Model for the Grasses, John Vogel
4. Advancing the Barley Genome, Timothy J. Close, Stefano Lonardi, Jeffrey L. Bennetzen, and Gary J. Muehlbauer
5. Genome Sequence for the Common Bean, Scott Jackson, Phil McClean, Jeremy Schmutz, and Dan Rokhsar
6. BeanCAP – A NIFA Coordinated Agricultural Project, Phillip E. McClean
7. Barley Coordinated Agricultural Project: Leveraging Genomics, Genetics and Breeding for Gene Discovery and Barley Improvement, Gary J. Muehlbauer
8. Bioinformatics Use in Advancing Plant Genomics, Genetics and Breeding: The Plant Breeders Perspective, David Francis

**List of Presentations**

1. AFRI Plant Genome, Genetics and Breeding Program, Ed Kaleikau

2. BeanCAP - A NIFA Coordinated Agricultural Project, Phil McClean

3. Genome Sequence for the Common Bean, Scott Jackson
4. Barley Coordinated Agricultural Project: Leveraging Genomics, Genetics and Breeding for Gene Discovery and Barley Improvement, Gary J. Muehlbauer

5. Advancing the Barley Genome, Tim Close et al


7. Brachypodium distachyon: A New Model for the Grasses, John Vogel

8. GRIN-Global: An International Project to Develop a Global Plant Genebank Information Management System, Peter Bretting

9. Introduction to the Plant Knowledgebase Workshop, Catherine Ronning
10. DOE Systems Biology Knowledgebase, Bob Cottingham
11. A Plant Breeding Perspective, David Francis
12. The iPlant Collaborative, Steve Rounsley
13. Leveraging Arabidopsis Data for Research on Other Plant Species, Eva Huala
14. US-EC Plant Bioinformatics, Doreen Ware
15. Genomes as an "Organizing Principle" for the Knowledgebase, Dan Rokhsar

Kbase Workshop Summary Report
Posted March 19, 2010
The DOE Genomic Science Contractor-Grantee and Knowledgebase Workshop will be held February 8-11, 2010 in Washington, D.C. Research projects supported by the DOE Genomic Science program are working towards achieving a predictive, systems-level understanding of plants, microbes, and biological communities, via integration of fundamental science and technology development, to enable biological solutions to DOE mission challenges, including energy, environment, and climate. The Genomic Science program objectives are:

1. Determine the genomic properties, molecular and regulatory mechanisms, and resulting functional potential of microbes, plants, and biological communities central to DOE missions.
2. Develop the experimental capabilities and enabling technologies needed to achieve a genome-based, dynamic systems-level understanding of organism and community functions.
3. Develop the knowledgebase, computational infrastructure, and modeling capabilities to advance the understanding, prediction, and manipulation of complex biological systems.

The DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop (Feb. 9, 2010)

**Workshop Report now available** - click here to download report

The DOE Genomic Science Contractor-Grantee and Knowledgebase Workshop will be held February 8-11, 2010 in Washington, D.C. Research projects supported by the DOE Genomic Science program are working towards achieving a predictive, systems-level understanding of plants, microbes, and biological communities, via integration of fundamental science and technology development, to enable biological solutions to DOE mission challenges, including energy, environment, and climate. The Genomic Science program objectives are:

1. Determine the genomic properties, molecular and regulatory mechanisms, and resulting functional potential of microbes, plants, and biological communities central to DOE missions.
2. Develop the experimental capabilities and enabling technologies needed to achieve a genome-based, dynamic systems-level understanding of organism and community functions.
3. Develop the knowledgebase, computational infrastructure, and modeling capabilities to advance the understanding, prediction, and manipulation of complex biological systems.

The DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop (Breakout Session D)

**Tuesday, February 9, 2:00-5:00pm**

This workshop will bring together researchers from the Genomic Sciences community in microbial systems biology, computational biology and bioinformatics. This workshop is open to all participants attending the DOE Genomic Science Contractor-Grantee Workshop, however, we ask that you pre-register at the meeting website: [http://www.orau.gov/gtl2010/](http://www.orau.gov/gtl2010/).

The goal of this workshop is to outline the near, mid and long
term trajectory of microbial sciences for energy and the environment. A second scientific objective is to map the associated workflows and data integration methods that can inform the specifications and requirements for the future development of the DOE Systems Biology Knowledgebase. A fully functional Systems Biology Knowledgebase is envisioned to be a cyber-infrastructure for systems biology information and data that not only includes data storage, retrieval and management, but also enables new knowledge acquisition and management, through free and open access to data, analysis tools, and information for the scientific research community.

