# DOE Systems Biology Knowledgebase Workshop Report from the 5th Annual JGI User Meeting

## Tuesday, March 23, 2010, 8:30 a.m. – 5:00 p.m.

**Convened by**

**The U.S. Department of Energy Office of Science as part of the DOE Joint Genome Institute's (JGI) Genomics of Energy and Environment 5th Annual User Meeting, Walnut Creek, California**

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

**Workshop Cochairs:** Victor Markowitz (JGI, Lawrence Berkeley National Laboratory) and Jill Banfield (University of California–Berkeley)

### Table of Contents

## Introduction

This report covers discussion and material from the Department of Energy (DOE) Systems Biology Knowledgebase workshop held on March 23, 2010, prior to the 5th Annual DOE Joint Genome Institute (JGI) User meeting. The focus of this knowledgebase workshop was to discuss scientific objectives and challenges for data handling and knowledge integration specific to the study of microbial communities or metagenomes. The topics also included some discussions and items pertinent to all development and initial implementation of knowledgebases for the broader biological community.

A brief table of contents for this report is provided above. First, there is a background summary of the purpose of the DOE Systems Biology Knowledgebase planning project. Next is a summary of several topics presented and discussed during the workshop. Many of these topics require more discourse than could be fully covered during the meeting itself. Several groups and individuals were assigned to elaborate on these topics for inclusion within the report. These expanded topics are in the next section but directly refer to topics in the preceding section. For example, in the discussion of science objectives, having illustrative examples of workflows for the study of microbial communities was desired. Finally, there are appendices containing the participants list and agenda.

Prior to the workshop, participants were asked to consider the following **charge questions**:

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

As part of the Knowledgebase planning project, DOE is sponsoring a series of community workshops to establish the requirements for the Knowledgebase and to outline a plan for implementing them. Previous meetings include the following, and the output from each is available online at www.systemsbiologyknowledgebase.org/workshops.

1. **Knowledgebase workshop at the Supercomputing conference** in Portland, Oregon (November 2009). Explored the potential for applying the cloud computing approach to systems biology research.
2. **Joint USDA-DOE Plant Genomics Knowledgebase workshop** at the Plant and Animal Genome meeting (January 2010). Addressed the Knowledgebase requirements necessary for developing data capabilities for plants.

3. **DOE Genomic Science Microbial Systems Biology Knowledgebase workshop** at the DOE Genomic Science Contractor-Grantee (PI) meeting in Crystal City, Virginia (February 2010). Outlined workflows and data integration methods pertaining to microbial sciences that can inform Knowledgebase specifications and requirements.

Since the goal of the Knowledgebase planning project is to develop an initial prioritized plan for a useful systems biology knowledgebase, there is a continued consensus that these initial efforts cannot be all things for all users. It is better to show strong success in a few areas than minimal progress in many areas. That this needs to move forward is also reflected in the standards discussion below. Having too broad an approach has stymied and slowed past efforts.

## 2. Background

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

### *Knowledgebase Vision and Background*

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

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

## 3. Topics Discussed at the Workshop

This section attempts to briefly summarize the wide-ranging discussion during the meeting. Where there appeared to be a general consensus, this is indicated. The level of discussion detail was not the same for all topics, and thus the level of detail in this report is uneven. Many of the topics were assigned to participants to develop further details after the workshop for inclusion in this report. Discussion of science objectives and the resulting workflows (Section 3a and related Sections 4a and 4b) was the primary focus of the meeting.

### *3a. Proposed Science Objectives for Microbial Community Analyses*

The earlier 2009 report summarized needs and visions for knowledgebases. Here we are challenged to define precise science objectives: What do we want to accomplish in the science now? These prioritized science objectives will be a mix of priorities for importance and for current feasibility. Most of these objectives will require the exchange of data and insights (i.e., knowledge). To drive this interoperation, the Knowledgebase must have challenge problems that require cooperation and integration. A number of science objectives were described and discussed at the workshop. More potential objectives were gathered from online input to the charge questions. It will be obvious that there are common themes within objectives articulated in the report and in the earlier workshops. However, there are some unique aspects with respect to metaomics, or microbial community studies.

**Some of these unique aspects with respect to microbial community studies are:**

- *Massive amounts of data.* There will be terabytes of data resulting from genomic sequencing and increasingly from other techniques.
- *Datasets that never "close."* Unlike a genome for a microbial isolate, one can never finish—more data will just provide deeper details and resolution without reaching an inherent endpoint.
- *Experimental protocols will continue to develop and rapidly change.* An example is the increased application and development of RNA sequencing technologies.
- *All studies are studies of populations.* Even a species within a natural community must be considered ultimately as a population of genetic individuals that will change and evolve.
- *Natural communities are closely linked to their environmental context.* Unlike a laboratory study, this environment will not be controlled and must be observed. Despite the best available knowledge to capture the most important measurements, these observations will be incomplete. This provides a serious metadata challenge.
    - Note: Metadata is the associated data and information that provides context for the primary dataset. For example, a microbial community is analyzed for its metagenome by 16SRNA (the genomic sequences are the primary dataset). The metadata would be, for example, the location, time, environmental conditions, method of genomic isolation,16SRNA.

**Four science objectives** for initial study of microbial communities were proposed and discussed during the workshop. These were broadly affirmed as valuable by workshop participants. However, these and the expanded list were not prioritized during this meeting. The prioritization of these and other objectives will be a primary goal of the final Knowledgebase workshop in June 2010. The objectives discussed were:

- Metagenome analysis workflows
- Genome-based prediction of culture conditions
- Linkage and feedback from transcriptomic and proteomic data to gene calls
- Expanding metabolic pathways from metabolomic data and linking to other datasets

**Metagenomic analysis workflows** were seen as important in both this workshop and the one held in conjunction with the Genomic Science PI meeting. This workflow discussion has been given its own section below (Section 4a). One example of this challenge problem is that the first phase in analyzing a metagenome is done at one site, export to the binning into analysis of organisms at another site, exporting for pathways analysis at another site, followed by regulatory analysis at another site. This would drive interoperation and connections between the different groups, resulting in great science. More participants liked this collaborative model, but some preferred an approach where analysis tool needs are identified *a priori*, the tools are developed and distributed via the Knowledgebase, data is analyzed using those tools, and feedback is provided to the developers.

The need to develop expanded workflows relevant to the science community studying the microbial communities was recognized at the PI meeting and at this meeting. A small group was assigned to work offline on describing such workflows—both present and needed. Their effort is almost a stand-alone report and is presented in Section 4b and briefly summarized below. *The recommendations from this sub-report should be expanded upon to create a more detailed initial guidance in the final workshop.*

From the perspective of the metagenomics community, the DOE Systems Biology Knowledgebase will need to fulfill a range of requirements to achieve the research community's envisaged goals. These include:

- Providing a common mechanism for collecting, organizing, annotating, analyzing, and distributing data that enables easy data sharing and comparative analyses.
- Facilitating ***dynamic*** interconnection of data types, data sources, applications, and workflows to allow ***data integration*** for biological insight.
- Enabling researchers to identify, assess, and access all relevant ***datasets*** worldwide.
- Allowing scientists and facilities to "publish" their data, applications, and workflows into the "live data network."
- Providing space for larger-scale data integration, analysis, and publishing.

- Providing scientifically accepted rewards for researchers who "publish" well-annotated, good quality data, applications, and workflows.

We suggest that this could be achieved through the development of:

- A set of community-accepted semantic description formats (ontologies)
- A peer-to-peer based system of data, metadata, ontology, analysis tools, and workflow registration repositories that are integrated in discovery, access, and utilization through common semantics.
- Guidelines and software libraries that allow scientists and facilities to "publish" their data, applications, and workflows into the Knowledgebase in a set of agreed forms.
- A mechanism that allows scientists and facilities to easily and rapidly annotate, change, and correct research results and annotations in the Knowledgebase, capturing source, reason, quality, and proof for changes.
- User-friendly interfaces (APIs and people) to access data and application modules, ***as well as*** derived data products, enabling other users to build novel solutions with the data.
- A framework of citable, unique identifiers for data, applications, workflows, and researchers.
- Guidelines, training, and workshops for all new products and concepts provided by the Knowledgebase.

**Genome-based prediction of culture conditions.** Here the challenge is: Using a partial single microbial genome found within microbial communities, can we predict how to cultivate (and isolate) this target species? Put another way, can we predict culture conditions from genomic information? This Knowledgebase tool will be very valuable in rapidly culturing currently "unculturable" isolates from microbial communities. This would expand the study of difficult-to-culture or new microbes with interesting properties. This could lead to better integration or new experiments where one could envision testing 500 isolates a day to achieve a goal of studying newly discovered organisms with unique properties faster and cheaper.

For example, if the genome identifies heterotrophic metabolism features, will this organism grow on lactate? Is it an auxotroph, or will it require some amino acid supplement? This tool would tell you what experiments are necessary to test the proposed metabolism hypotheses. Further development of this concept would be needed including: What aspects of this tool could be automated? After the success or failure of the initial experimental cultivation tests, what information should come back to you? How do you incorporate knockout data, and can you predict the effects of knockouts? This becomes a capability tools and challenge for both the informatics and experimental communities.

The prediction of culture condition is the initial goal, but this scientific objective can be seen as the first step to a broader scientific goal in the area of genome-based functional prediction. This high-level goal would move knowledge from genetic information (which is more and more easily available compared to other data) into molecular or protein function, then to organismal function, and on to community function. It is complexity across scales. These studies are a prerequisite for investigating the function of both microbes and microbial communities. At a higher level, this would also provide potential data to feed back into improved annotation and validation. However, as stated in many other objectives, the consensus among workshop participants was that this initial effort will move most rapidly if used to address a specific problem. Each of these objectives would require detailed workflows to be developed.

**Linkage and feedback from transcriptomic and proteomic data to gene calls.** This scientific objective is a subset of the broader need to improve gene calls or annotation. The higher-level needs to move annotation beyond simple homology inferences were well described throughout the 2009 Knowledgebase report. The challenge here is using the massive amounts of data from transcriptomic and proteomic measurements to improve gene calls. This data is already used in the most straightforward manner—to promote gene calls from hypothetical to putative when a transcript or protein signature is observed. However, even this use does not often extend beyond the specific metagenome or genome under study. We need to find ways to draw further functional confirmations to improve gene calls, to invalidate and correct false calls, and to provide better descriptions for use in further homology searches.

