DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop

This workshop was part of the U.S. Department of Energy Office of Science's 2010 Genomic Science Contractor-Grantee and Knowledgebase Workshop in Crystal City, Virginia, February 7–10, 2010.

Organized By: Susan Gregurick (U.S. Department of Energy)
Robert Cottingham (Oak Ridge National Laboratory)

Co-Chairs: Adam Arkin (Lawrence Berkeley National Laboratory; University of California, Berkeley)
Robert Kelly (North Carolina State University)

Report Contents

Section I: Knowledgebase Concept and Workshop

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

Knowledgebase Vision and Background

The emergence of systems biology as a research paradigm and approach for DOE missions has resulted in dramatic increases in data flow from new generations of experimental technologies in areas such as genomics and imaging. While some resource centers are generating large datasets with workflows designed to answer specific scientific questions, there is also a great increase in data production, generally from individual laboratories. New scientific questions arise and can be answered by combining and analyzing such data across laboratories and projects. Great value has derived from the ability to combine sequence and structure data across producers, and in some research communities, such as the yeast field, general access to functional genomic data has greatly accelerated discovery and technology development. Over the last decade, BER—through its Genomic Science program—has sought to solve bioenergy, environmental remediation, and carbon sequestration challenges that demand understanding biological activities exhibited by complex populations and the individuals within them. Since we seek to understand the molecular basis of these dynamics and activities on scales from individual genomes through cellular networks to community function and evolution, these projects are generating multiscale information that could be organized more effectively to aid

the science of individual projects and to synergize data across projects with related missions. Perhaps even more important, the data from multiple, possibly unrelated programs could be flexibly reorganized and analyzed to aid new scientific discoveries and provide insight to researchers in environmental microbiology and biotechnology generally.

Enabling the community to serve, query, combine, and analyze these diverse data types is therefore imperative, as is building a blueprint and system to enable the design, implementation, and use of new analytical tools and frameworks for working with such data. To manage and effectively use this ever-increasing volume and diversity of data, the Genomic Science program is developing the DOE Systems Biology Knowledgebase—an open, community-driven cyberinfrastructure for sharing and integrating data, analytical software, and computational modeling tools. Historically, most bioinformatics efforts have been developed in isolation by people working on individual projects, resulting in isolated data and methods. An integrated, community-oriented informatics resource such as the Knowledgebase would provide a broader and more powerful tool for conducting systems biology research relevant to BER's complex, multidisciplinary challenges in energy and environment. It also would be easily and widely applicable to all systems biology research.

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

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

Beyond these perspectives from the last report, the Knowledgebase is now envisioned as a robust, flexible, and well-documented open architecture. This architecture would allow for both organized and distributed community development, facilitate the sharing of data and tools for data transfer, integration, query, analysis, and visualization, and be committed to interoperating with community resources and standards.

The Knowledgebase would differ from current informatics efforts by integrating data and information across projects and laboratories—tracking diverse, multiscale biological data from the genome through molecular networks, to cellular populations and communities, to environmental function, and combining data centralization with distributed data. Integration

implies that the Knowledgebase be a community effort rather than a monolithic project overseen and contributed to by only a few people. The Knowledgebase also will need to be more standardized than today's informatics resources. Although standardized components may not always be "cutting edge," they will be more interoperable, enabling comparisons among different laboratories and thus yielding important new insights. Standardization will involve not only data but also experimental protocols.

Another fundamental feature is that Knowledgebase development will have a more mature software engineering approach. In the past, biologists not necessarily trained in state-of-the-art computational techniques reasonably applied the computational tools of the day to their research. However, the dramatic increase in the amount of sequencing and other data being generated requires the support of a more robust computational infrastructure, with analyses that no longer are carried out in an *ad hoc* manner. Many current development efforts are based on computational technologies created 10 to 15 years ago. More modern analytical technologies are needed. To be useful, these new techniques must be developed by the entire research community rather than by informatics specialists working in isolation.

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

Several existing systems and applications can serve as reference models for thinking about Knowledgebase development. Exemplifying the concept of an open-source development is the computer operating system Linux, which is being built by a community of software developers working collaboratively to create a sophisticated and fairly successful system. Other familiar examples include iPhone or Google apps that enable users to pick and choose the kinds of features and capabilities they want and integrate them into a phone or other device. We are familiar with user interfaces that show layering of data from Google maps and Google Earth annotations (e.g., locations of landmarks and restaurants). Experimental design and research in the future will be conducted in the context of a user model similar to these successful systems. As research users gain new insights in systems biology from experiments and analyses, their interaction with the Knowledgebase populates new detail in the biological systems, forming the basis for new referential insight.

Wikipedia development also is open source and open development, allowing individuals or groups to contribute content. It has an editorial model, and, over time, the quality of its content evolves and improves. For the Knowledgebase, such an open-development environment

conceivably would enable noncomputing experts to play a role in the project's development and evolution.

Although these historical examples are approximations of the Knowledgebase vision, they provide a notion of possibilities in their commonly understood characteristics of flexible community development, data layering, editorial control, and peer review integration. The take-away lesson is that we see the initial Knowledgebase development like an operating system kernel that provides a platform on which open contribution of new applications can occur while the Knowledgebase simultaneously is managed to provide core functions like protection of legacy data and development of the underlying access and sharing model and architectural methods.

Workshop Description, Goals, Inputs, and Outputs

Although the 2009 Knowledgebase report describes a vision and long-term objectives for the Knowledgebase, it does not provide details about a plan to implement the system. To that end, DOE has launched an R&D project to establish the requirements for the Knowledgebase and to outline a plan for implementing them. As part of this project, DOE is sponsoring a series of community workshops. The first—held in conjunction with the November 2009 Supercomputing conference in Portland, Oregon—explored the potential for applying the cloud computing approach to systems biology research. The second workshop—held prior to the January 2010 Plant and Animal Genome meeting—addressed the Knowledgebase requirements necessary for developing data capabilities for plants. The output for these and subsequent workshops is now or will soon be posted online at www.systemsbiologyknowledgebase.org/workshops. As the third event in this series, the DOE Genomic Science Microbial Systems Biology Knowledgebase workshop was held Feb. 9–10, 2010, during the DOE Genomic Science Contractor-Grantee meeting in Crystal City, Virginia.

The goals of this workshop were to outline the near-, mid-, and long-term trajectory of microbial sciences for energy and environment and to map the associated workflows and data integration methods that can inform Knowledgebase specifications and requirements. Participants were asked to provide responses to six charge questions:

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

The workshop featured presentations discussing the current, near-, and long-term prospects for microbial systems biology research in the context of the Knowledgebase. Formal presentations were given by Robert Cottingham (Oak Ridge National Laboratory) describing Knowledgebase background and objectives, by Robert Kelly (North Carolina State University) on the "Near-Term Prospects for Functional Microbial Genomics: Moving Beyond the Monoculture Paradigm," and by Adam Arkin (University of California, Berkeley, and Lawrence Berkeley National Laboratory) on "From Pathways to Populations and Back Again: Long-Term Prospects for the Microbial Systems Biology Knowledgebase."

Kelly indicated the rapidity with which new genome sequence information appears in public databases is presenting a growing challenge for the data storage, analysis, and utilization necessary to foster scientific and technological advances. The systems biology framework has arisen in response to this challenge, but new computing strategies are needed to take advantage of this new context for examining microbial biology.

Kelly also pointed out that most of what is now known about microbial biology was learned from the study of pure laboratory cultures. The "monoculture" paradigm has been quite productive and will continue to be at the heart of microbiology. However, monocultures are not representative of how microbial systems exist in nature. To this end, metagenomics has provided a means for examining microbial complexity, but complementary functional information is still needed to understand the "metaphenotype."

Illustrating the need for microbial community studies is the hypothesis that a significant portion of every microbial genome encodes elements designed to regulate and mediate intercellular interactions. These elements may not be responsive in laboratory monocultures and may be triggered only by certain environmental and ecological stimuli. Do these genomic elements exist? What are the studies needed to make this determination? If these genomic elements exist, how can they be identified, characterized, and manipulated? If multispecies systems are to be examined via systems biology, what are the consequences in terms of experimental design and analysis? What is the best way to construct a systems biology knowledgebase for multispecies (multiphenotype) investigations?

Over the next several years, efforts are needed to link the complexity reflected in metagenomes to what is already known from monoculture studies. Kelly indicated this learning curve will necessarily start with relatively simple systems because even co-cultures can exhibit phenotypes not easily predicted from pure culture information. Extending functional microbial genomics beyond monocultures was discussed with a view toward the integration of experimental design, experimental methods, and data analysis strategies. Kelly used hyperthermophile communities to illustrate some challenges that arise when moving beyond monocultures.

In his presentation, Arkin indicated the grand challenge to predict phenotype from genotype is particularly difficult in the microbial world. At its core, this challenge seeks to understand the principles of biological architecture and function sufficient for predicting behavior and, of course, for changing it. A systems biology knowledgebase should grow into an indispensible tool for molecular, environmental, evolutionary, medical, and epidemiological microbiologists

and for biotechnologists to understand and engineer their systems. However, there are challenges in accomplishing this that are found in few other systems.