**Important Topics Relevant To This Workshop**
- Near, mid and long term scientific goals of a Systems Biology Knowledgebase
- Development of workflows for systems biology
- Bridging data and integration

**Workshop Organizers**
Susan Gregurick, Department of Energy
Bob Cottingham, Oak Ridge National Laboratory

**Co-Chairs**
Adam Arkin, University of California
Bob Kelly, North Carolina State University

**Charge Questions and Responses**
We are soliciting responses to the charge questions listed below (not more than 2 pages). Please email your response to kbasewiki@ornl.gov by January 24th, 2010. Please use PDF, ASCII or DOC formats. Responses will be posted here.

1. For systems biology of interest to Genomic Sciences, what are scientific objectives that a knowledgebase could address in the 5 year timeframe and in the longer timeframe?
2. What are the key workflows that could be developed to accomplish these goals? Provide comprehensive usage examples that lead to scientific objectives.
3. What types of data are required to accomplish the objective?
4. What are bottlenecks to data integration and data usability that need to be addressed to accomplish these goals?
5. What are the bottlenecks in bioinformatic and computational algorithms that need to be addressed to accomplish these goals?
6. What would success look like? What would the benefit be?

**List of Charge Question Responses**
(Click here to read responses to the Charge Questions)
1. Response to Charge Questions, Louis Sherman
2. Response to Charge Questions, Kerstin Kleese Van Dam, et al.
3. Response to Charge Questions, Costas Maranas
4. Response to Charge Questions, Christopher Chang
5. Response to Charge Questions, Brian Davison
6. Response to Charge Questions, Jake McKinlay
7. Response to Charge Questions, Frank Collart
9. Response to Charge Questions, James Brainard
10. Response to Charge Questions, Michael Seibert
11. Response to Charge Questions, William Cannon
12. Response to Charge Questions, David Pletcher
13. Response to Charge Questions, Paul Blum
14. Response to Charge Questions, Ronan Fleming
15. Response to Charge Questions, Jeffrey Lewis
16. Response to Charge Questions, Dylan Chivian
17. Response to Charge Questions, David Reiss
18. Response to Charge Questions, Xiao-Jiang Feng
19. Response to Charge Questions, Tatiana Karpinets

**Agenda (Breakout Session D)**

**2:00 - 2:30 p.m.**

**Bob Cottingham, Microbial Systems Biology Knowledgebase: Scientific Objectives and Current Prospects**

*Focus on examples of scientific objectives, benefits and outcomes*

**2:30 - 3:00 p.m.**

Discussion

**3:00 - 3:30 p.m.**

**Bob Kelly, Near-Term Prospects for Functional Microbial Genomics: Moving Beyond the Monoculture Paradigm**

*One organism to two organism, adding complexity*
The DOE Joint Genome Institute's (JGI) Genomics of Energy & Environment 5th Annual User Meeting was held March 24-26, 2010 in Walnut Creek, California (website). The JGI invited scientists interested in the application of genomics to bioenergy and environmental issues, as well as all current and prospective users and collaborators, to attend the annual JGI Genomics of Energy & Environment meeting. This international gathering of researchers with an interest in sequence-based science offered three days of presentations, tours, workshops, and poster sessions. The emphasis of this meeting was on the genomics of renewable energy strategies, carbon cycling, environmental gene discovery, and engineering of fuel-producing organisms. The meeting featured presentations by leading scientists advancing these topics.

The Knowledgebase Workshop was held on Tuesday, March 23rd and is part of a series of meetings aimed at engaging the scientific community in discussing the scientific goals and development strategy for the Kbase. The workshop was scheduled in conjunction with JGI's Annual User Meeting, and thus will allow setting the Kbase discussions in the context of JGI’s genome and metagenome sequence data generation and analysis services. The main theme of this workshop was to discuss the Kbase as a system that would build on existing "omics" data management and analysis systems while achieving a higher level of support for the scientific community.

Research projects supported by the DOE Genomic Science program are working towards achieving a predictive, systems-level understanding of plants, microbes, and biological communities, via integration of fundamental science and technology development, to enable biological solutions to DOE mission challenges, including energy, environment, and climate. The Genomic Science program objectives are:

1. Determine the genomic properties, molecular and regulatory mechanisms, and resulting functional potential of microbes, plants, and biological communities central to DOE missions.
2. Develop the experimental capabilities and enabling technologies needed to achieve a genome-based, dynamic
systems-level understanding of organism and community functions.