With the rapid improvement of techniques such as RNA sequencing, it is clear that transcriptomic data for metagenomic communites soon will not be limited by the current requirement for an *a priori*—determined metagenome for that community. This will also enable better proteomic data analysis. This will require improved cluster analysis and the inference of pathways and function. Localization data from parts of the community (such as using laser dissection to gather small samples) will be needed to create estimates of community structure and function.

**Expanding metabolic pathways from metabolomic data and linking to other datasets.** There is a clear, if sometimes difficult, path from genomic to transcriptomic and proteomic datasets. Each is linked by the underlying gene. There is a different challenge in taking metabolomic data and validating and expanding metabolic pathways, as well as linking these pathways to the proteins and regulation. Since metabolites are pathway oriented, not genome oriented, the challenges of metabolomics will be largely similar, whether dealing with single microbes, communities, or plants. A related issue and challenge is extending metabolite concentration data into flux estimates. Due to tightly controlled multistep pathways, key intermediates can be present at very small levels, while the flux through that intermediate is large. With metabolomics, thousands of metabolites might be detected. However, there may be no final answer, and the dynamic range issue can confound the depth of analysis (concentrations can range from mM to single molecules). On the positive side, while there are thousands of metabolites potentially present, most experimental research targets, particular processes, or pathways (with the identification and quantification of tens of metabolites) are all that is needed. Still most metabolomic techniques are untargeted (i.e., they try to measure everything).

Challenges here include the positive identification of detected metabolites. For example, in the synthesis of lignocellulosic biomass, there are many similar compounds such as sugar isomers. The gold standard is the purification of synthesis of a compound for use as a standard for identification. As identification libraries continue to expand, do we need to save raw data to allow later identification of metabolites from saved data?

Another challenge is to link confirmed metabolites with the measured proteins that catalyze that reaction. (Note that this requires the correct functional identification of the protein.)

Clustering, visualization, and other tools are needed to extract insights from metabolomic data. We need to have these both for microbes and for observing the change of function within a community. This is needed to determine how the rest of the microbial community environment influences the pathways of member organisms and how they utilize their genetic potential. These tools should also highlight apparent "gaps" in pathways where either metabolites or enzymes do not appear to be present. This can help identify needed experiments to fill in important pathways.

**Other potential science objectives** have been proposed from several other sources. Workshop participants were reminded to return to the broad objectives in the 2009 Knowledgebase roadmap. There were also a number of potential science objectives suggested in response to the charge questions and posted by participants on the Knowledgebase wiki site (www.systemsbiologyknowledgebase.org). We are continuing to extract these objectives and will place them in the final report in Section 4f.

### *3b: Standards: The Role of Standards-Setting in the Knowledgebase*

Standards to expedite data and file sharing are important. Gene sequence data is relatively established as a standard. mRNA expression (MIAME) and other standards are being developed. However, participants had a range of opinions on the priority of standards (i.e., when do we focus on the standards?). Historically, standards development committees by community consensus have taken a very long time, and there is a need for this effort to move faster. Part of this long duration is driven by the desire to make the standards do all things for all people and uses. For example, required metadata lists quickly become wish lists of all possible information. There have also been "dictatorial" attempts at setting standards. These can lead to frustration as they are outgrown, such as in the file formats used for annotation for the last decade. Nevertheless, at a minimum, there was agreement in the need to have some standards for file-sharing formats to expedite transfer (I/O protocols). On the other side, there is the sense that if we do the needed work, the standards will sort themselves out. If the data exists, and there is a need to share, "someone" will create a protocol for sharing, which in effect is a small *de facto* standard. The challenge here is that this leads to duplication and balkanized tools. Within the context of this workshop, the range of consensus was narrower after the discussion. *Standards are important, but standards-setting is not the first task or top priority of building a*

*Knowledgebase community. This workshop, the developed workflows and the final workshop report need to focus on science needs and what the initial Knowledgebase version 1 will do. If some standard setting is required as part of this implementation, it can be addressed at that time.* There was an agreement that this group not be distracted into spending time in the actual standards discussions. Beyond the need for I/O, it was not clear that major effort was required in standards-setting in the first year or two. Broadly, the first two years of the Knowledgebase should focus on implementation data and tools to enable specific science.

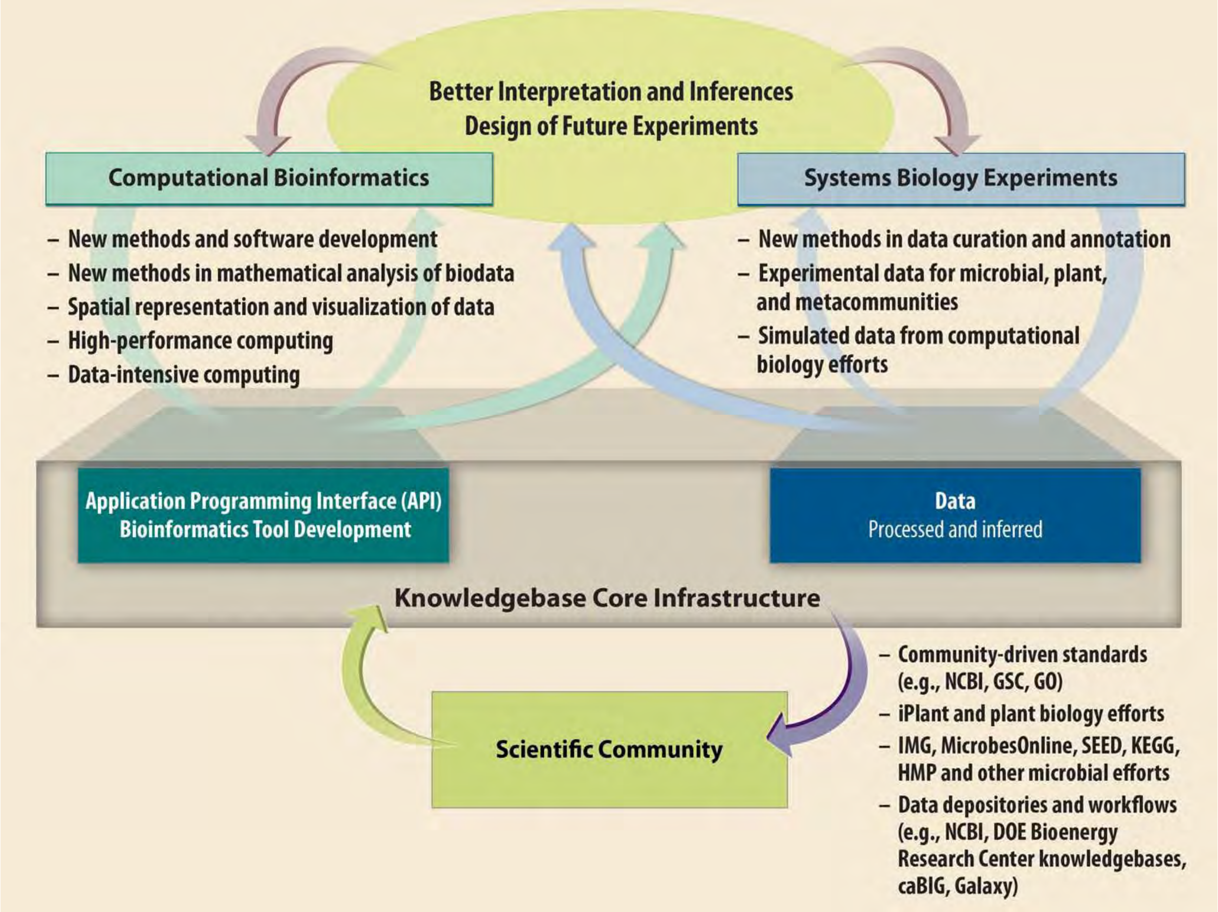
This I/O-focused approach is re-emphasized below in the API interface discussions and workflows. There is a minimalist view that standards are actually formalized file formats, but the discussions of required metadata move beyond that interpretation.

### *3c. Tool Builders and Data Generators: The Need to Engage the Various Scientific Communities*

Another consensus was on the goal of knowledgebases. The Knowledgebase will enable better understanding and interpretation by the "experimental" biologist and will enable testing and development of new analysis tools by the computational biologist. This reaffirms the goals stated in the 2009 report and showcases two critical science communities essential for the Knowledgebase: (1) the computational biologist or bioinformaticians who build the tools and (2) the systems biology data generators who design and run the experiments and usually provide initial interpretations. Both need to provide insights and inferred knowledge to each other through the Knowledgebase then out to the broader scientific community. This concept is presented at a high level in the Fig. 1. A challenge for both groups is the need for confidence versus just information. This was well articulated as: "I'd rather have less data but be more confident that the data is "real. I'd rather see less data with higher quality." This data would be used to create processed interpretations, like the calling of a gene. This is a challenge in assessing quality and confidence in the sea of data. For example, it is hard to assess and utilize negative experimental data because publications release only what worked. Elsewhere, frustration was expressed at the loss of underlying information when the data is processed. For example, more information goes into the calling of a gene than is saved in BLAST (i.e., intermediate analysis is lost). Also, the identification of a protein from three peptide fragment hits will lose the possible post-translational modification data hidden in an MS spectra from a "missing" peptide fragment. The combination and cross-correlation of multiple datasets from different sources into a synthesizing computational analysis struggle with different qualities of data and unreported conditions. An example recent work shows that errors in genome annotation are propagating.[1]

[1] Schnoes, A. M., S. D. Brown, I. Dodevski, and P. C. Babbitt. 2009. "Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies," *PloS Computational Biology* **5**(12), e1000605. doi:10.1371/journal.pcbi.1000605.