Microbes rarely work alone but operate in complex communities that form spatial and temporal webs of mutual support, parasitism, and predation. Perhaps unique to microbes and their communities are the astonishingly rapid mechanisms for evolution and the deeply intertwined ecology of mobile genetic elements that aid in the preservation, diversification, and dissemination of function and may be central drivers themselves of the architecture of microbial networks.

The Knowledgebase, in the long term, will be faced with capturing and interrelating data about all these processes at scales from molecules to meters. Sequencing technologies reveal information on the identities of microbial players in these communities and can hone in on some aspects of gene expression. Structural techniques can provide key information on molecular identity and sometimes function. New imaging technologies can give us information on the arrangements and interactions among molecules, cells, and their environment. However, the complexity of the data increases greatly when moving beyond the sequence of single genomes and crystal structures of single proteins. The data also become far more conditional on unmeasured conditions and interactions and less precise and accurate metrologically, all of which present challenges for organizing and navigating this information. Arkin presented an example process outlining how such information could be assembled, navigated, and used in a knowledgebase. At each level, the challenges and acuteness of need for the community were described.

In ensuing discussions at the workshop, emphasis was placed on establishing agreed-upon scientific objectives that will result in a successful, community-driven Knowledgebase. To build a system that helps achieve important scientific goals, informatics experts need input from and frequent dialogue with the research community on what these goals are, including how the research technologies, data types and quantities, and goals change over time (see Fig. 1.1. Knowledgebase R&D Project).

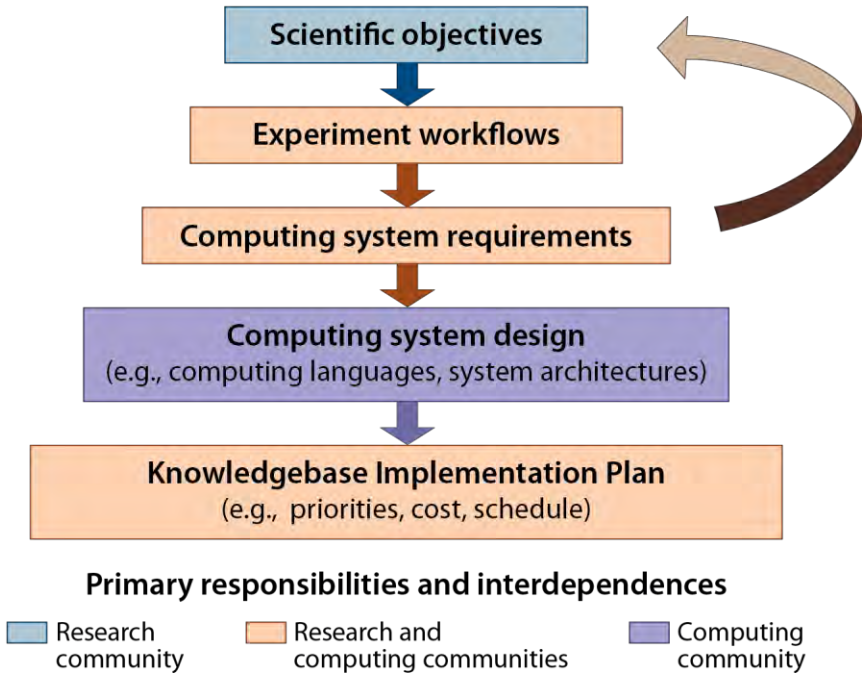


Fig. 1.1. Knowledgebase R&D Project: Scientific Objectives, Intense Collaboration Critical to Successful Knowledgebase Implementation Plan. The final product of this project, the Knowledgebase Implementation Plan, is being developed to incorporate the components and functionality necessary for the systems biology research community to meet its defined scientific objectives. To do this, the research and computing communities must work closely together to define—realistically and at a significant level of detail—the scientific objectives and experiment workflows (protocols) necessary for defining computing system requirements and design and for completing the implementation plan for a robust, durable Knowledgebase.

Workshops, such as this one, provide opportunities to discuss and identify appropriate and community-generated scientific objectives. Any and all input was welcomed, and participants were encouraged to contribute to the final R&D report at www.systemsbiologyknowledgebase.org. To be effective, scientific objectives must be credible, impactful, and achievable in a few years. Participants were asked to discuss objectives based on current research activities and consider candidates and priorities to recommend.

Several examples of potential scientific objectives related to microbes were presented to stimulate discussion. The first was improved prediction of gene regulatory networks based on integrating genomic sequences from phylogenetically related organisms with high-resolution expression (RNA-Seq) data from multiple biological states. Suppose the goal was to predict gene regulation in a particular situation. What are the Knowledgebase capabilities necessary for predicting gene regulation in a subsystem? One need would be the ability to upload raw RNA-Seq sequence data or provide access it. Another need would be tools to process raw sequence into standard formats. A third involves data visualization capabilities.

The limit in the future might be how many biological samples are available to be assayed by RNA-Seq and not the availability or cost of the technique. As cost rapidly declines, it is conceivable that thousands of states could be measured. From plots of expression profiles, genes that are statistically represented in a particular biological state can be readily visualized.

Which genes are in probable states or pathways that are of particular research interest? This should all be readily available to any researcher. Even small regulatory sequences are visible. Imagine doing RNA-Seq analysis on a set of phylogenetically related organisms and comparing them based on genomic structure. This new data will reinforce past experiments in these same organisms. Based on the alignments, we can find promoters and make predictions about gene regulatory binding sites. This illustrates the type of understanding achievable with a knowledgebase characterized by good algorithms and data integration technologies that have been built up over time. When using Google maps to find the nearest Starbucks location, users rely on a series of technologies that have been developed in such a way. A set of standards allows this information to be mapped together, enabling the system to generate the appropriate directions. The data integration underlying Google maps is analogous to many of the current challenges associated with integrating biological data.

A second example of a scientific objective would be integrating phenotypic response with specific genotypes or pathways so that regulatory or genetic changes could be predictably associated with microbial behavior and response. The idea of relating phenotypes to genotypes and putting that information in context is of wide interest. What are the sources of data? How do we transform them? What are the analytical steps, and what tools are currently available?

As with any scientific objective considered for the Knowledgebase, these two examples would be evaluated to determine if they are credible, impactful, and achievable. If a particular objective meets these three criteria, then community input would help set priorities for the development and implementation timeline of the Knowledgebase.

Section II: Workflows—Knowledgebase Use Cases

Workflows as a Bridge from Bench to Computer

The focus of this workshop, particularly on the second day, was on creating workflows. In research, a scientific objective is satisfied by creating hypotheses and doing one or more experiments depending on the scope of the objective. For every experiment, there are rationales, protocols to be executed, a number of data inputs (data sources) and outputs (results), and analysis tools. Workflows describe this information. Detailed workflows are bridges between the research and computing communities and thus are key to translating research into computing requirements that will most effectively advance the science.

Workflows provide important details for Knowledgebase design, both in terms of the underlying data as well as the experimental or analytical objective. Knowledgebase architecture will have layers such as data repositories, workflow management, and output visualization, all of which relate to workflows developed by the scientific community participating in this Knowledgebase development process. Workflows are essentially communication mechanisms that exchange ideas and information between the researchers and those who actually build the computing system. Included in this report are six workflows drafted to satisfy research objectives important in DOE systems biology. These workflows encompass diverse problem-solving methodologies representative of the broad scientific community and are works in progress—presented here to stimulate discussions between the research and bioinformatics

communities so that robust computing system requirements and an implementation plan can be developed.

Developing an executable Knowledgebase Implementation Plan must be a community effort—involving both the experimental and computing research communities—where we integrate across projects and research laboratories. Fully developed, robust workflows will foster this integration and lead to a more standardized approach. To handle a new level of biological complexity, we need to embrace more strategic software engineering approaches; we can no longer afford to create isolated and ad hoc systems.

As the key products of this workshop, workflows are critical inputs to the participants of the final workshop (June 1–2, 2010, Crystal City, Virginia). Prior to and during the final workshop, representatives from the computing and biological research communities will work closely together to refine the scientific objectives and workflows and to translate the workflows into computing system requirements. These requirements will form the basis of the Knowledgebase design—a prerequisite to the Knowledgebase Implementation Plan, which is the final product of the DOE Systems Biology Knowledgebase Research and Development Project.