3. Develop the knowledgebase, computational infrastructure, and modeling capabilities to advance the understanding, prediction, and manipulation of complex biological systems.

**Important Topics Relevant to this Workshop**

- Future directions for Metagenomics and implications for the DOE Systems Biology Knowledgebase
- Building toward an open community Knowledgebase to meet scientific objectives

**Registration for Workshop**

To register for this Workshop, please click this link and provide your name, organization, telephone number and e-mail address.

**Workshop Organizers**

Susan Gregurick, Department of Energy
Bob Cottingham, Oak Ridge National Laboratory

**Co-Chairs**

Victor Markowitz, JGI and Lawrence Berkeley National Lab
Jill Banfield, University of California, Berkeley

**Charge Questions**

We are soliciting responses to the charge questions listed below (not more than 2 pages). Please email your response by March 22, 2010. Please use PDF, ASCII or DOC formats. Responses will be posted.

1. What are key experimental and computational next steps that build on the sequencing data and information provided by JGI and are feasible for an initial Knowledgebase implementation associated with research in microbial communities?
2. What types of data and information are currently available or required to accomplish these objectives?
3. How these research goals are hindered by an inability to access and integrate data from various sources or of other types?
4. **What are the bottlenecks in bioinformatics and computational algorithms that need to be addressed to accomplish these goals? Specifically, is there a benefit to closer collaboration between sequencing analysis and the downstream analysis?**

**Response to Charge Questions**

1. Response to Charge Questions, Louis Sherman
2. Response to Charge Questions, Ernest Szeto
3. Response to Charge Questions, Mavrommatis Konstantinos
4. Response to Charge Questions, Robert Landick
5. Response to Charge Questions, Matthias Hess
6. Response to Charge Questions, Igor Brown
7. Response to Charge Questions, Janet Jansson
8. Response to Charge Questions, Patrik D’haeseleer
9. Response to Charge Questions, Steven Hallam
10. Response to Charge Questions, Kerstin Kleese van Dam

**Agenda**

Walnut Creek Marriott, California Ball Room A, B, and C

8:30 a.m. – 8:40 a.m.        Welcome, Susan Gregurick

8:40 a.m. – 9:00 a.m.        Introduction to Knowledgebase Initiative /Workshop Objectives   **Bob Cottingham**

9:00 a.m. – 9:30 a.m.        Science Presentation on Metagenomics: Current Experience and Future Expectations for the Kbase  **Phil Hugenholtz**

9:30 a.m. – 10:00 a.m.       Science Presentation on Metagenomics: Current Experience and Future Expectations for the Kbase  **Jill Banfield**

10:00 a.m. - 10:30 a.m.      Panel Summary/Audience Questions

10:30 a.m. – 11:00 a.m.      Break

11:00 a.m. – 12:30 p.m.      Informatics Perspectives and Roundtable: How to Transition from the Present Towards an Open, Shared, Integrated Kbase

Discussion Leads: Adam Arkin, Folker Meyer, Ed Uberbacher, Nikos Kyrpides, Peter Karp, Tatiana Tatusova, Victor Markowitz, Bob Cottingham

12:30 p.m. – 1:00 p.m.       Working Lunch

1:00 p.m. – 1:30 p.m.        Presentation of Metagenomic Workflow Example, Jill Banfield

1:30 p.m. – 3:00 p.m.        Panel Discussion, Jill Banfield

3:00 p.m. – 3:30 p.m.        Break

3:30 p.m. – 4:30 p.m.        Panel Discussion, Jill Banfield
4:30 p.m. – 5:00 p.m.                Conclusions and Adjourn, Bob Cottingham

List of Presentations

1. Metagenomics: Current Experience and Future Expectations for the Kbase, Jill Banfield
2. Introduction to the Knowledgebase Initiative, Bob Cottingham
3. Introduction to the Joint Genome Institute DOE Systems Biology Knowledgebase Workshop, Susan Gregurick

Picture from Workshop
Click on link above to view

Kbase Workshop Summary Report
Posted May 28, 2010. See attachment below to download.
The three-day **Knowledgebase Systems Development Workshop** will focus on discussion and planning to determine the systems biology software and architecture necessary for a fully functioning Knowledgebase and will document the design, cost, and schedule for achieving it.

**Workshop Goal and Agenda**

The goal of this workshop is to establish initial, actionable plans to create the Knowledgebase. Below is a list outlining the focuses of each day. To see the full agenda, click here.