**Fig. 1. Relationship between the DOE Systems Biology Knowledgebase and the Larger Scientific Community**


Better Interpretation and Inferences
Design of Future Experiments
Computational Bioinformatics
– New methods and software development
– New methods in mathematical analysis of biodata
– Spatial representation and visualization of data
– High-performance computing
– Data-intensive computing
Systems Biology Experiments
– New methods in data curation and annotation
– Experimental data for microbial, plant, and metacommunities
– Simulated data from computational biology efforts
Application Programming Interface (API) Bioinformatics Tool Development
Data
Processed and inferred
Knowledgebase Core Infrastructure
Scientific Community
– Community-driven standards (e.g., NCBI, GSC, GO)
– iPlant and plant biology efforts
– IMG, MicrobesOnline, SEED, KEGG, HMP and other microbial efforts
– Data depositories and workflows (e.g., NCBI, DOE Bioenergy Research Center knowledgebases, caBIG, Galaxy)

There are two communities that must be both served and enabled by the Knowledgebase. One focus needs to be on the biologists sitting at their computers having done an experiment, who want to understand their results. Another focus is enabling tool builders. All agreed that this is not an ever-growing but static archive. *It is a combination of new* experimental data and tools that accesses a growing reference data. By having common access to quality data, tool-builders will also have the transformations of the data products in one place. This should accelerate the evolution of transformations and provide a better process for designing new data products. Some innovative ideas in this arena were suggested. These included tools registries, challenges and challenge grants to answer "tools needs," and Facebook-style entries of "my experiment" to advertise. There is a more detailed section on how a tools registry might function (see Section 4c). A vision is to improve data analysis sufficiently that experimental sample generation becomes a bottleneck, despite the massive amounts of data generated per experimental sample.

This can start with well-defined workflows leading to a mapped interface with access to data and tools. Then we can string tools together to do powerful operations without having to worry about the data formats and come up with an answer. This answer may be a better interpretation or a better designed experiment. This may link into web services and other models. This interface or API should be practical, with not much investment early on.

An early priority should be to develop initial APIs to get to the data with an interface to the tools. The API should make it easier to develop and test new tools for biologists to use and to add them into the interface. There was broad consensus on the importance of the API—that it should be modular and interchangeable. An open question for the architecture and the data transfer challenges relates to how much analysis is done where the data resides versus at the viewer's site (download and I/O concerns).

There is also the need to consider architecture that relates to the massive data transfers that we are potentially considering, especially in a federated or distributed approach. This can be considered as the datasets increase; it is more and more difficult to perform "all versus all" comparisons. A number of web-based applications for metagenomics exist that do not currently support large-scale sequence analysis, including, but not limited to, on-demand clustering of user-provided datasets. Thus, the evolution of centralized data repositories and analytical services in metagenomics is currently not in sync with the accumulation of next-generation sequence data as it relates to end-user capabilities. To put things in perspective, consider the initial ScalaBLAST calculation in which comparing 1.6 million proteins in the IMG 1.6 database against the 3.2 million proteins in the nonredundent database consumed approximately 5 years of CPU time. Individual investigators simply cannot achieve these calculations due to technical or infrastructure limitations, and even if they could, the visualization tools needed to interpret and compare next-generation metagenomic and metaproteomic datasets do not scale with data volume and complexity. While good progress has been made in developing tools to inventory and, to a lesser extent, to compare microbial community structure and function, there is no comprehensive tool that allows integrating and comparing multimolecular datasets (e.g., DNA, RNA, protein, and metabolites), which are needed to fully realize the vision of microbial systems ecology.

***There is continued consensus for a federated model***. However, this federation cannot be the current system of separate unconnected sites. Here, federated means distributed resources, data, and tools but integrated and coordinated in a manner to be apparently seamless to an outside user. Some very mature examples are the current genome data repositories (e.g., GenBank), which actually are distributed in three sites (the United States, Europe, and Japan) but appear as one to the science community. Of course, reaching this level of integration will take a long time and effort and is beyond an initial plan. The use of a federated model brings with it the underlying challenge of how much centralization is required in deposition, curation, or "advertizing." (Note: "advertising" was discussed as a possible mechanism to draw attention to new datasets or tools in the ongoing development of the Knowledgebase.) A possible consensus in this group was that this does not matter as long as the access and the goals can be accomplished.

The development of an API allows the potential of an "open-source" system. The potential and challenges of "open" systems are discussed in more detail in Section 4d.

This use and development and data deposition in knowledgebases must be balanced with the need for some level of public/private embargo and the need to further the careers of bioinformaticists and experimentalists. This was deemed important and is covered in more detail in Section 4e.

### *3d. NIH Interactions: Data Resources and Leverage*

There was discussion about the need for awareness, linkage, and leverage with NIH-led efforts, in particular NCBI. Current and planned NCBI efforts are described elsewhere. Workshop consensus was that we should leverage resources as much as possible. In particular, we should use both existing and under-development NCBI capacities as an archive and repository as much as possible. But there will always be a gap in filling current BER Genome Science needs and challenges, therefore we will also need our own efforts and to link them with other projects.

## 4. Expanded Discussion from Workshop: Assignments

### *4a. Workflows and the Systems Biology Knowledgebase*

In bioinformatics, complex biological analyses frequently require large-scale computations that compose standard tools and methods into a pipeline, or workflow, that runs a series of tasks to achieve a specific outcome. There are two major types of workflows, namely:

1. **Ad hoc Interactive:** In ad hoc interactive workflows, the biologist is fundamental in driving the steps involved in the workflow. Interactive tools (e.g., Cytoscape, DMV, R scripts) are used to analyze and visualize data, and the results from one tool become the inputs to the next tool in the workflow. The biologist typically drives the transition between tools based on his or her observations of the state of the analyses, and data is moved between tools either manually (e.g., saving files in specific formats) or by using a lightweight data transfer tool like Gaggle (www.systemsbiology.org/Technology/Data_Management/Gaggle).

2. **Automated:** Automated workflows, also called pipelines, take a set of input data and apply a series of analyses to the data to produce outputs. No human intervention is necessary to invoke the next step in the workflow and to transfer data between computations. Automated workflows can take anything from seconds to weeks to execute, and the steps in the workflow are commonly controlled by scripts or workflow tools like Taverna (www.taverna.org.uk/).

While the precise software mechanisms used to coordinate the steps in a workflow vary between the interactive and automated cases, the ease of construction of workflows in both cases is hampered by two fundamental technical issues:

- **Tool heterogeneity**: Standard tools and algorithms are not created using a common software framework so that they can be readily “plugged together” to form a workflow.

- **Data heterogeneity:** Standard tools and algorithms consume and produce data in a variety of different data formats. Feeding the outputs from one tool into another commonly requires data transformations to produce inputs in a format that a given tool is expecting.

For these reasons, creating effective bioinformatics workflows is non-trivial and requires considerable effort from biologists and software engineers alike in order to meet scientific objectives.

The Systems Biology Knowledgebase is an opportunity to address the current complexity of creating both interactive and automated workflows. The Knowledgebase can create a lightweight, flexible software infrastructure that enables tool developers to "componentize" their existing and new algorithms, providing standard interfaces that can be used to compose tools into workflows. In addition, the Knowledgebase infrastructure can support the flexible, discoverable definition of data formats that tools produce and consume. By describing a given tool's data requirements using metadata, converting data from one tool to another becomes simpler, and potentially automatable.

**We therefore recommend the Knowledgebase implements a set of simple programming interfaces that enable much more effective workflow construction and reliable execution.** By reducing the "levels of pain" experienced by biologists and software engineers in creating workflows, we envisage the creation of a software ecosystem in which useful workflows can be rapidly built, deployed, and shared with the community through the Knowledgebase infrastructure. This would be analogous to social networking sites such as Facebook, which encourage development and sharing of new applications based on the software infrastructure and programming tools that Facebook makes available. This model, which is expanded upon in the next section, is designed to (1) encourage development of many tools that provide multiple approaches to solving a particular problem and (2) enable the end-users to determine which approaches survive. Applications that accurately solve a problem in a particularly elegant or succinct manner will become highly adopted, and others will slip into oblivion.

A primary issue to be addressed under the Knowledgebase plan is the motivation of the developer. Platforms like Facebook, Twitter, or the Apple App Store can provide a financial incentive for *de novo* application development. Knowledgebase infrastructure and early applications will need to be developed under more conventional models. But as the programming platforms become established, the project will need to consider funding models designed to maintain and expand innovation over the long term. This may include a combination of standard funding models and models designed to reward *de novo* application development. Failure to address this basic issue will almost certainly result in stagnation of the development cycle.

## *4b. Metagenomics Systems Biology Knowledgebase: Workflows, Background, Design Goals, and Recommendations*

### Integrating Metagenomic-Enabled Workflows

Most metagenomic data come from microbial ecosystems. Data derive from a broad range of environment types—from the deep subsurface to the human gut—motivating many questions such as how are microbial communities structured, and how do they function? Do genetic profiles vary across environment types? Using metagenomics to answer such questions will require the effective integration of information about metabolic potential (genomic sequence); metrics for function (proteomics, transcriptomics, metabolomics); contextual information; data that define the physical and chemical environment (metadata); and methods to consistently and accurately update annotation as new evidence becomes available (see Fig. 2). Metagenomic data may be collected from one or many samples, whereas proteomic, metabolomic, and transcriptomic data typically stem from a diversity of experiments such as time series, environmental perturbation, and genetic manipulation.

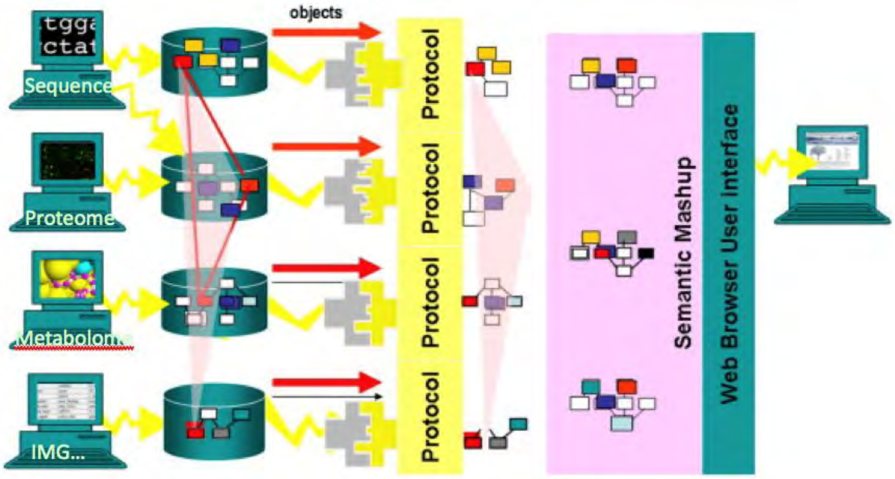


**Fig. 2. Data Warehouse.** Data are linked by common or discovered identities and shared annotations (tags) drawn from controlled vocabularies, and managed by identity and ontology authorities. The data are accessed through simple application programming interfaces (APIs) and aggregated through browser scripting based on common identities and tags. [Slightly modified from Goble and Stevens 2008.]