The workflows described in this section are critical to the success of systems biology research and reflect the data inputs, outputs, and experiments being carried out in the DOE-sponsored research community. Over the next several months, assessments will be made to ensure that the highest priority workflows, as identified by community consensus, will be included in the Knowledgebase Implementation Plan. The workflows generated in this workshop are:

1. Metabolic Network Reconstruction (Ines Thiele)
2. Metabolic Flux Analysis via Isotope Labeling (Hector Garcia Martin)
3. Inference of Gene Regulatory Networks (Adam Arkin and Nitin Baliga)
4. Signaling (Aindrila Mukhopadhyay and Loren Hauser)
5. Structural Biology (Paul Adams)
6. Imaging Bioinformatics (Bahram Parvin)

To foster further interactions among the biology research communities, both experimental and computational, most of these have been included as originally submitted as a snapshot in time showing the current range of thought on what a workflow is and how the concept relates to various researchers and areas of research. [Note: An additional workflow on microbial community science is under development and will be available in May. This workflow is based on discussions from the Knowledgebase workshop held in conjunction with the DOE Joint Genome Institute's annual user meeting (March 23, 2010). Workflows associated with the microbial community scientific objectives will be discussed by the interdisciplinary participants at the June Knowledgebase workshop where Knowledgebase system requirements will be discussed and drafted.]

To facilitate workflow development, participants in this workshop were instructed to focus on describing several workflow components: data and sources (inputs), process steps (transformation rules or algorithms), and results or output. They also were asked to explain why

the workflow is important to the research endeavor and how it might be improved. As an example, consider the first workflow on Metabolic Network Reconstruction starting on p. 11. This example lists input data and even outlines how to obtain it. The process diagram associated with this workflow identifies each process step (see Figure 1. Detailed Workflow for Metabolic Network Reconstructions, p. 12). Many of these steps are common bioinformatic transformations that could readily be included in a future Knowledgebase. As the authors note, many steps are not precise and require some type of manual intervention such as curation. This identifies areas for improvement in either the underlying data or a need for better standards. Some of the process steps are experimental and produce specific results. Again, issues of data quality and accuracy can be important. Although not entirely automatable, this process is of wide utility and interest. This presents an excellent example of a workflow that the research community could prioritize to focus on in the Knowledgebase. By having a range of researchers focused on the bottlenecks, there would likely be improvements not only for metabolic reconstruction, but for other areas of research that depend on similar process steps.

Workflow 1: Metabolic Network Reconstruction

Summary

The metabolism workflow consists of two parts:

1. The metabolic network reconstruction protocol [1] and required data and
2. The protocol to obtain fluxomic data required by the metabolic network reconstruction protocol.

Genome-scale metabolic network reconstructions are biochemically, genetically, and genomically (BiGG) structured knowledgebases, the goal of which is to formally represent the metabolic activities of a specific organism. Genome-scale metabolic networks have been published for more than 30 organisms to date, though they are of varying quality and completeness. Reconstructions are useful because they can be mathematically converted into constraint-based models, allowing important predictive calculations like flux balance analysis to be performed. This comprehensive workflow details nearly 100 iterative steps in the following categories:

1. Draft reconstruction
2. Refinement of reconstruction
3. Conversion of reconstruction into computable formats
4. Network evaluation
5. Data assembly and dissemination

The output of this workflow is a highly curated, accurate, and comprehensive representation of biochemical transformation taking place in the organism of interest. It is not yet possible to automate all steps within the process without loss of accuracy or correctness.

We also attached the comprehensive standard operating procedure (SOP) for biochemical network reconstruction [1] to this workflow.

Input

Required organism-specific data

- Gene information (ID, coordinates, function)
- Protein information (function, location, complex formation)
- Enzymatic reaction (stoichiometry at cellular pH, substrate specificity, cofactor specificity, location, directionality)
- Biomass composition (fraction of macromolecule, molecular composition of macromolecules)
- Phenotyping data (growth medium composition, other growth conditions – e.g. temperature, pH, etc)
- Knock-out strain information (growth phenotypes, other characteristics)

- P/O ratio

How to obtain this information

Online resources:

- Genome database containing genome annotation (i.e., locus ID, gene coordinates, (putative) annotation) – e.g., GOLD, TIGR, SEED, etc.
- Biochemical reaction database for metabolic reactions – e.g., KEGG, BRENDA
- Transport database for transport reaction mechanisms – e.g., Transport DB
- Organism-specific databases – e.g., EcoCy, PyloriGene, GeneCards
- Protein location prediction – e.g., PSORT, PA-SUB
- Thermodynamic information (estimation of standard Gibbs free energy of formation ($\Delta_f G'°$) and of reaction ($\Delta_r G'°$)) – e.g., Web GCM
- CMR database (estimation of DNA, RNA and protein composition)

Tools:

- Blast (if not or insufficient genome annotation is available), other gene function annotation tools

Bibliome:

- Primary and review literature about organism, its metabolic characteristics and its components (proteins, genes)
- Biochemical textbooks
- Organism-specific books

Experiments:

- Measurement of biomass composition (lipids, amino acids, nucleotides, cofactors, etc.)
- Measurement of growth environments (e.g., biolog)
- Measurement of single and double knockout mutants
- Measurement of possible secretion products (and ratios) at different growth environments
- Omics data: Metabolomics, fluxomics, proteomics, transcriptomics
- Transcriptional regulatory information – which pathways are active under which conditions

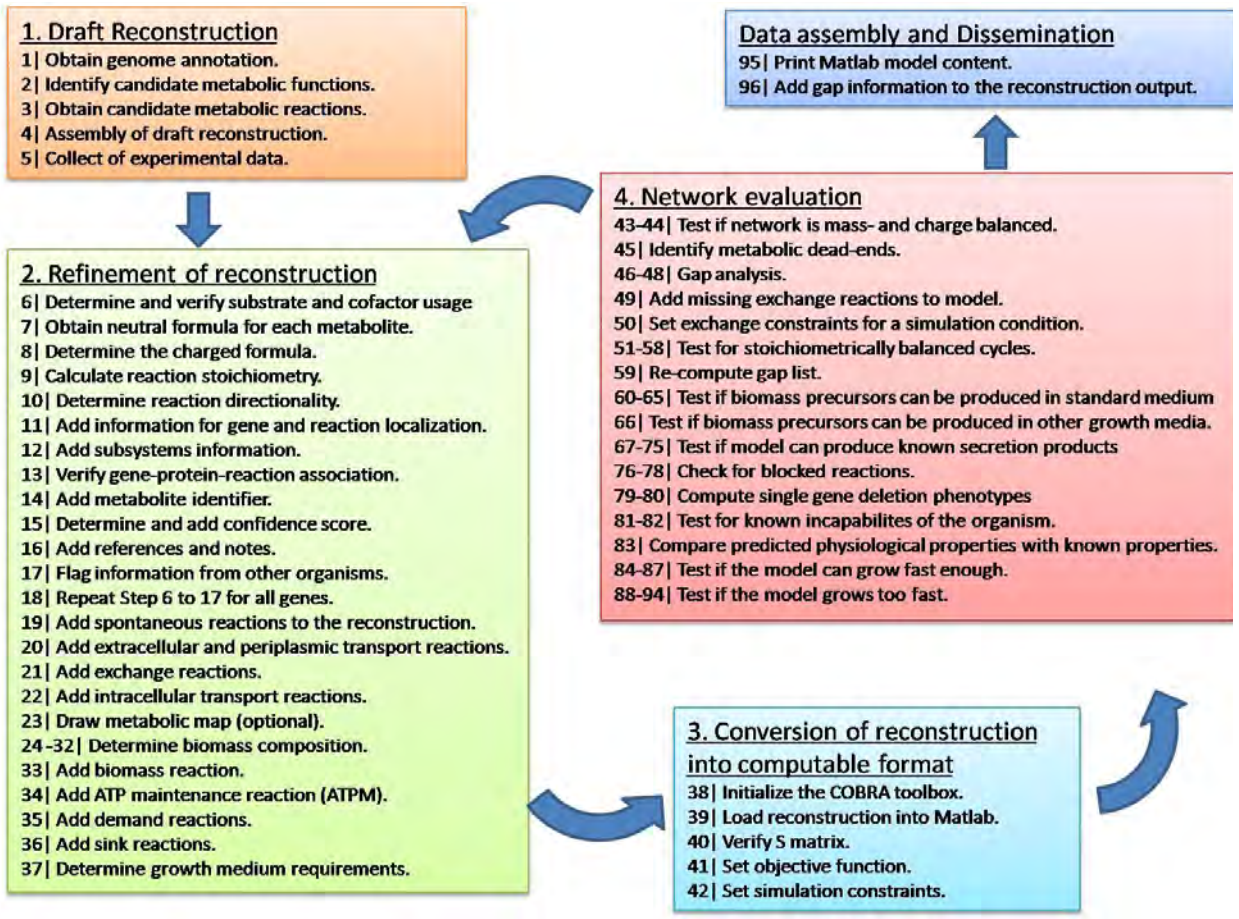


Figure 1. Detailed Workflow for Metabolic Network Reconstructions.

Workflow Process to Metabolic Network Reconstruction

The biochemical network reconstruction process is well established for metabolism and has been applied to many model organisms. The same approach can also be applied for other cellular functions, such as signaling [2, 3] and macromolecular synthesis [4]. The reconstruction process has been reviewed by numerous groups [5-8]. More recently, it has been formulated in the form of a standard operating procedure (SOP), or protocol, which explains the necessary stages and steps in great details [1]. Readers interested in reconstruction are advised to also refer to the SOP.