<table>
<thead>
<tr>
<th>Date</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>June 1</td>
<td>Prioritization of clear scientific objectives and specific requirements or the Knowledgebase derived from these objectives</td>
</tr>
<tr>
<td>June 2</td>
<td>Development of an implementation plan, system architecture, and governance for the initial system</td>
</tr>
<tr>
<td>June 3</td>
<td>Finishing writing assignments leading to the report which will be the plan for creating the Knowledgebase</td>
</tr>
</tbody>
</table>

**Pre-Workshop Documents**

Click here to read information and supporting materials for the workshop, including background reports, contributed white papers, and abstracts.

*JUST POSTED* - the **Scientific Objective and Requirements templates and example documents and the Scientific**
Objectives Index.

Contact Information
Bob Cottingham, 865-241-0554 office; cottinghamrw@ornl.gov
Brian Davison, 865-574-0955 office, davisonbh@ornl.gov

Workshop Organizers
Susan Gregurick, Department of Energy
Bob Cottingham, Oak Ridge National Laboratory

Registration for Workshop
Workshop openings have all been filled. Registration has been closed.

Presentations
Click here to read the presentations from this Workshop.

Kbase Workshop Summary Report
Posted June 30, 2010. See attachment below to download.
## Draft Outline of Conceptual Design Report

1. **Purpose**

   1. **Vision**: KBase provides a computational environment for researchers to contribute data and analysis methods to model dynamic cellular systems at a high level of accuracy and, by extension, many of these systems within a cell and within a community of cells and organisms, all interacting with their environment. Ultimately, KBase will allow users to perturb a cellular system and observe the result.

   2. **Detailed scientific objectives and how KBase provides a solution**: What systems biology problems will KBase address? (from science workshops)

1. **Requirements**

   1. **Use cases**: Comprehensive usage examples to meet scientific objective. Scenarios or vignettes; provide scenarios that tie multiple use cases together (an end-to-end picture that meets science objective) (from science workshops)

   2. **Which scientific objective(s) and scenario(s) will be addressed first, and which will be addressed thereafter?**

   3. **What data would be contributed?**

   4. **What analysis methods exist or need to be developed?**

   5. **What kind of hardware will be needed?**

2. **System architectural details**

   1. **Architectural attributes**

      1. Integration
      2. Interoperability
3. Data description
4. Usability
5. Performance

2. Available tools to implement KBase
   1. Software
   2. Databases
   3. Hardware
   4. Relevant architectural approaches
   5. Successful open development methodologies
   6. Existing systems such as IMG, RAST, CAMERA
   7. Community de facto tools and databases available from organizations

3. Gaps in existing tools and hardware
4. Data management, modeling, and representation
5. Distributed computing model
6. Architectural strategies and decisions that affect the overall organization of KBase. These include things such as:

   1. Programming languages, database systems, libraries, etc. to be supported.
   2. Reuse of existing software components to implement various parts/features of the system (which tools listed in 3.1.1 and 3.1.2 above).
   3. Data and software synchronization across sites if a distributed model is embraced. External databases vs. local copies.
   4. User interface paradigms (web, thin clients, thick clients, others?).

   7. How the system is divided up into subsystems, what is unique to each subsystem, and which elements of the subsystems are shared.

8. How will subsequent scientific objectives and scenarios affect the KBase architecture?

3. Design (this is distinct from governance in that governance decides what, the design indicates how)
   1. How would data be contributed?
   2. How would new types of data be contributed?
   3. How would new analysis methods be contributed?
   4. How would KBase be modified and improved?

4. Implementation Plan [include elements of software project management (PM), software configuration management (CM), and software quality assurance (SQA)].
   1. Review specific scientific objective and scenarios will be addressed.
   2. Process by which KBase will be implemented (PM).
   3. Introduce development environment on pilot projects.
   4. How will it be operated?
   5. How will it be maintained?
   6. How will software be tested (testing strategy SQA)?
7. What will the change process look like, and how will the change control and bug tracking be implemented (CM)?
8. How will software be compiled, deployed, and released (deployment and release strategy CM)?

5. Governance Model – Governance in the enterprise software domain of KBase can be thought of as the development and enforcement of policies and procedures. Policies in this context can be thought of as design decisions combined with enforcement. Since a primary goal of a good architecture is to define a modular system and well-defined abstractions, choices that are made along the way in this regard need a level of enforcement. Governance starts with a vision of what the governance process will accomplish. This vision should be a collective effort of the people who will use, design, build, and pay for KBase. Most of all, tolerance must be the social norm in the KBase governance model.