One of the most important aspects of metagenomic investigations is that sequence information is (or will be) intimately linked with proteomic, metabolomic, and transcriptomic data. Commonly, metagenomic workflows begin with sequence information, but a Knowledgebase—an artificially intelligent tool that provides a mechanism for collecting, organizing, analyzing, and distributing data—must be designed to facilitate dynamic interconnection of these data types to allow data integration for biological insight (see Fig. 3). The need for dynamic interconnection is underlined by the observation that data can exist in many states: "there is live data, living data (more live than live), stale data (archived?), dead data (archived?), lost data, vandalized data (valid data overwritten by non-valid data)."[2] For example, when data consumers download a specific dataset from a resource and put it into a new form for their own purposes, the data become disconnected from the original source in the absence of dynamic linking or a provenance system. It therefore cannot benefit from changes or upgrades in the source (i.e., it becomes "stale" or "dead"). Examples of changes at the source include reassignment of a gene function, re-searching of a proteomic spectra database with new genomic sequence, and changed identification of a metabolite due to the addition of new, standard metabolite profiles to reference databases. These types of data insertions, deletions, and mergers represent problems for all subsequent users of a resource.

[2] Web commentary, www.ted.com/talks/tim_berners_lee_on_the_next_web.html

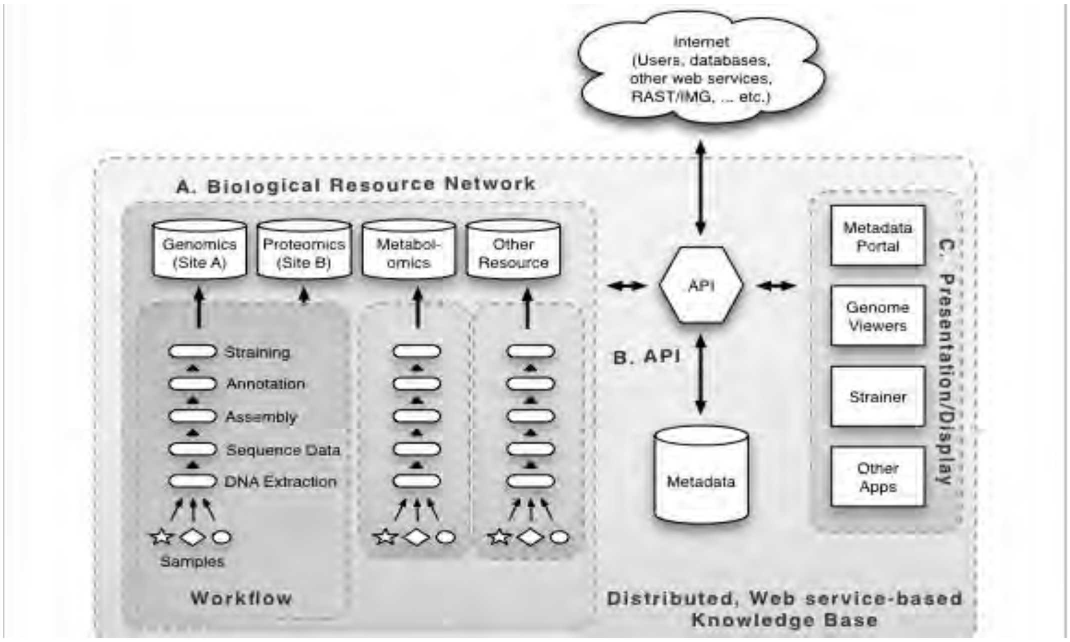


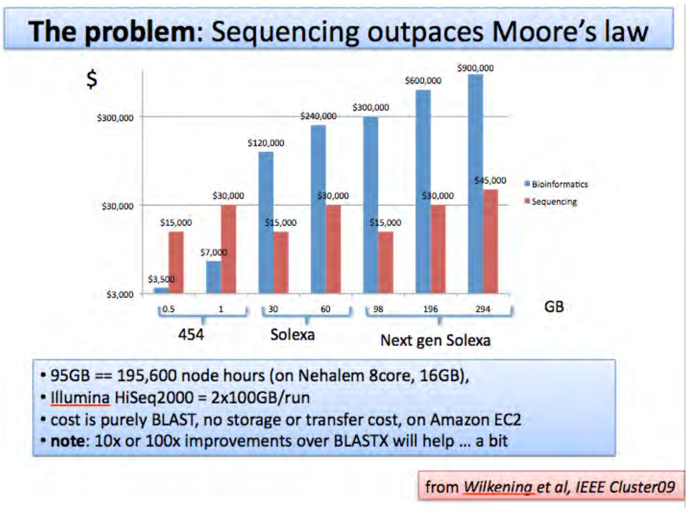
**Fig. 3. Overview of Metagenomic-Enabled Workflows Feeding into a Knowledgebase System.** This figure provides examples of data-generating experiments, data types, file formats, and processing steps. Note that not all studies include other "omics" methodologies. Data derive from a common source (a natural sample or series of samples, an experiment or series of experiments), and data are integrated via tools to answer specific questions. Lines with arrowheads represent data flow, cylinders indicate a data resource, and rectangles indicate an application (stand-alone or web-based). Data analyses can draw on a wide variety of existing and newly developed tools (e.g., assembly programs, gene prediction, functional annotation, analysis of regulatory structure), as well as tools developed specifically for metagenomics (e.g., example methods to visualize and analyze strain variation).

## Metagenomic Data Challenges

Five key attributes associated with metagenomic data pose challenges that require special consideration.

**(1) Volume of Data Generated.** Sequencing of metagenomes generates somewhat to highly fragmentary datasets, often with low redundancy levels and potentially high error rates (due to low genomic coverage with error-prone sequencing).

The growth in sequencing capabilities has led to a flurry of metagenome sequencing and analysis projects in recent years. Interestingly, computational analysis costs are now quickly outpacing data generation costs (Goble and Stevens 2008). As shown in Fig. 4, running similarity searches (BLASTX) for data generated by one run on an Illumina GS-FLX instrument (costing approximately $15,000) will take 60,000 hours of compute time on a recent machine (or cost approximately $120,000 if run on Amazon's EC2 service).



**The problem**: Sequencing outpaces Moore's law

- 95GB == 195,600 node hours (on Nehalem 8core, 16GB),
- Illumina HiSeq2000 = 2x100GB/run
- cost is purely BLAST, no storage or transfer cost, on Amazon EC2
- **note**: 10x or 100x improvements over BLASTX will help ... a bit

from *Wilkening et al, IEEE Cluster09*

**Fig. 4. Cost Comparison.** The computational costs (blue) are rising significantly with growing sequencer yield. Already they can be up to 10 times the sequencing cost (red) on some instruments. Costs shown are for running BLASTX against the current National Center for Biotechnology Information (NCBI) non-redundant protein database. [From Goble and Stevens 2008.]

An associated problem is that while hundreds of metagenome datasets are publicly available, these are widely distributed across the world, and it is at times challenging to identify and access all relevant datasets. Many more experimental data sources are not publicly discoverable. Therefore, few high-profile studies have emerged that attempt the systematic comparison of available public data because this task is difficult (and, in part, because currently the primary focus of most investigators is on their own datasets). One way of enabling both general discovery and access (e.g., via web services, as well as "local" access to the data) is via a peer-to-peer based framework of dataset registries. Unique identifiers such as Digital Object Identifiers (DOI) known from publications, URLs, or persistent URLs (PURLs) could enable the citation of high-quality, well-annotated datasets, and offer identification and linkage between datasets. Any such framework also would encourage metadata and semantic enrichment, enabling queries such as "Display all soil metagenomes from the Midwest" or even "What is the number of short-read metagenome datasets currently available from a specific sequencing platform."However, without accompanying metadata, the sequences may be next to useless for some of the broader purposes.

**(2) Never-Ending Analysis.** With metagenomic data investigations, the analysis is never finished. Unlike an isolate genome where a final set of open reading frames and relatively stable functional annotation is generated, metagenomic datasets are prone to continual revision resulting from new sequence data, additional manual curation, new reference genomes, and more. Gene numbers, for example, may change multiple times, and so can organism assignments of genome fragments. Although curated metagenomic datasets share many features with isolate datasets, there is a major difference: there is no single answer in almost all cases because populations are not clonal cultures. In many current studies, the extent of metagenomic sequence curation is minimal, but this must change. A Knowledgebase system must be ***dynamic*** in order to be able to deal with this key attribute.

**(3) Sequence Variation.** One of the most important features of a metagenomic dataset is the sequence variation that is captured via the sequencing reads. Typically, each sequencing read derives from a different individual, and thus a metagenomic dataset provides information about sequence heterogeneity. Although it is arguable whether or not the primary sequence data must be retained (or just the base-calling scores), ***a system to tie read-based data to the composite, assembled sequence is essential***. In addition to providing access to such data, an annotated and permanent record of research steps taken during curation will be essential to establishing a powerful, relevant, and long-lasting scientific Knowledgebase. To our knowledge, this capacity is not available via any shared resource. Sequence-variation analyses will be critical to population genetics and evolutionary studies, for efforts to identify the reasons for fine-scale variation in functional attributes, and to locate potentially interesting gene variation for targeted bioengineering applications.

**(4) Tools and Computational Infrastructure Are Required for Data Sharing and Comparative Analyses**. As the number of sequencing and “omics” instruments available for metagenomic research grows, ever increasing the volume of available data, the community needs both the computational infrastructure to analyze and compare the data as well as the tools to analyze the results.

An important approach to tool development is the generation of a series of modular components (as opposed to large, integrated pipelines suited to run on a centralized computational facility). In addition to portability, an advantage of the “components”-based approach is that the user retains considerably more flexibility with regard to the way in which the data are processed. An example of current interest involves software for correcting homopolymer errors in 454 sequences. Currently, this capability is contained within a complex package within a pipeline. For practical reasons, it is undesirable to send the entire dataset to a staff member at the centralized facility for homopolymer correction, and, more importantly, data reprocessing in a new pipeline will unlink information already associated with the sequence.