The metabolic reconstruction process can be grouped into 5 major stages (see Figure 1):

1. **Generation of a draft reconstruction based on genome annotation and biochemical databases.** Generally, the genome annotation is downloaded from a repository (e.g., NCBI) or the sequencing center (e.g., TIGR), and it should list at least a unique identifier, genome coordinates, and potential gene product function. Many of genome resources have also enzyme commission (EC) numbers for the genome encoded enzymes. These EC numbers along with key words can be used to compile a sublist of potential metabolic functions in the target organism. This list can be then used to obtain from

biochemical databases (e.g., KEGG [9]. BRENDA [10]) the metabolic reactions catalyzed out by the enzymes. This list represents the draft reconstruction. The characteristics of this draft reconstruction are that it is incomplete (missing or wrong annotations) and it has an organism-independent reaction list: KEGG, as well as partially BRENDA, list all possible metabolic transformation catalyzed by a particular enzyme. However, the enzyme of the target organism may be able to bind to a subset of the listed substrates, or only one of the listed coenzymes can participate in the reaction in the target organism. This substrate and coenzyme pluripotency is one of the main reasons why manual curation is necessary.

2. **Refinement and expansion of the draft reconstruction through manual curation and extensive use of biochemical literature specific for the target organism.** Starting from the draft reconstruction, every entry will be evaluated for the following criteria:

   a. Is the assigned function of the gene product correct? Use of biochemical literature, enzyme purification studies, a more detailed, phylogeny based annotation are helpful to answer this question.

   b. What is the substrate and coenzyme specificity of the target organism's enzyme? Use of biochemical data, enzyme assays and protein structure will be helpful for answering this question. Finding evidence for this issue can be difficult. The use of closed relative organisms can be helpful.

   c. Is the biochemical reaction(s) mass- and charge balanced? Therefore, the neutral formula of each metabolite in the reaction has to be obtained (e.g., from KEGG or PubChem [11]). The charged formula has to be determined for each metabolite for a set pH value (e.g., pH 7.2) by determining the protonation state of each functional group within the metabolite. Software tools are available to assist this step (see Thiele and Palsson for details [1]). Once the charged formula has been determined for each metabolite, the occurrences of each element (e.g., C, H, N, S, O, P), as well as the charge, on the left- and right-hand side has to be counted. Stoichiometric coefficients may need to be adjusted such that the same amount of each element appears on both sides of the reaction. In some cases, protons (H+) or water may be added to the reactions to obtain a mass- and charge balanced reaction.

   d. The reaction directionality needs to be determined using thermodynamic information (refer for details to Thiele and Palsson [1], Feist et al [12], and Fleming et al [13]).

   e. Localization of reaction needs to be determined, especially, if multiple compartments are considered (e.g., human metabolic network accounts for eight cellular compartments, while many bacterial reconstructions account for two or three compartments, which are extracellular space, periplasm, and cytosol). Information about reaction location may be obtained from the genome sequence if it encodes for a signal peptide (for protein export) or by targeted experiments (e.g., using GFP tagging and fluorescence microscopy).

f. Gene-protein-reaction (GPR) association needs to be determined: while the genome annotation indicates that the gene product has a particular function, one should investigate if further gene products are needed for function, as is the case for protein complexes, or if alternate gene products exist that can carry out the function, i.e., isozymes. The reconstruction contains these GPR associations in form of Boolean rules: for example, a protein complex is encoded as 'gene_A & gene_B', while isozymes are encoded as 'gene_A or gene_B'. Any combination of these rules is possible. Beside genome annotation, biochemical data, protein purification, and/or structural genomics can provide information regarding the GPR association.

g. Confidence score, references, and notes: The steps listed above collect valuable information for a particular enzyme or function in the target organism. This information should be associated with the network reaction (e.g., in special columns in the spreadsheet). This information is thought to increase the traceability of reaction/gene evidence as well as highlight/summarize the amount of knowledge currently available. Often, a confidence scoring system is employed, which allows easy identification of high-confidence/low confidence reactions in the network. This is of particular value during the network debugging and evaluation stage (see below). The highest confidence score (4) is given to reactions that have biochemical evidence (e.g., protein purification, protein assays, protein structure information). A score of 3 is given if genetic data is available (e.g., knock-out mutant characterization, knock-in experiments, over-expression of a protein). A score of 2 is given if either physiological data (e.g., secretion products, growth capability on substrate) or (high confidence) sequence annotation is available. A low confidence score of 1 is given if reactions are included for modeling purposes without any of the aforementioned evidence. In some cases, a confidence score of zero is also employed, which highlights reactions that have not yet been evaluated for supporting evidence.

h. Finally, different information should be collected in this stage of the reconstruction process to facilitate the following stages. This information includes the biomass precursors, necessary to produce a new cell (target organism) which is ideally derived from experimental data (see Thiele and Palsson for a detailed description on how to compile this information). Furthermore, information about enzyme reaction rates (v_{max}) should be collected, as many biochemical publications contain this information. Information about growth media should be also collected.

3. **Conversion of the manual curated metabolic reconstruction into a mathematical model.** The reconstruction process is an iterative process as shown in Figure 1, where the initial reconstruction is converted into a mathematical format, the so called stoichiometric (S) matrix. This model conversion also includes the addition of balances and bounds. Balances in biochemical networks can be, for example, mass- and energy conservation. For instance, the majority of modeling applications of metabolic models assume the system to be in quasi steady state. This assumption implies that the sum of producing reactions for a particular metabolite is equal to the sum of consuming reactions. Bounds on metabolic reactions can include maximal reaction rates based on

the catalyzing enzyme's properties, thermodynamic information (e.g., reaction directionalities), etc. Often, the mathematical models are stored and computed in Matlab (Mathwork, Inc). Commonly, metabolic reconstructions and models are stored in the systems biology markup-language (SBML) format [14], which is platform independent and can be loaded in numerous systems biology applications.

4. **Network debugging and evaluation to ensure that the metabolic model has similar phenotypic properties as the target organism.** Once the metabolic reconstruction is converted into a mathematical format and balances and bounds are applied, a comprehensive investigation of the model's properties begins. Most reconstructions contain initially numerous dead-end metabolites (i.e., metabolites that are only produced or consumed in the network). Due to the balance constraints, reactions which contain such dead-end metabolites cannot carry any reaction flux in any simulation conditions. A detailed evaluation of these dead-end metabolites is necessary to identify whether these metabolites can be connected to the remaining network by adding one or more reactions to the reconstruction. However, one has to be careful, as arbitrary filling of the so-called gaps will alter significantly the model's properties. All added reactions should have experimental, genome and/or physiological data as supporting evidence. Some dead-end metabolites may remain in the network, as current knowledge does not support any filling of gaps they are causing. In addition to these 'knowledge gaps' the reconstruction can contain 'scope gaps.' In the case of scope gaps, reactions are known, which could connect the dead-end metabolite, but they are either non-metabolic or not within a previously defined scope of the reconstruction (e.g., tRNA charging with amino acids).

   Once all dead-end metabolites have been characterized and partially connected to the network by repeating part of the second and third stage, the model's capability to produce biomass precursor is evaluated. This process will lead to further identification of network gaps, which need to be filled. This step can be quite time-consuming, and detailed evaluation of dead-end metabolites in the earlier step will directly pay off. When the model can produce all biomass precursors, one can compile them into one reaction (the biomass reaction) by considering their fractional contributions to cell composition. This stage also includes further (i) quality tests, such as the model's capability to grow on known carbon, nitrogen, phosphor and sulfur sources; (ii) the capability to reproduce accurately measured growth rates and to secrete known by-products. The list of tests depends on the properties of the target organism as well as the availability of experimental data (e.g., phenotyping data, knock-out mutant growth phenotype data, etc.). Note that this stage is iterative, in which network reactions will be added (by repeating partially, or in full, stage 2 and 3) or in some cases reactions will be removed from the metabolic reconstruction. This stage is deemed to be finished if the model reproduces accurately the target organism's phenotypic characteristics and/or experimental data is exhausted.

5. **Prospective use of the reconstruction and the metabolic models. This stage is certainly the most exciting part of the reconstruction process.** Numerous applications have been developed over last decade or so, including biological discovery [15], metabolic

engineering [16-17], prediction of outcome of adaptive evolution [18], network topology [19], and assessment of phenotypic behavior [20-22]. Some of these applications have been summarized in a recent review [23-24].

Output

The output of this workflow is a highly curated, accurate and comprehensive representation of biochemical transformation taking place in the organism of interest (Figure 2). Note that to date, it is not possible to automate all steps within the 5 stages without loss of accuracy or correctness.