1. Define a tolerant governance model.
2. Define a board of review that will be responsible for the development, maintenance, and change of policies.
3. Develop an interoperability framework that discusses standards and the details of these standards.
   1. An interoperability framework should list the standards that KBase will use, point to reference information, and indicate the status of the choice (such as approved, de facto, emerging, sustained, being phased out, being phased in, etc.)
   2. A standard that has a status of 'sustained' is a special case where although a newer standard or alternative standard has been chosen, the sustained standard will be supported by KBase.
4. Identify centers of excellence within the community that can support governance and interoperability standards.
5. Define different classes of participants (i.e., users, developers, committers, subproject managers) and their roles in the open KBase development process, in governance, in project management, etc.
6. Define a governance process that will produce policies, xml schemas, wsdl documents, ontologies, and other artifacts that must be distributed to the KBase community of users, developers, etc.
7. Make the products of governance searchable, versioned, and easily referenced (URI). And in many cases, make the products of governance machine-readable.
8. How and where will we create a registry of governance artifacts? If using Netbeans or Eclipse, does a plug-in exist that links the registry to the IDE?
9. What tools are available to automate as much of the governance process as possible?
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10. On the issue of enforcement or compliance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. Role of PM team to keep subprojects compliant</td>
</tr>
<tr>
<td></td>
<td>2. Role of executive team to keep project compliant</td>
</tr>
<tr>
<td></td>
<td>3. Link funding and compliance where possible</td>
</tr>
<tr>
<td></td>
<td>4. What are the corrective actions when a part of a subsystem or system is found to be non-compliant?</td>
</tr>
<tr>
<td></td>
<td>5. How are exceptions to policy handled and supported?</td>
</tr>
<tr>
<td>11. Transition from largely independent efforts to community working together.</td>
<td></td>
</tr>
<tr>
<td>13. Editorial organization and process.</td>
<td></td>
</tr>
<tr>
<td>14. Deciding which data and data types are to be contributed (proposed by users...).</td>
<td></td>
</tr>
<tr>
<td>15. Deciding which new analysis methods are to be contributed (proposed by users...).</td>
<td></td>
</tr>
<tr>
<td>16. Deciding which standards to endorse (proposed by users...).</td>
<td></td>
</tr>
<tr>
<td>17. How to keep the KBase user community engaged in the software development life cycle and invested in it so they feel some ownership of success. Who are the enthusiastic communities?</td>
<td></td>
</tr>
</tbody>
</table>

6. Appendices

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Workshop Report</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. Supercomputing</td>
</tr>
<tr>
<td></td>
<td>2. Plant and Animal Genome</td>
</tr>
<tr>
<td></td>
<td>3. DOE Genomic Science PI Meeting</td>
</tr>
<tr>
<td></td>
<td>4. JGI User Meeting</td>
</tr>
<tr>
<td></td>
<td>5. Knowledgebase Systems Development</td>
</tr>
<tr>
<td>2. Specific scientific objectives from the workshops or proposed by others. Rather than include multiple examples in the main document, each example will have its own section.</td>
<td></td>
</tr>
<tr>
<td>3. Technical software engineering documents</td>
<td></td>
</tr>
</tbody>
</table>

# Wiki for Drafting Requirements for the DOE Systems Biology Knowledgebase

*Community-driven cyberinfrastructure for sharing and integrating data and analytical tools*

## Related Resources

### Documents and Reports
- Systems Biology Knowledgebase: 4-page brochure
- Systems Biology Knowledgebase for a New Era in Biology: Report from the May 2008 Workshop
- DOE Genomic Science Program Overview
- Genomics:GTL 2008 Strategic Plan (Note: the DOE Genomic Science program was previously called Genomics:GTL program)
- Genomics:GTL Roadmap 2005 (See PDF of computing section)

### Websites
- U.S. Department of Energy Office of Biological & Environmental Research
- U.S. Department of Energy Genomic Science Program
- U.S. Department of Energy Genomic Science Program - Computing
- U.S. Department of Energy Advanced Scientific Computing Research

### Papers

### Press Releases
- DOE to Explore Scientific Cloud Computing at Argonne, Lawrence Berkeley National Laboratories
- A Deluge of Data Shapes a New Era in Computing By John Markoff (click on attachment below)
- Security in the Ether, David Talbot
- Is Our Data Too Vulnerable in the Cloud? Nick Bilton (click on attachment below)
### Contact Us

If you have any comments, suggestions or questions, we would like to hear from you. Click here and provide your name, organization, phone number and comments.

**Contact Bob Cottingham, cottinghamrw@ornl.gov; 865-241-0554**