**(5) Diversity of Data Types, File Formats, and Processing Steps.** Scientific research itself has become more specialized on the individual level and more collaborative and international on the community level, making it desirable and necessary to relate one’s own local research results meaningfully to the geographically distributed, multifaceted results being compiled elsewhere in the world. Systems biology has a long tradition of utilizing diverse research results from experimental and computational methods that stem from varied and distributed sources. Commonly, these research results are locally integrated and synthesized with the scientist’s own findings, then published as yet another valuable source of information. Over the years, the community has created a wealth of outstanding data sources and tools for access, integration, and analysis. Unfortunately, these ***sources of scientific knowledge and analysis are mostly characterized by a diversity of data formats, data representation, metadata, and access methods,*** making it difficult to identify all of the relevant data sources for a given topic, assess their quality, and integrate them into the scientific research process.

**Architecture Design Goals and Knowledgebase Adoption Strategies**
As experiences in other scientific communities have shown, it could take many years to change working practices and move to a central deposition system based on community-agreed, standard data formats, metadata, and semantic data descriptions, with associated discovery and analysis tools. This type of integration has been most successful in slower-moving fields with less diversity in their research methods than metaomics, fewer data sources, and more standard analysis software. In these fields, such integrations have been very successful in terms of making data more widely known, used, and effectively analyzed by the community. Despite the larger challenges faced by our community, such central deposition systems have their place in the Knowledgebase, but it appears unlikely that the more rigid structures of a central data and applications repository could effectively meet all systems biology needs. Instead, a more flexible approach is needed. We advocate combining the benefits of the more rigid, standardized frameworks with a “live data network” of shared experimental results.

Ideally, the live data network component of the Knowledgebase would be compatible with current, more distributed working practices, while at the same time assisting with greater integration of resources. A peer-to-peer based system of data, metadata, analysis tools, and workflow registration repositories, integrated through common semantics would seem desirable. In this scenario, users could “publish” their data, applications, and workflows into the live data network by describing their provenance, content, location, access, and usage methodologies (both for other users and computer applications) in one of a set of community–agreed semantic description formats (ontologies). In designing these semantic descriptions and underlying metadata, it will be important to focus on the data content and ontologies. Important concepts that must be included, rather than the particular local implementation, need to be identified. Semantic ontologies will allow the mapping between different expressions of the same concepts (within reason), as well as linking concepts where their expressions do not overlap but are related.

If a “local” format contains desired information, it is relatively simple to map these data to the shared semantic description and make the local knowledge available in a community format. These conversion interfaces are not difficult, but they can be time consuming to establish. Therefore, offering different levels of community participation would be desirable. Initially, one might submit only enough information to allow others to discover one’s resource. More functionality can be added over time to support full integration. Alternatively, central data centers might offer services to smaller groups to integrate their data into the archive and “publish” it for them (including data analysis, such as normalization and filtering, annotation, and more). This type of tiered approach to Knowledgebase participation is critical because it would help to bring about adoption by the wider community by supporting the twin (and sometimes competing) goals of (1) defining relevant data standards and formats for participation in a more centralized repository and (2) the necessity of allowing dynamic integration to be done at a smaller, local, and scientific inquiry-driven level.

Similar to the underlying resources (see Fig. 5), links between people, data, applications, and workflows could be explicitly recorded and published, or discovered following the semantic description “trail.” This approach is analogous to the LinkedData Web proposed by Tim Berners Lee but is extended with more domain-specific information about the data, especially the applications and workflows, to aid directed scientific discovery and experimentation. An added benefit could result if the Knowledgebase provided space for larger-scale data integration, analysis, and publishing—following the same format described above—to aid smaller institutes and preserve important results after funding for their further curation elapses.

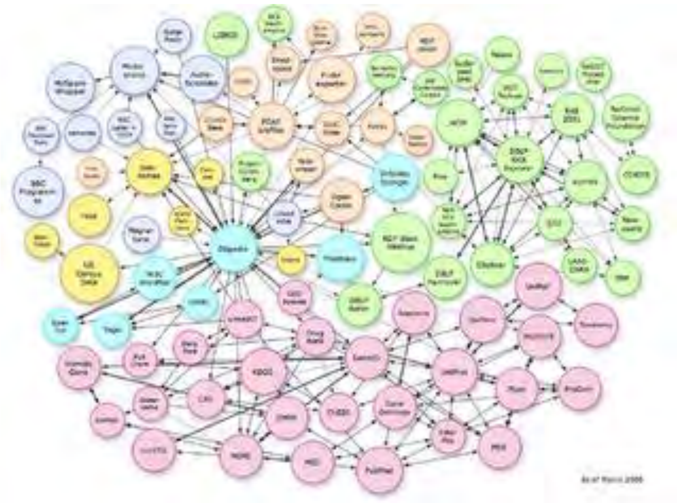


**Fig. 5. Linking Open Data.** This cloud diagram gives an overview of published datasets and their interlinkage relationships.

When designing the community-agreed metadata and semantic formats to be utilized, advice from the data curation community should be considered. Their aim is the development and provision of methodologies and tools that allow the perpetual reuse of data by its designated user communities, by keeping it "living" (i.e., adapting its representation to changing community trends without affecting its integrity). Data policy examples, data life cycle models, archival reference models, and more will help to define critical components of the data, analysis software, and workflow descriptions.

The Knowledgebase's success will depend strongly on the quality of the data, tools, and workflows that are available through it, as well as the ability of researchers to identify, assess, and integrate resources quickly. Much of the latter will depend crucially on the quality and extent of the metadata and semantic information about the resources. Herein lies a potential problem. Although good data or tools will usually lead to direct scientific rewards in the form of publications and resulting citations, the time-consuming annotation and preparation of sharable datasets is not directly rewarded by the community in similar fashion. This often results in a lack of motivation to provide this vital information, or to provide it with the required due diligence. One way to resolve this issue might be to enable scientists to "publish" their datasets, following similar style, content, and peer-review standards that they are required to follow throughout the publishing world (e.g., see DataCite at www.datacite.org/). This way, datasets would be citable and earn the owner well-deserved recognition and rewards.

Similarly, to encourage participation in the Knowledgebase, it will be critical to ensure that utilized data, analysis tools, and workflows are correctly attributed to their owner, an often difficult task due to similar names and changing names and affiliations. Recently, the research community has initiated more concerted efforts to develop unique identifier systems for researchers, although most of them are focused on publications (e.g., International Standard Name Identifier (ISNI) being developed by the International Organization for Standardization (ISO) as Draft International Standard 27729, and Open Researcher Contributor ID (ORCID) being organized by the ORCID Initiative) and not data elements. Additionally, protocols for maintaining permanent data identifiers are being developed and increasingly deployed in the biological domain. The combination of Uniform Resource Names (URNs) and Uniform Resource Locators (URLs) are one example. URNs represent a persistent, location-independent identifier, and they promote mapping to other namespaces. A URL is the specific location of a URN. For example, a specific protein name and its annotation from the Kyoto Encyclopedia of Genes and Genomes (KEGG) can be represented by both a unique identifier and a location. Using the Life Science Identifiers system (lsids.sourceforge.net/), a URN in the Knowledgebase might be "urn:lsid:doekb.gov:eco:b1743", which uniquely names a specific gene in *E. coli*. The location of the data element is accessed using the URL www.genome.jp/dbget-bin/www_bget?eco:b1743. This combination of URN and URL provide an unchanging name (the URN) and a location (URL) of data about the URN. Using a common naming system allows for data linkages and for the development of rich semantic descriptions. Another identifier system is the persistent URL system, or PURL. This system achieves the same goal of specifically naming and locating data, but it does not directly describe the location. Instead, it references an intermediate location that redirects the request to the proper location. PURLs depend on a master system for redirection and on contributors to maintain their links. As an example, the same *E. coli* protein described previously can be accessed from the UniRef Database using a PURL system from this link: purl.uniprot.org/uniprot/P77754.

Using unique identifiers (DOI, URN/URL, or PURLs) for data, and potentially applications and workflows, will make it possible to utilize the services the library and web technology communities offer for information discovery and access. Furthermore, different data publications and publishers could be easily linked through citations, as could the history and connectivity of data elements in the Knowledgebase.

### Some Data Sharing and Analysis Needs

As data sharing is becoming more common, data storage is essential, but as with the World Wide Web, it need not be centralized. As noted previously, repositories must be dynamically linked to information sources; otherwise, we run the risk of "stale data." To enhance research, each repository should consist of user-friendly interfaces (APIs and people) to access data and application modules, as well as derived data products, enabling other users to build novel solutions. One such product provides genomic neighborhood views across multiple genomes. Both the DOE Joint Genome Institute's Integrated Microbial Genomes (IMG) system and Argonne National Laboratory's Rapid Annotation Subsystem Technology (RAST) offer these views (see Fig. 6). However, such foundational capabilities need to be extended to meet real systems biology needs. For example, a user may need to integrate proteomic, metabolomic, or transcriptomic data in a display such as that shown in Fig. 6. Semantic data interfaces could provide vital support in this endeavor, as they insulate tool developers (including those of user-friendly interfaces and data products) from differences and changes in local data formats, data organization, and access mechanisms. Based on the community-agreed semantic descriptions (ontologies), any software can search and request data using semantic concepts; for example, "Give me the first protein of the second metabolic pathway that the system identified in KEGG for *Shewanella.*" This type of request would still continue to work even when the underlying data sources change, as long as the data sources maintain their semantic interfaces to the Knowledgebase. Therefore, any tool development effort can focus on new functionality, rather than redundantly expending effort on data search and access methods. Similarly, it will be much easier to combine development efforts that previously would have only benefitted selected data sources. Application maintenance will again require reduced resources, as changes to underlying data sources must not affect the usability of any tools, unless they want to benefit from any additional information (new concepts, not more data in the same structure).

In general, we anticipate building upon tools such as those shown in Fig. 6 to address new needs and provide additional capabilities. For example, a user may need to answer a specific question such as the temporal distribution of strain genotypes. Thus, the "display" tool may need the ability to output specific data characteristics as inputs to other programs (e.g., the library of origin of a set of sequencing reads, or the environment type from which specific gene contexts derive). A benefit of a web service—based system is that it allows for the development of tools for such specific questions. In general, continually capturing ***innovation*** by the broader community for purpose-specific application development provides a way to address the unavoidable limitation that no program can ever meet all possible user needs.