A

Col 1	*Col 2*	*Col 3*	*Col 4*	*Col 5*	*Col 6*	*Col 7*
Rxn name	**Rxn description**	**Formula**	**GPR**	**Genes**	**Subsystem**	**Reversible**
PFK	phosphofructokinase	atp[c] + f6p[c] -> adp[c] + fdp[c] + h[c]	(b3916 or b1723)	b1723 b3916	Glycolysis/Gluconeogenesis	0

Col 8	*Col 9*	*Col 10*	*Col 11*	*Col 12*	*Col 13*	*Col 14*	*Col 15*
LB	**UB**	**Objective**	**CS**	**E.C. number**	**rxnKeggID**	**Notes**	**References**
0	1000	0	4	2.7.1.11	R00756	E. coli has to genes for PFK (pfkA and pfkB) where pfkA is the major form.	PMID: 6310120;149128;6310120

B

Col 1	*Col 2*	*Col 3*	*Col 4*	*Col 5*	*Col 6*	*Col 7*	*Col 8*	*Col 9*	*Col 10*
Metabolite name	**Description**	**Formula**	**Charge**	**Compartment**	**KeggID**	**PubChemID**	**CheBI ID**	**Smiles**	**InChi**
glc-D[c]	D-Glucose	C6H12O6	0	Cytoplasm	C00031	3333	17634	OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O	1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6?/m1/s1

Figure 13:
Figure 2. Data contained in the metabolic network reconstruction after completion of the presented workflow.

References

1. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction.** *Nature protocols* 2010, **5**(1):93-121.
2. Papin JA, Palsson BO: **The JAK-STAT Signaling Network in the Human B-Cell: An Extreme Signaling Pathway Analysis**. *Biophysical journal* 2004, **87**(1):37-46.
3. Li F, Thiele I, Jamshidi N, Palsson BO: **Identification of potential pathway mediation targets in Toll-like receptor signaling**. *PLoS Comput Biol* 2009, **5**(2):e1000292.

4. Thiele I, Jamshidi N, Fleming RM, Palsson BO: **Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization**. *PLoS Comput Biol* 2009, **5**(3):e1000312.
5. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO: **Reconstruction of biochemical networks in microorganisms**. *Nature reviews* 2009, **7**(2):129-143.
6. Reed JL, Famili I, Thiele I, Palsson BO: **Towards multidimensional genome annotation**. *Nature reviews* 2006, **7**(2):130-141.
7. Notebaart RA, van Enckevort FH, Francke C, Siezen RJ, Teusink B: **Accelerating the reconstruction of genome-scale metabolic networks**. *BMC Bioinformatics* 2006, **7**(1):296.
8. Durot M, Bourguignon PY, Schachter V: **Genome-scale models of bacterial metabolism: reconstruction and applications**. *FEMS microbiology reviews* 2009, **33**(1):164-190.
9. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic Acids Res* 2006, **34**(Database issue):D354-357.
10. Barthelmes J, Ebeling C, Chang A, Schomburg I, Schomburg D: **BRENDA, AMENDA and FRENDA: the enzyme information system in 2007**. *Nucleic Acids Res* 2007, **35**(Database issue):D511-514.
11. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2008, **36**(Database issue):D13-21.
12. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Molecular systems biology* 2007, **3**:121.
13. Fleming RM, Thiele I, Nasheuer HP: **Quantitative assignment of reaction directionality in constraint-based models of metabolism: Application to Escherichia coli**. *Biophys Chem* 2009.
14. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A *et al*: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics (Oxford, England)* 2003, **19**(4):524-531.
15. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO: **Systems approach to refining genome annotation.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(46):17480-17484.
16. Lee SY, Kim JM, Song H, Lee JW, Kim TY, Jang YS: **From genome sequence to integrated bioprocess for succinic acid production by Mannheimia succiniciproducens**. *Appl Microbiol Biotechnol* 2008, **79**(1):11-22.

17. Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO: ***In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid**. *Biotechnology and bioengineering* 2005, **91**(5):643-648.

18. Ibarra RU, Edwards JS, Palsson BO: ***Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth**. *Nature* 2002, **420**(6912):186-189.

19. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL: **Global organization of metabolic fluxes in the bacterium *Escherichia coli***. *Nature* 2004, **427**(6977):839-843.

20. Thiele I, Price ND, Vo TD, Palsson BO: **Candidate metabolic network states in human mitochondria: Impact of diabetes, ischemia, and diet**. *J Biol Chem* 2005, **280**(12):11683-11695.

21. Reed JL, Palsson BO: **Genome-Scale In Silico Models of E. coli Have Multiple Equivalent Phenotypic States: Assessment of Correlated Reaction Subsets That Comprise Network States**. *Genome Res* 2004, **14**(9):1797-1805.

22. Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ, Smid EJ: **Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model**. *J Biol Chem* 2006, **281**(52):40041-40048.

23. Feist AM, Palsson BO: **The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli***. *Nat Biotech* 2008, **26**(6):659-667.

24. Oberhardt MA, Palsson BO, Papin JA: **Applications of genome-scale metabolic reconstructions**. *Molecular systems biology* 2009, **5**:320.

Workflow 2: Metabolic Flux Analysis via Isotope Labeling

Summary: Metabolic fluxes are a key determinant of cellular physiology, representing the final functional output of the interaction of all the molecular machinery (genes, proteins, metabolites) studied by the other "omics" fields. This workflow (a schematic of which is presented in Figure 1) describes the input data required for measuring fluxes using an isotope labeled feed, along with the expected output and the processes needed to obtain it. The main input data are metabolite labeling patterns after a carbon labeling experiment, a metabolic reconstruction, and measured extracellular and biomass fluxes. The desired output is the rate (i.e., number of molecules through the reaction) for each of the reactions considered in the model, along with confidence intervals. Here, we will focus on the most common and well-established form of flux analysis through isotope labeling: ^{13}C Metabolic Flux Analysis (^{13}C MFA) from proteogenic amino acids in the exponential phase. Nonetheless, the modular nature of the workflow presented here will allow for other varieties of ^{13}C MFA in development to be easily incorporated.

Figure 1. Workflow for ^{13}C metabolic flux analysis. Blue blocks indicate processes (e.g., experiments or algorithms), and the green blocks indicate datasets or physical objects. See text on next page for callout to this figure.

The following workflow for metabolic flux analysis though isotope labeling will focus on its most common and established form: 13C Metabolic Flux Analysis (13C MFA) from proteogenic amino acids in the exponential phase. This is not to say that it is the most important, but rather the most mature and where agreement on a common workflow is most likely. That having been said, the modular nature of the workflow presented here allows for other varieties to be easily incorporated, some of which are still in development. For example, if intracellular metabolite labeling were to be used instead of amino acid labeling, this data (and the necessary metabolite concentrations) could be easily inserted at the same point in the diagram as amino acid labeling

(along with a connection to the metabolomics workflow). Other such variations such as flux analysis in a non-steady state [1], labeling of atoms other than carbon [2], or usage of NMR data [3] can be added in a similar fashion.

The workflow described here is an important tool for us as researchers for several reasons: 1) it explains the process to new members of the group as well as collaborators, 2) it helps define the standards for stored data in order to replicate and compare results in the future 3) it defines the steps used to track project completion and to help plan and develop high-throughput experiments.

The first step we include in the workflow is the characterization of the strain growth, a process not exclusive to 13C MFA. This characterization produces two sets of data that will be useful for planning the isotope labeled experiment: the growth curve and the concentration of extracellular metabolites. The growth curve provides the mid-log point used for sampling, and the extracellular metabolite concentration provides a rough idea of which metabolic pathways are important in addition to measured transport fluxes for later use. An example of a possible data input of extracellular metabolite concentration is shown in Figure 2. Useful metadata involves strain details, including plasmid and genetic modifications, along with materials and methods for OD and extracellular metabolite measurements.

Compound Summary

Sequence start: 2/20/2009 11:22:46 AM
Operator:

Method file name: C:\CHEM32\1\DATA\FEB-20-2009 2009-02-20 11-22-22\YEAST_FERMENTATION_JENNIFER.M

Sample Name	Sample Amt [g/L]	Multip.* Dilution	FileName .D	RetTime [min]	Amount [g/L]	Compound
Blank	0.00000	1.0000	001-0801	9.222	-	-
				9.325	-	-
				9.876	-	-
				12.810	-	-
				13.083	-	-
				13.769	-	-
				14.072	-	-
				14.345	-	-
				15.418	-	-
				15.693	-	-
				22.392	-	-
1-1	0.00000	1.0000	011-0901	9.217	0.032	Pyruvate
				9.230	16.379	Glucose
				9.876	-	-
				12.809	5.33320e-1	Lactate
				13.085	5.23185e-1	Lactate
				13.777	9.03251e-3	Glycerol
				14.087	6.99238e-2	Formate
				14.344	6.28119e-2	Formate
				15.423	3.43505e-1	Acetate
				15.701	3.34890e-1	Acetate
				22.410	1.87606e-1	EtOH
1-2	0.00000	1.0000	012-1001	9.215	0.027	Pyruvate
				9.229	16.774	Glucose
				9.876	-	-
				12.807	4.39012e-1	Lactate
				13.083	4.06618e-1	Lactate
				13.769	-	-
				14.087	5.33141e-2	Formate
				14.343	3.99210e-2	Formate
				15.420	2.81520e-1	Acetate
				15.699	2.65478e-1	Acetate
				22.408	1.16859e-1	EtOH
1-3	0.00000	1.0000	013-1101	9.215	0.030	Pyruvate
				9.225	16.606	Glucose
				9.876	-	-
				12.807	4.11458e-1	Lactate
				13.079	3.89868e-1	Lactate
				13.769	-	-
				14.087	5.74127e-2	Formate
				14.342	4.45182e-2	Formate
				15.419	3.00310e-1	Acetate
				15.694	2.99122e-1	Acetate
				22.401	1.36531e-1	EtOH
2-1	0.00000	1.0000	014-1201	9.215	0.028	Pyruvate
				9.226	16.606	Glucose

Figure 2. Example of output of HP-LC analysis used as input of extracellular metabolite concentration. A standard format for this information would be useful.