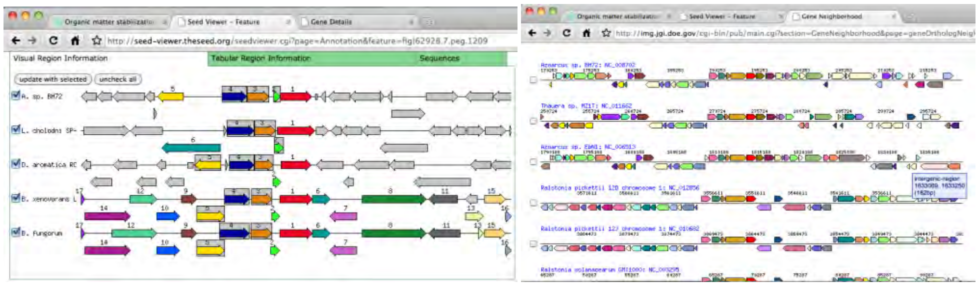

Organic matter stabilizatio
Seed Viewer - Feature
Gene Details
http://seed-viewer.theseed.org/seedviewer.cgi?page=Annotation&feature=fig|62928.7.peg.1209
Visual Region Information
Tabular Region Information
Sequences
update with selected
uncheck all
A. sp. BH72
L. cholodni SP-
D. aromatica RC
B. xenovorans L
B. fungorum
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Organic matter stabilizatio
Seed Viewer - Feature
Gene Neighborhood
http://img.jgi.doe.gov/cgi-bin/pub/main.cgi?section=GeneNeighborhood&page=geneOrthologNeig
Azoarcus sp. BH72: NC_008702
Thauera sp. MZ1T: NC_011662
Azoarcus sp. EbN1: NC_006513
Ralstonia pickettii 12D chromosome 1: NC_012856
Ralstonia pickettii 12J chromosome 1: NC_010682
Ralstonia solanacearum GMI1000: NC_003295
intergenic-region
1633089..1633250
(162bp)

**Fig. 6. Data Displays.** The chromosomal context of a protein in *Azoarcus sp.* BH72 as shown by the SEED-Viewer (left) and IMG (right). Both systems provide a similar view of data that are computationally expensive.

With the growing volume of data, there is an opportunity to introduce "lightweight standards" to allow the exchange of many sequence datasets and to reduce the cost of computational analysis. The Genomics Standards Consortium (GSC) has presented the Minimum Information about a Metagenomic Sequence/Sample (MIMS) standard, which allows the exchange of contextual data for metagenomes (e.g., location, sampling method, and biome description). This provides an initial description of the sample, but only minimal information about computational sample processing is included. The GSC's M5 working group has presented a draft metagenome interchange standard (MTF) that includes computational results and MIMS metadata (see Fig. 7). However, as important as standards are, they can have a downside. It is also important to ensure that standards are flexible enough to adapt to cutting-edge advances in the field, allowing specific users to add additional information.

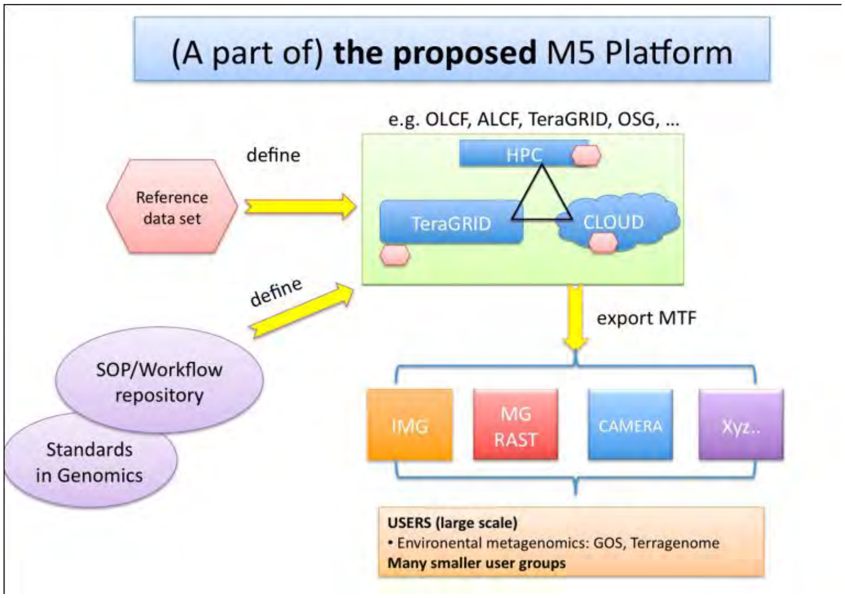


**Fig. 7. Standard Development.** The proposed M5 platform will include a standardized processing pipeline that can be executed by third parties (e.g., large supercomputer facilities), and, via a reference non-redundant protein database, it will enable many groups to use the results. JGI's IMG/M and MG-RAST, as well as the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA), are working on the M5 standard.

**Integrating data across multiple disciplinary Knowledgebase subcomponents.** An important concept for a DOE-wide Knowledgebase is that it will interlink components tailored to specific areas of systems biology investigations that have been or will be designed by those with special expertise with the various data types and needs. A possible solution is a Metagenomics Knowledgebase (MKb) developed on a database structure flexible enough to be ported to other laboratories and populated with data from any of a wide diversity of systems.

Such an MKb must enable free flow of data into and out of a larger DOE-wide Knowledgebase structure, as well as into and out of many other publically available databases (e.g., KEGG and Pride). One component will be data repositories (e.g., from a specific research group or team) with one or more data resources. Many studies will include both metagenomic sequence and other data forms (e.g., proteomic, metabolomic, or microarray data). Each data repository will have specialized tools, so that one might visualize links from an MKb to, for example, a project-specific transcriptomic dataset in the Transcriptomics Knowledgebase (TKb), or the Proteomics Knowledgebase (PKb). Developing guidelines for assisting local groups in creating resources such as these is an important goal for the DOE Knowledgebase. A system that describes and publishes data dynamically will greatly aid in this step, and incorporating the tiered participation option mentioned previously will ensure that all stakeholders are represented.

One of the most important obstacles to overcome is the lack of data integration in a form that enables data mining by groups not previously involved in specific kinds of experiments. To some extent, web-services type interfaces to the data and a comprehensive dataset registry will allow only a certain subset of queries and satisfy just a subset of researchers. "Web services" typically means an API that can be accessed over a network and executed on a remote system hosting the service. It is usually in one of two forms: "big web services" (e.g., KEGG) and representational state transfer services (RESTful services; e.g., UniRef and GO).

Currently, "the integration requirements of biologists working with unpublished data are not being widely addressed by the community."[3] To illustrate the opportunities associated with data integration across experiment types, the different data types generated in different laboratories by different methods (e.g., metabolomic data and proteomic data) must be considered. Another researcher might be interested in a specific protein of unknown function suspected to play a role in a certain metabolism. So long as the data are accessible, patterns may emerge through the integration of this information with information from other research groups. This can be accomplished as long as the experimental groups use a consistent "resource description framework" (RDF), a data description document that describes and links the data (e.g., gene A is translated into protein B).

Both centralized, high-performance computing (HPC) services and dynamic, "local" web services can be envisioned as important coexisting and linked parts of the MKb. For example, each of the experimental platforms (e.g., metagenomic sequencing, proteomics, and transcriptomics) requires extensive HPC, but the calculations and analysis only need to be done once, and the results then are shared (e.g., by using a common data description and a RESTful service). In fact, the raw data could be made publically available upon generation, and as the local system converts raw data to processed data, the broadcast version is continually updated ("live data"). This should be feasible within the framework of the laboratories that participate in data sharing, although it is likely necessary that central data warehouses will create new releases on a regular basis.

In the specific case of metagenomics, raw sequence is generated in vast quantities. Over time, these sequences are assembled and annotated. In parallel, comparisons among reads and assembled fragments reveal within populations variation in gene sequence and gene content. A whole suite of computational tools is required to collect and present the data as part of the in-house analysis, but this work product (the added value) currently is never distributed. For mass spectral datasets, a resource description could be generated that enables a researcher to access the up-to-date analysis and download components (e.g., reads, contig sequence, and single-nucleotide polymorphism (SNP) concentrations). In this way, long-running, CPU-intensive analyses can be part of a Knowledgebase approach that allows biological data integration via web service protocols. This will accomplish the important goal of keeping those responsible for data generation and data upkeep ("live data") connected to widely distributed data analysis tasks. Furthermore, the data are continuously enhanced as the network of links to new experiments expands.

[3] Anwar, N. and E. Hunt. 2009. "*Francisella tularensis novicida* Proteomic and Transcriptomic Data Integration and Annotation Based on Semantic Web Technologies," *BMC Bioinformatics* **10**(Suppl 10:S3), PMCID: 2755824. Electronic resource.

Annotations change as our understanding of protein families and complete genomes improves over time. Work on specific genes or proteins builds a body of functional evidence around each of these entities and their families. An optimal annotation system should permit easy and rapid updating of specific sequence annotations in response to new experimental data, and provide a straightforward method to record the ***source and quality*** of the updates. Automated annotation systems should be able to use the quality data to inform new annotations and update overlapping datasets. Such a system would dramatically reduce the time from discovery to annotation and enable very richly annotated descriptions of genes, pathways, and organism function. It also would facilitate collaborative research by disparate laboratories focused on a common genetic system and reduce propagation of erroneous annotations into new datasets.

The primary problem with such a system is ensuring the accuracy of changes made by the broad research community. In the 1980s, GenBank dealt with this problem by restricting annotation changes to the individual that submitted the sequence. However, the MKb must move beyond this static model to embrace systems used by organizations such as Wikipedia to validate changes. Therefore, in addition to recommending a MKb that focuses on sharing of "live" experimental and associated data, our working group recommends finding a mechanism in which new knowledge deposited within the data warehouses (e.g., NCBI) can be updated easily.

**Role of grand challenges in linking community, shaping Knowledgebase development.** Transition to a scientific framework in which data sharing and data integration is facile will present many challenges. At this time, small groups are beginning to address parts of the problem (e.g., metabolomic-metagenomic-proteomic data integration on a small scale), but the effort at the "omics" community level has a long way to go. A recommendation of our working group is to motivate the formation of linkages and overall architecture of a Knowledgebase with this goal via "grand challenges" that require data integration and sharing. We find this approach preferable to tasking a group of bioinformatics experts with establishing a system that will be later used by the community.