The main experimental process in the workflow is the performance of the labeling experiment, the workflow for which has been described by Zamboni et al [4] (see Figure 3). The necessary input for planning and performing the experiment includes the growth curve and extracellular metabolite concentrations, which has been discussed above, and the feed labeling, which affects the range of fluxes that can reliably be determined [5] [6]. The output includes the main piece of data needed to constrain the metabolic fluxes: the amino acid labeling pattern. The labeling information should include as metadata details of the experiment including sampling points, initial feed labeling and materials, and methods for labeling measurement. Examples of amino acid labeling data in terms of the derivatized fragments [7] or amino acid backbone labeling can be seen in Figs. 4 and 5 [8] [9].

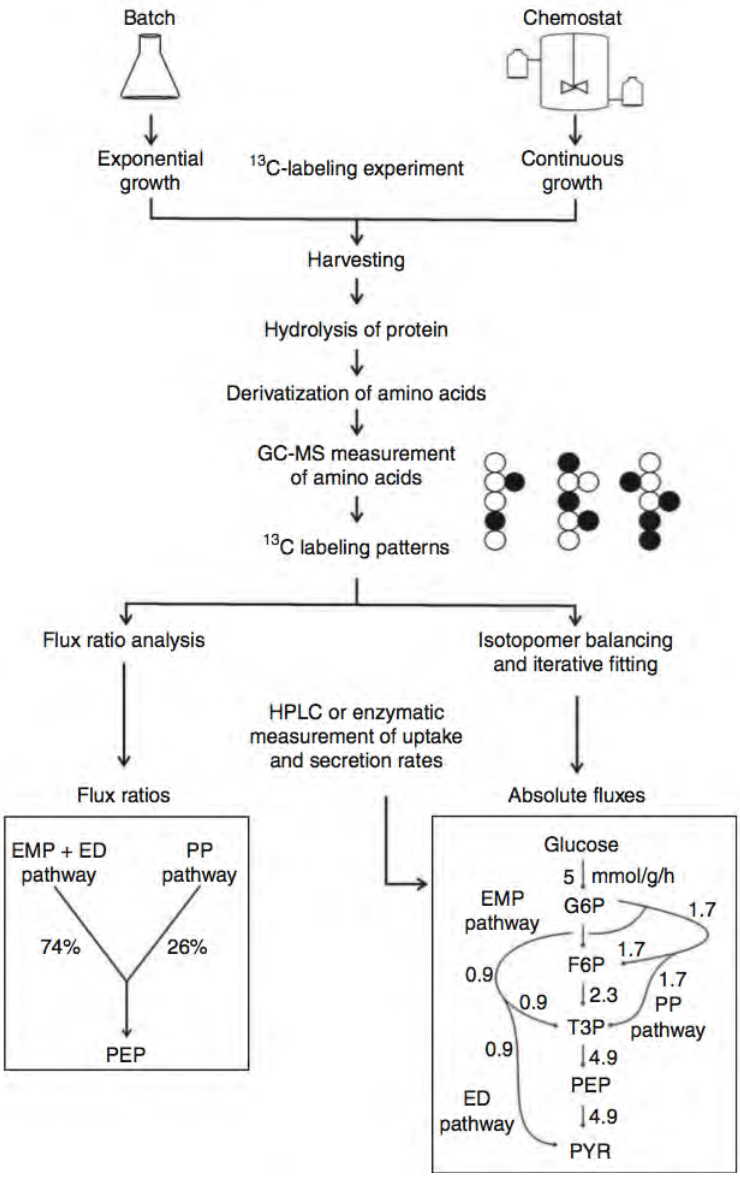


Figure 3. Workflow for carbon labeling experiment as per Zamboni et al [3] showing the two types of methods to obtain flux profiles: through flux ratio analysis or isotopomer balancing and iterative fitting.

SAMPLE:	WT+MET	WT+MET	WT+MET	MET28	MET28	CBF1	CBF1
REPLICATE	A	B	C	A	B	A	B
TBDMS FRAGMENT							
Ala232 (M0)	0.5666	0.5638	0.5633	0.5757	0.5755	0.5632	0.5699
Ala233 (M1)	0.3151	0.3174	0.3170	0.3102	0.3135	0.3173	0.3133
Ala234 (M2)	0.0907	0.0912	0.0916	0.0876	0.0857	0.0912	0.0898
Ala235 (M3)	0.0238	0.0235	0.0240	0.0227	0.0214	0.0242	0.0234
Ala236 (M4)	0.0034	0.0035	0.0036	0.0035	0.0037	0.0035	0.0032
Ala237 (M5)	0.0004	0.0005	0.0005	0.0003	0.0002	0.0005	0.0004
Ala260 (M0)	0.5603	0.5570	0.5542	0.5678	0.5626	0.5582	0.5595
Ala261 (M1)	0.3171	0.3192	0.3197	0.3115	0.3155	0.3174	0.3161
Ala262 (M2)	0.0930	0.0934	0.0944	0.0913	0.0898	0.0941	0.0939
Ala263 (M3)	0.0250	0.0255	0.0269	0.0247	0.0269	0.0258	0.0258
Ala264 (M4)	0.0040	0.0040	0.0040	0.0041	0.0045	0.0040	0.0041
Ala265 (M5)	0.0007	0.0008	0.0007	0.0007	0.0006	0.0005	0.0006
Ala266 (M6)	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Gly218 (M0)	0.7614	0.7592	0.7579	0.7612	0.7621	0.7585	0.7605
Gly219 (M1)	0.1603	0.1605	0.1613	0.1589	0.1575	0.1606	0.1596
Gly220 (M2)	0.0664	0.0685	0.0682	0.0664	0.0643	0.0687	0.0671
Gly221 (M3)	0.0098	0.0100	0.0104	0.0109	0.0128	0.0102	0.0107
Gly222 (M4)	0.0019	0.0017	0.0019	0.0021	0.0025	0.0019	0.0020
Gly223 (M5)	0.0002	0.0001	0.0003	0.0005	0.0008	0.0001	0.0002
Gly246 (M0)	0.7479	0.7491	0.7479	0.7529	0.7581	0.7511	0.7529
Gly247 (M1)	0.1694	0.1683	0.1696	0.1673	0.1640	0.1667	0.1664
Gly248 (M2)	0.0706	0.0710	0.0708	0.0687	0.0667	0.0703	0.0693
Gly249 (M3)	0.0103	0.0099	0.0100	0.0093	0.0094	0.0103	0.0096
Gly250 (M4)	0.0018	0.0016	0.0018	0.0018	0.0019	0.0017	0.0017
Val260 (M0)	0.4119	0.4087	0.4130	0.4269	0.4181	0.5802	0.6023
Val261 (M1)	0.3806	0.3828	0.3795	0.3737	0.3797	0.2760	0.2637
Val262 (M2)	0.1532	0.1531	0.1525	0.1482	0.1475	0.1105	0.1039
Val263 (M3)	0.0451	0.0447	0.0436	0.0430	0.0447	0.0273	0.0248
Val264 (M4)	0.0086	0.0094	0.0098	0.0082	0.0099	0.0056	0.0046
Val265 (M5)	0.0006	0.0013	0.0015	0.0000	0.0000	0.0003	0.0006
Val288 (M0)	0.4098	0.4054	0.4083	0.4247	0.4139	0.5779	0.5984
Val289 (M1)	0.3812	0.3839	0.3802	0.3729	0.3776	0.2771	0.2638
Val290 (M2)	0.1545	0.1552	0.1559	0.1497	0.1550	0.1120	0.1062
Val291 (M3)	0.0440	0.0448	0.0448	0.0428	0.0441	0.0272	0.0255
Val292 (M4)	0.0091	0.0093	0.0095	0.0087	0.0088	0.0052	0.0055
Val293 (M5)	0.0013	0.0014	0.0013	0.0012	0.0005	0.0007	0.0007
Leu274 (M0)	0.3110	0.3088	0.3100	0.3312	0.3190	0.7240	0.7293
Leu275 (M1)	0.3885	0.3898	0.3890	0.3845	0.3895	0.1899	0.1854
Leu276 (M2)	0.2111	0.2114	0.2100	0.2035	0.2063	0.0748	0.0723
Leu277 (M3)	0.0689	0.0688	0.0699	0.0624	0.0662	0.0098	0.0113
Leu278 (M4)	0.0170	0.0175	0.0174	0.0157	0.0161	0.0015	0.0017
Leu279 (M5)	0.0031	0.0032	0.0032	0.0025	0.0029	0.0000	0.0000
Leu280 (M6)	0.0005	0.0005	0.0004	0.0002	0.0000	0.0000	0.0000

Figure 4. Amino acid labeling for different derivatized fragments, taken from [7]. The name on the left column corresponds to the amino acid and the fragment type [6]. Each of the following columns corresponds to the fraction of molecules with 0,1,2... extra mass units incorporated due to isotopic variation (from carbon or other atoms).

Table 1. Relative Intensity and Error Associated with the Measurement of Each Amino Acid Isotopomer[a]

amino acid	rel intens (%)	error (ppm)	amino acid	rel intens (%)	error (ppm)	amino acid	rel intens (%)	error (ppm)	amino acid	rel intens (%)	error (ppm)
Gly			Val			Lys			Phe		
M0	1.30	0.21	M0	11.60	0.07	M0	0.90	0.15	M0	0.65	0.17
M1	4.20	0.22	M1	42.40	0.01	M1	4.20	0.09	M1	2.60	0.21
			M2	1.80	0.00	M2	8.00	0.10	M2	5.60	0.20
						M3	0.30	0.09	M3	3.70	0.21
Ala			Thr			Glu			Arg		
M0	4.70	0.00	M0	1.20	0.02	M0	8.30	0.16	M0	0.30	0.27
M1	19.50	0.01	M1	6.00	0.02	M1	23.30	0.08	M1	1.00	0.27
M2	0.40	0.00	M2	11.50	0.03	M2	1.00	0.09	M2	1.80	0.37
			M3	0.20	0.03				M3	0.07	0.21
Ser			Leu (and Ile)			Met			Tyr		
M0	2.30	0.03	M0	100	0.06	M0	0.15	0.17	M0	nd[b]	
M1	7.30	0.05	M1	71.70	0.01	M1	1.00	0.18	M1	0.10	0.25
M2	0.20	0.00	M2	4.20	0.01	M2	4.40	0.14	M2	0.20	0.42
			M3	0.10	0.04	M4	8.20	0.13	M3	0.15	0.15
Pro			Asp			His					
M0	19.00	0.03	M0	1.45	0.01	M0	2.40	0.12			
M1	55.20	0.02	M1	7.30	0.03	M1	3.80	0.12			
M2	2.50	0.02	M2	15.10	0.03	M2	0.90	0.09			
M3	0.06	0.40	M3	0.30	0.07	M4	0.30	0.08			

[a] M0, M1, M2, etc., refers to isotopomers with 0, 1, 2, etc., ^{13}C incorporated in the backbone of the amino acid. A RSD of ≤2% is associated with each relative inensity measurement. Errors refer to one single measurement. A variation of 10% is associated with it. [b] Not detected.

Figure 5. Amino acid labeling for carbon backbone. M0, M1, M2... indicate the fraction of molecules with 0,1,2.... labeled carbons incorporated [8].

Extracellular metabolite concentrations from the growth characterization experiment are used to derive the transport fluxes, (i.e. uptake and secretion rates). The calculation from extracellular metabolites is straightforward, and it involves calculating the change in metabolite concentration in the media. Another important set of known fluxes is the fluxes to biomass production, obtained from the change in OD and the cell composition.

Fitting the fluxes to the labeling data is the main computational process in the workflow. A variety of methods are available to do this [7] [10] [11]. Some involve determination of local flux ratios, and some are based on iterative fittings for the whole metabolic network under consideration (see Figure 3). Among the latter, the fit can either be performed in a search space involving fluxes and labeling, with the labeling pattern included as a constraint [12], or in a search space involving only fluxes, with the labeling determined for each flux profile. Labeling corresponding to each flux profile can be produced using several methods, including isotopomer mapping matrices [13], cumomers [14] or elementary metabolic units [15], to name a few. The search through the flux phase space can, as well, be carried over via a variety of techniques, including genetic algorithms, sequential quadratic programming and simulated annealing [7]. Software for flux calculations include 13CFLUX [16] [4] and FIATFLUX [17], none of which are available in open source format, and openFLUX [18], a recent application based on elementary metabolic units available in open source format. For the purpose of designing a workflow, what is important is not the differences among these methods but the fact that they all require the same input: 1) transport and biomass fluxes, 2) amino acid labeling, 3) initial feed labeling, and 4) the carbon transitions included in a metabolic reconstruction. Amino acid and initial labeling patterns, and measured fluxes have all been discussed above. The metabolic reconstruction has been discussed at length above; the only required condition is that it includes atomic transitions (see example in Figure 6 [19]). This metabolic model may be a coarse grained version of the models considered above. See, for example, Figs. 6 and 7, where

some reactions have been clumped together to ease calculations. Recently, a new tool for sharing, storing, and constructing these atomic transitions embedded in a metabolic reconstruction has become available [20]. Standard formats used in this program are the 13CFLUX format and SBML.

Supplementary Table S-2 Reaction lists:

```
% Lactate to pyruvate

1: ABC(1) -> ABC(2)

% Pyruvate to Acetyl-CoA

2: ABC(2) -> BC(3) + A(17)

% Oxaloacetate to citrate

3: AB(3) + abcd(11) -> dcbBAa(5)

% Citrate to isocitrate

4: ABCDEF(5) -> ABCDEF(6)

% Isocitrate to 2-oxoglutarate

5: ABCDEF(6) -> ABCDE(7) + F(17)

% 2-oxoglutarate to succinyl-CoA

6: ABCDE(7) -> BCDE(8) + A(17) :: ABCDE(7) -> EDCB(8) + A(17)

% Succinate to Malate

7: ABCD(8) -> ABCD(9)

% Fumarate -> malate

8: ABCD(9) -> ABCD(10)
```

Figure 6. Example of atomic transitions input needed form 13C MFA [18]. The first number indicates the reaction number as per Figure 7, and the numbers in parenthesis indicate the metabolite numbers. Carbon transitions are indicated as strings of letters: e.g., ABC -> AB + C indicates that the first two carbon in the reactant end up as the two carbons in the first product and the last carbon goes to the second product.

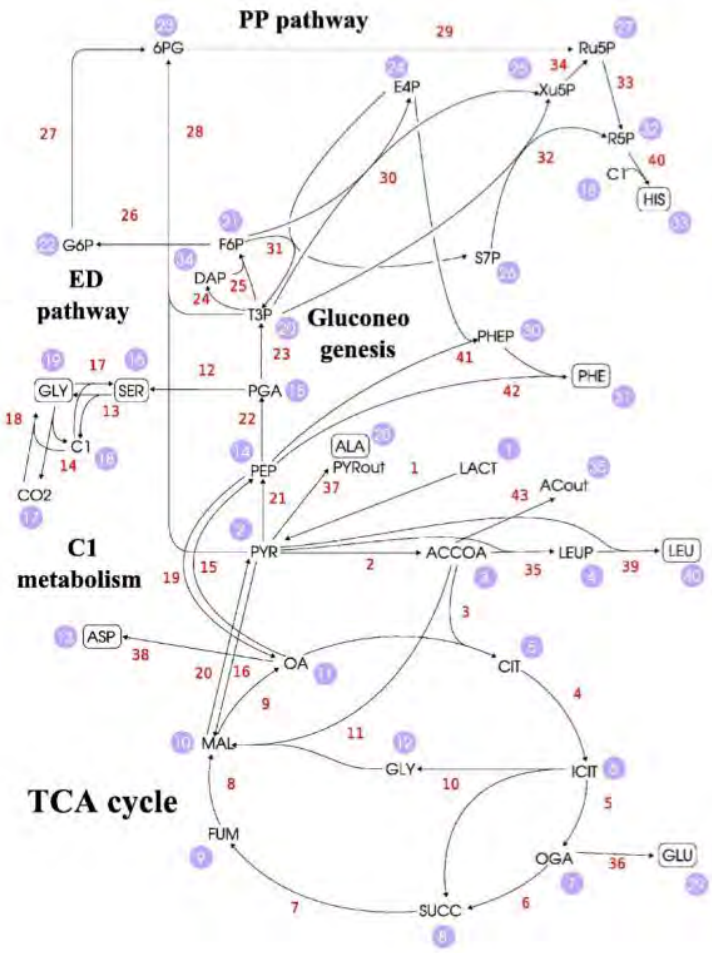


Figure 7. Reaction network for *Shewanella* central carbon metabolism [18]. Notice how some of the reactions (e.g., g6p to Ru5P in the penthose-phosphate pathway) have been clumped together to ease calculations).

The output should include the flux profile giving the best fit for the experimental data, a confidence interval, and the computed labeling patterns. The metabolic flux profile gives the best guess (compatible with the data) of the rate for each of the metabolic reactions considered in the metabolic model of the cell. This information is useful *per se* as a widely recognized highly relevant characteristic of the phenotype [21], and has numerous applications in (e.g.) metabolic engineering [11]. As with every experimental measurement, it is also desirable to assign confidence intervals to flux estimates, and a variety of algorithms are available for this purpose [7].