As one example of a grand challenge, consider the potential for data integration to improve protein annotation [g1]. Currently, most genomes encode a significant number of lineage-specific proteins that have not been studied biochemically. These proteins may hold considerable significance for DOE efforts in environmental remediation and bioenergy, as they may be involved in novel pathways for metal redox transformations or degradation of complex organic compounds. Similarly, there are probably many small non-coding RNAs coded on genomes and genome fragments for which annotations are lacking in public databases. Consequently, many gene predictions are uncertain, and a significant number of predicted proteins discovered via short-read sequencing may be corrupted by frameshifts. A single confident identification of a hypothetical protein via proteomics (or transcriptomics) converts a "hypothetical" to a "protein of unknown function" (an annotation that could be amended with the words "validated by proteomics"). If all the curated proteomic data from all samples worldwide could be integrated in a single analysis, many annotations in public databases could be updated. In addition, detection of the first peptide in a protein can confirm either the start site or the truncation status of the mature protein (e.g., due to cleavage of signal peptides). High-throughput improvement of start-site information from either proteomics (or transcript sequencing) will provide important constraints for better gene prediction and regulatory models. Such tools and models are essential for confident systems biology studies.

### Recommendations

The Systems Biology Knowledgebase will need to fulfill a range of requirements to achieve the research community's envisaged goals. These include:

- Providing a common mechanism for collecting, organizing, annotating, analyzing, and distributing data that enables easy data sharing and comparative analyses.
- Facilitating ***dynamic*** interconnection of data types, data sources, applications, and workflows to allow ***data integration*** for biological insight.
- Enabling researchers to identify, assess, and access all relevant ***datasets*** worldwide.
- Allowing scientists and facilities to "publish" their data, applications, and workflows into the "live data network."
- Providing space for larger-scale data integration, analysis, and publishing.
- Providing scientifically accepted rewards for researchers who "publish" well-annotated, good quality data, applications, and workflows.

We suggest that this could be achieved through the development of:

- A set of community-accepted semantic description formats (ontologies).
- A peer-to-peer based system of data, metadata, ontology, analysis tools, and workflow registration repositories that are integrated in discovery, access, and utilization through common semantics.
- Guidelines and software libraries that allow scientists and facilities to "publish" their data, applications, and workflows into the Knowledgebase in a set of agreed forms.
- A mechanism that allows scientists and facilities to easily and rapidly annotate, change, and correct research results and annotations in the Knowledgebase, capturing source, reason, quality, and proof for changes.
- User-friendly interfaces (APIs and people) to access data and application modules, ***as well as*** derived data products, enabling other users to build novel solutions with the data
- A framework of citable, unique identifiers for data, applications, workflows, and researchers
- Guidelines, training, and workshops for all new products and concepts provided by the Knowledgebase

### References for Section 4b

Anwar, N. and E. Hunt. 2009. "*Francisella tularensis novicida* Proteomic and Transcriptomic Data Integration and Annotation Based on Semantic Web Technologies," *BMC Bioinformatics* **10**(Suppl 10:S3), PMCID: 2755824. Electronic resource.

Cheung, K. H., H. R. Frost, M. S. Marshall, E. Prud'hommeaux, M. Samwald, J. Zhao, and A. Paschke. 2009. "A Journey to Semantic Web Query Federation in the Life Sciences," *BMC Bioinformatics* **10**(Suppl 10:S10). Electronic resource.

Goble, C., and R. Stevens. 2008. "State of the Nation in Data Integration for Bioinformatics," *Journal of Biomedical Informatics* **41**(5)*,* 687–93.

Wilkening, J., A. Wilke, N. Desai, and F. Meyer. 2009. "Using Clouds for Metagenomics: A Case Study," *IEEE Cluster 2009.*

### *4c. Toolkit Registry Development*

The development and management of a Systems Biology Knowledgebase will provide a unique resource to integrate experimentation, modeling, and bioinformatics across disparate levels of biological inquiry. Foremost, the successful implementation of a modular cyberinfrastructure to service the informatic needs of the systems biology community must actively recognize the broad potential user base of the resource. This recognition is critical to populate the resource with the appropriate data, tools, workflows, and corresponding literature consistent with the expectations of the user community and commensurate with the varying expertise of the Knowledgebase clientele.

Fundamental to the success of the Knowledgebase will be the bioinformatic services and resource sharing potential of the portal. At present, non-computational biologists seeking to incorporate high-level informatic investigations into their research program do not have access to a resource analogous to that which is envisioned in the development and deployment of the Knowledgebase. Most often direct consultation with expert bioinformaticians is required to initiate such investigations, effectively dissociating experimentalists with their data in the pipeline of biological discovery. To improve the efficiency of bioinformatic investigations, a centralized resource announcing the availability and utility of the various tools, applications, and algorithms available will stand as a tremendous advance toward minimizing the opacity of "–omics"-based data analysis and interpretation. To meet this need, it is recommended that a Knowledgebase-hosted toolkit registry be developed to provide a comprehensive inventory of software available to the user community. The construction of such a registry will also assist in defining the architectural strategy of the Knowledgebase engine, including compute resources, need for portage, development of APIs for web-based analysis, and demarcation of analysis subsystems specific to particular tasks and objectives.

Below is a brief list of thematic applications and additional resources that could initiate inventory within the Knowledgebase resource:

Sequence assembly

Sequence annotation

Comparative genomics

Phylogenetics

Cluster analyses

Statistical analysis

Metadata integration

Geographic Information System (GISP deployment

Data QA/QC

Transcriptomic resources

| | |
|---|---|
| Proteomic resources | Genetic database resources |
| Metabolomic resources | Metabolic subsystem database |

Strategically, each of these resources should contain a brief description of the type of data or analysis provided, its product, and rationale for why a user might be interested in using a given tool or database. Importantly, the Knowledgebase must also exist as a central forum where researchers and developers can exchange ideas to nucleate new tool development. Curation and expansion of the toolkit registry should be guided through an interactive Wiki environment where users can post comments and suggestions for including new resources into the Knowledgebase. Additionally, resource examples and workflows should be included to assist users. An exquisite example of a tool registry has been developed as part of the Neuroscience Information Framework accessible at neuinfo.org/nif_tools/nif_registry.shtm.

Ultimately, content within the Knowledgebase must be adequately indexed to allow integration across the multi-dimensions of available data. Key drivers of this need include incorporating genetic, genomic, transcriptomic, proteomic, and metabolic datasets with phylogenetic, metabolic, imaging, ecological, and geospatial information. Although such considerations are beyond the scope of the toolkit registry described here, it is critical that a composite inventory of analysis tools be available to augment discovery and seed the data integration process.

### *4d. The Knowledgebase as an Open Development Platform*

Enable tool development and integration, by providing an open developer platform inspired by the Facebook Platform / Google Apps API. Allow outside developers to produce novel analysis and visualization tools that can query the database directly (with appropriate access controls) and display and exchange results through a common UI. There will always be disagreement between research communities on which analysis is the best for any particular data type. DOE should not be in the position of enshrining one type of analysis over another. It should provide the platform, let the individual researchers develop the tools, and let the community reach a consensus.

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

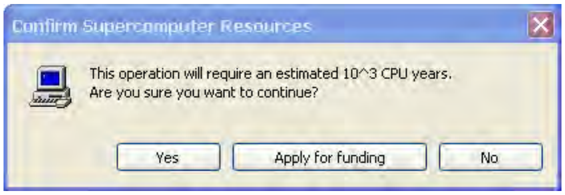
**User Empowerment and Community Collaboration.** Regardless of the size and quality of the Knowledgebase development team, there will inevitably be more developers, talent, and ideas (not to mention time to implement) "outside" than "inside." We should aim to leverage the talent within the Knowledgebase user community to develop and choose the best tools. Many novel bioinformatics tools suffer from a "failure to launch," never reaching beyond the initial journal publication, due to a lack of a web-based implementation or lack of marketing skills of the developer. By enabling tool developers to integrate their tools with the Knowledgebase platform and tie directly into its user interface, we can expose a wider variety of tools to a wider variety of users and enable more users to become tool developers themselves.

**Components, Scripts, and Open-Source Development.** Individual tools may be as simple as calculating the GC content of a DNA sequence or displaying a matrix of numbers as a heatmap. More complex tools can then be constructed by combining these elementary components, piping data from one tool to another—similar to unix shell scripts that are composed of elementary data manipulations such as “grep” and “sort,” with some control logic to pipe data between them. At the extreme end, entire processing pipelines could be encapsulated into a single tool, calling upon dozens of other analysis and visualization tools. By making these tools open source, we can enable even relatively novice programmers to tinker with and improve upon them (e.g. swapping out one statistical test for an improved one or adding a novel visualization tool).

**Reputation, Attribution, and Credit.** If we open up tool development to the world, some mechanisms are needed to enable the community to disseminate, vote, and prioritize the highest quality tools. The reputation of a tool may be based on various factors, including usage statistics (how many other tools incorporate this tool and how frequently is it actually called), direct votes by users, and the reputation of its developer. Developer reputation in turn would depend primarily on the reputation of the tools they have contributed. Community reputation may be a powerful incentive for contribution, especially for junior members. A credible mechanism for attribution and credit could potentially also be used to drive funding and even tenure decision for tool developers, on par with journal impact factors.

***Some important issues which need to be resolved:***

**Computing Resources.** Some tools can easily be run on the user’s own computer, some should be run on the server side because of higher CPU or storage requirements (e.g., BLAST against NR), others may require substantial high-performance (or cloud) computing capabilities. How do we throttle processes to achieve an equitable distribution of resources? How do we keep users from making an expensive mistake? How should we deal with a “poorly behaved” tool? Can we estimate a priori (e.g., based on previous usage statistics) how much computing power a specific tool will require? How do we fund additional computing time on this system? Can users simply buy more compute power on the cloud?



Confirm Supercomputer Resources

This operation will require an estimated 10^3 CPU years.
Are you sure you want to continue?

**Incentives for Contribution.** How do we encourage an active and vibrant developer ecosystem? Some of the larger components such as the Knowledgebase platform and early applications will need to be developed under more conventional funding models. But as the programming platforms become established, the project will need to consider funding models designed to maintain and expand innovation over the long term. Significant attention should be paid during the design phase of the Knowledgebase platform to social engineering and design of interactions between tool developers and scientists. How do scientists share and evaluate tools? How can we leverage existing networks of interaction to enhance community buy-in and involvement?

**A suitable pilot project for an open Knowledgebase development platform** might be to mirror part of an existing genome database (e.g. IMG, MicrobesOnline, SEED, etc.), implement rudimentary tool development infrastructure, make this platform accessible online, and then invite an external user to write a novel analysis or visualization tool not already found in the original database.

***Sidebar: the Facebook Platform***

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

### *4e. Institutional and Career Considerations Surrounding Open-Source Development*

This section describes some considerations with regard to institutional technology transfer philosophy and the career impact of open-source software development and use in the DOE BER Knowledgebase program.

Science assumes a clear account of methodology that is repeatable by others. Open source provides a definitive account of methods where software was used in analysis. Open source has become widely practiced in federally funded research. An Open Source Policy would be at least in keeping with the spirit of recently proposed legislation to provide free open access to all federally funded research within 6 months of publication. Open source does not have to mean immediate release. The Open Source Policy could be similar to the Data Release Policy where there is some allowance for limited access before being made public. At a glance, open source seems like an obvious choice for the Knowledgebase, but there are real issues associated with making all software immediately open source. Key issues are discussed below:

(1) Experience has shown that conflict arises between the open-source concept and the desire by home institutions to license IP in the process of facilitating technology transfer. This has in the past involved situations where the home institution licensed software to companies and

earned revenue from software developed under DOE grants. This issue can arise within universities, national laboratories, and private industry. While resolvable, this will need to be addressed. In the open-source scenario, where software is publicly available, licensing efforts are defeated, and the contractors performance and thus award fee depends partly on their success with technology transfer. This will depend in part as to whether the software has unique value or is merely routine in nature. While an open-source policy might seem in conflict with an expectation of technology transfer placed on institutions, such policy is intended to encourage rapid transfer of IP to provide a basis for new business development and benefit society. An open-source policy would accomplish that objective.

(2) The DOE Genomic Science program has a strong history of doing bioinformatics research—that is, developing new algorithms or solving the hardest computational problems in new ways. Examples include gene finders, transcription binding predictions, protein structure prediction, protein dynamics and function, metabolic pathway simulations, and large-scale cellular process modeling. To a researcher in bioinformatics, such products represent the publishable results of research, and such investigators have a right to publish their algorithms and their performance in the literature prior to open release in a manner similar to experimental scientists. As we work toward large infrastructure and in many current national laboratory SFAs (Science Focus Areas), we see the role of bioinformatics changing in part from research to that of programming and operational support, for example in building databases or websites for projects. It is clear that we have or perhaps should have two classes of bioinformatics tasks: (i) publishable research, which develops new algorithms or methods for key problems, and (ii) infrastructure support and development, which is likely to be much less publishable and where methods are more likely to be mature. The infrastructural element is analogous to core facilities for major experimental capabilities such as sequencers.

Fig. 1 illustrates the potential collaborative nature and continuum of interests and capabilities across the scientific community between the pure experimentalists and pure computationalists that become evident in the context of the Knowledgebase as a platform for future research that brings different groups and communities together. Publishable tools development is an aspect of research, while the infrastructure development is more linked with the development of the Knowledgebase itself. Infrastructure development and deployment are much more amenable to immediate open-source standards, with rewards to such individuals much less likely to be publications or novel results. A concern is that in large projects and in the push toward open source, we forget that the research mission in bioinformatics (tools development) is very important and the types of individuals that do such research are vital to the DOE Genomic Science program. There needs to be a strong research activity to generate solutions for next-generation problems in bioinformatics. We need to identify proper incentives for both paths and encourage top people in both for sustained careers. Ultimately both are required, working together, to attain the ambitious scientific objectives of the future.

### *4f. Other Potential Science Objectives and Knowledgebase Features Drawn from Responses to Preworkshop Charge Questions*

These potential science objectives were drawn from the online responses to the charge questions. This section needs further expansion and revision for the final report. The development of these lists from this and prior workshops will speed the community's process of identification and then prioritization of the set of scientific objectives that will be developed in much greater detail needed for the final workshop and ultimately for the final report—the Knowledgebase Implementation Plan.

Regarding data quality and annotation:

- Use data quality indicators.
- Use experimentally verified data only.
- Create a clean computing system using some model organisms with only experimentally verified data.
- We need the ability to update annotations in genome sequence repositories.
- Need statistical correlations of datasets.
- We need standard quantitative approaches for dealing with the different data types for normalization and assessment of statistical significance.

Regarding microbial community omics data integration, we need the following capabilities:

- To integrate short read sequence data (Illumina data) and proteome data.
- To routinely integrate all omics data for every newly sequenced organism to minimally include: RNA sequences, proteomics, metabolic phenotype (Biolog) profile. Move beyond gene-based annotations to pathways.
- Comprehensive tools that allow integrating and comparing multimolecular datasets, which are needed to fully realize the vision of microbial systems ecology.
- To integrate sequencing data and downstream analysis into a common analytic pipeline that enables end users at different skill levels to interrogate their data in interactive ways in real time. Controlled vocabularies or ontologies to leverage metadata across different organisms or samples.
- To link genotype with biogeochemistry or biogeography.
- Under defined conditions, to compare proteins expressed in related strains of bacteria to predict metabolic potential of microbial communities and resolve physiological differences; use this information to identify biomarkers diagnostic for specific biogeochemical processes. Intercompare spectral libraries to identify unique peak profiles as new gene models are added to the protein database.

## Appendix 1: Agenda

## DOE Systems Biology Knowledgebase Workshop

### Walnut Creek, California
### Tuesday, March 23, 2010

| | |
|---|---|
| 8:30 a.m. – 8:40 a.m. | Welcome, Susan Gregurick |
| 8:40 a.m. – 9:00 a.m. | “Introduction to Knowledgebase Initiative and Workshop Objectives”<br>Bob Cottingham |
| 9:00 a.m. – 9:30 a.m. | “Science Presentation on Metagenomics: Current Experience and Future Expectations for the Knowledgebase”<br>Phil Hugenholtz |
| 9:30 a.m. – 10:00 a.m. | “Science Presentation on Metagenomics: Current Experience and Future Expectations for the Knowledgebase”<br>Jill Banfield |
| 10:00 a.m. - 10:30 a.m. | Panel Summary and Audience Questions |
| 10:30 a.m. – 11:00 a.m. | Break |
| 11:00 a.m. – 12:30 p.m. | “Informatics Perspectives and Roundtable: How to Transition from the Present Towards an Open, Shared, Integrated Knowledgebase” |
| | **Discussion Leads:** Adam Arkin, Folker Meyer, Ed Uberbacher, Nikos Kyrpides, Peter Karp, Tatiana Tatusova, Victor Markowitz, Bob Cottingham |
| 12:30 p.m. – 1:00 p.m. | Working Lunch |
| 1:00 p.m. – 1:30 p.m. | “Presentation of Metagenomic Workflow Example”<br>Jill Banfield |
| 1:30 p.m. – 3:00 p.m. | Panel Discussion, Jill Banfield |
| 3:00 p.m. – 3:30 p.m. | Break |
| 3:30 p.m. – 4:30 p.m. | Panel Discussion, Jill Banfield |
| 4:30 p.m. – 5:00 p.m. | Conclusions and Adjourn, Bob Cottingham |

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

## Appendix 2: Participants and Observers

### Participants

Adams, Paul (LBNL)

Allen, Eric (University of California, San Diego

Allgaier, Martin (LBNL, JBEI)

Anderson, Gordon (PNNL)

Arkin, Adam (LBNL)

Baker, Scott (PNNL, JGI)

Banfield, Jill (University of California, Berkeley)

Benton, David (University of Wisconsin)

Bhaya, Devaki (Stanford University)

Bowen, Ben (LBNL)

Bownas, Jennifer (ORNL)

Brettin, Tom (ORNL)

Bristow, Jim (LBNL, JGI)

Broughton, Jeff (LBNL)

Canon, Shane (LBNL)

Chen, Amy (LBNL, JGI)

Chivian, Dylan (LBNL, JBEI)

Collart, Frank (ANL)

Cottingham, Bob (ORNL)

Davison, Brian (ORNL)

D'haeseleer, Patrik (LLNL)

Drell, Daniel (DOE BER)

Foust, Cheri (ORNL)

Gorton, Ian (PNNL)

Gregurick, Susan (DOE BER)

Grossman, Arthur (Stanford University)

Hallam, Steven (University of British Columbia)

Heidelberg, Karla (University of Southern California)

Hess, Matthias (JGI)

Hugenholtz, Phil (JGI)

Jansson, Janet (LBNL)

Karp, Peter (SRI International)

Kleese van Dam, Kerstin (PNNL)

Kodner, Robin (University of Washington)

Konstantinos, Mavrommatis (JGI)

Kosky, Anthony (JGI)

Kuske, Cheryl (LANL)

Kyrpides, Nikos (JGI)

Land, Miriam (ORNL)

Landick, Robert (GLBRC, University of Wisconsin-Madison)

Liolios, Konstantinos (JGI)

Lykidis, Thauos (LBNL, JGI)

Mansfield, Betty (ORNL)

Markowitz, Victor (LBNL)

McCue, Lee Ann (PNNL)

Mead, David (Lucigen Corporation)

Meyer, Folker (ANL)

Muyzer, Gerard (Delft University of Technology)

Palaniappan, Krishna (LBNL)

Pletcher, David (LBNL)

Raymond, Jason (University of California, Merced)

Richmond, Kathryn (GLBRC, University of Wisconsin)

Sczyrba, Alexander (JGI)

Slater, Steven (University of Wisconsin)

Stepanauskas, Ramunas (Bigelow Laboratory for Ocean Sciences)

Swingley, Wesley (University of California, Merced)

Szeto, Ernest (LBNL)

Tatusova, Tatiana (NIH)

Thomas, Brian (University of California, Berkeley)

Tringe, Susannah (JGI)

Uberbacher, Edward (ORNL)

Wang, Zhong (LBNL)

Woyke, Tanja (JGI)

### Acronyms

| | | | |
|---|---|---|---|
| ANL | Argonne National Laboratory | LBNL | Lawrence Berkeley National Laboratory |
| DOE | U.S. Department of Energy | NIH | National Institutes of Health |
| GLBRC | Great Lakes Bioenergy Research Center | NREL | National Renewable Energy Laboratory |
| JBEI | Joint BioEnergy Institute | ORNL | Oak Ridge National Laboratory |
| JGI | Joint Genome Institute | PNNL | Pacific Northwest National Laboratory |