A simple list of Fluxes with their corresponding confidence intervals for each metabolic reaction can be very difficult to make productive use of, particularly for large models. Hence, visualization is an important part of the workflow and several possibilities are available [22] [23] [24] [25], although not all of them allow flux visualization for models with clumped reactions.

Finally, a useful visual check that the fit is appropriate is to compare computational predictions with experimental data, as shown in Figure 8.

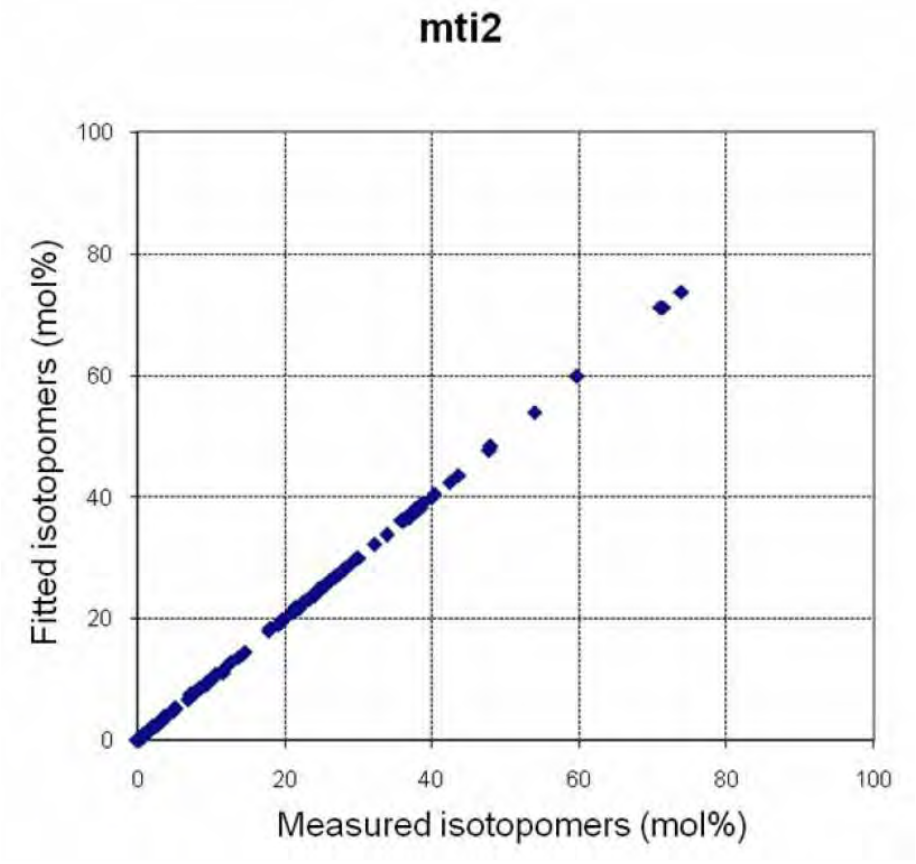


Figure 8. Comparison between computed and measured labeling data [7]. A good fit does not deviate from the diagonal.

References

[1] Katharina Nöh, Aljoscha Wahl, and Wolfgang Wiechert, Metab. Eng **8**, 554-577 (2006).

[2] Jie Yuan, William U Fowler, Elizabeth Kimball, Wenyun Lu, and Joshua D Rabinowitz, Nat Chem Biol **2**, 529-530 (2006).

[3] Ari Rantanen, Juho Rousu, Paula Jouhten, Nicola Zamboni, Hannu Maaheimo, and Esko Ukkonen, BMC Bioinformatics **9**, 266 (2008).

[4] Nicola Zamboni, Sarah-Maria Fendt, Martin Rühl, and Uwe Sauer, Nat Protoc **4**, 878-892 (2009).

[5] YoungJung Chang, Patrick F. Suthers, and Costas D. Maranas, Biotechnology and Bioengineering **100**, 1039-1049 (2008).

[6] M Möllney, W Wiechert, D Kownatzki, and A A de Graaf, Biotechnol. Bioeng **66**, 86-103 (1999).

[7] Yinjie J Tang, Hector Garcia Martin, Samuel Myers, Sarah Rodriguez, Edward E K Baidoo, and Jay D Keasling, Mass Spectrom Rev **28**, 362-375 (2009).

[8] Joel F. Moxley, Michael C. Jewett, Maciek R. Antoniewicz, Silas G. Villas-Boas, Hal Alper, Robert T. Wheeler, Lily Tong, Alan G. Hinnebusch, Trey Ideker, Jens Nielsen, and Gregory Stephanopoulos, Proceedings of the National Academy of Sciences **106**, 6477-6482 (2009).

[9] Francesco Pingitore, Yinjie Tang, Gary H Kruppa, and Jay D Keasling, Anal. Chem **79**, 2483-2490 (2007).

[10] Michael Dauner, Curr Opin Biotechnol (2010).

[11] Shintaro Iwatani, Yohei Yamada, and Yoshihiro Usuda, Biotechnol. Lett **30**, 791-799 (2008).

[12] Patrick F Suthers, Anthony P Burgard, Madhukar S Dasika, Farnaz Nowroozi, Stephen Van Dien, Jay D Keasling, and Costas D Maranas, Metab. Eng **9**, 387-405 (2007).

[13] Schmidt, Marx, de Graaf AA, Wiechert, Sahm, Nielsen, and Villadsen, Biotechnol. Bioeng **58**, 254-257 (1998).

[14] W Wiechert, M Möllney, N Isermann, M Wurzel, and A A de Graaf, Biotechnol. Bioeng **66**, 69-85 (1999).

[15] Maciek R. Antoniewicz, Joanne K. Kelleher, and Gregory Stephanopoulos, Metab Eng **9**, 68-86 (2007).

[16] Wolfgang Wiechert, Michael Möllney, Sören Petersen, and Albert A. de Graaf, Metabolic Engineering **3**, 265-283 (2001).

[17] Nicola Zamboni, Eliane Fischer, and Uwe Sauer, BMC Bioinformatics **6**, 209 (2005).

[18] Lake-Ee Quek, Christoph Wittmann, Lars K Nielsen, and Jens O Krömer, Microb Cell Fact **8**, 25-25 (n.d.).

[19] Yinjie J. Tang, Héctor García Martín, Paramvir S. Dehal, Adam Deutschbauer, Xavier Llora, Adam Meadows, Adam Arkin, and Jay. D. Keasling, Biotechnol. Bioeng. **102**, 1161-1169 (2009).

[20] Esa Pitkänen, Arto Akerlund, Ari Rantanen, Paula Jouhten, and Esko Ukkonen, J Integr Bioinform **5**, (2008).

[21] Uwe Sauer, Mol. Syst. Biol **2**, 62 (2006).

[22] Suzanne M Paley and Peter D Karp, Nucleic Acids Res **34**, 3771-3778 (2006).

[23] Nobuaki Kono, Kazuharu Arakawa, Ryu Ogawa, Nobuhiro Kido, Kazuki Oshita, Keita Ikegami, Satoshi Tamaki, and Masaru Tomita, PLoS ONE **4**, e7710 (2009).

[24] F Le Fèvre, S Smidtas, C Combe, M Durot, Florence d'Alché-Buc, and V Schachter, Bioinformatics **25**, 1987-1988 (2009).

[25] Eva Grafahrend-Belau, Christian Klukas, Björn H Junker, and Falk Schreiber, Bioinformatics **25**, 2755-2757 (2009).

Workflow 3: Inference of Gene Regulatory Networks

Summary

Gene regulatory networks (GRNs) are the "on-off" switches and rheostats of cells that operate at the gene level. They dynamically orchestrate the level of expression for each gene in the genome by controlling whether and how vigorously that gene will be transcribed into mRNA. Understanding how GRNs work is key to systems biology and its successful applications. An array of input data types exists. Knowledgebase users should be able to select an organism, upload, broadcast, or import expression data from public repositories, and submit a request for gene regulatory network inference. Meta-information on experiment design should be automatically parsed from public data, or the user should be prompted to upload this information. Users may want to start with a set of genes or a metabolic process and ask which factors are its regulators. Another use scenario is that researchers may want to know the gene targets of regulatory elements.

Genes will be grouped into putative regulatory modules whose transcription is correlated under specific conditions. For each module, the user selects a subset of known transcription factors and environmental factors that best predict the transcription levels. Additional inputs, such as motifs or protein interactions, may be statistically integrated in the clustering step or the network inference step, and shared regulatory motifs can be computed. Several algorithms have been devised for clustering and discovery of regulatory influences. Results can be exported as raw data or presented to the user in a searchable and browsable form. Subnetworks can be graphically displayed along with views of expression profiles and regulatory motifs and the gene content of individual clusters. Useful output will also include the ability to compute and present predictions (and confidence estimates on predictions) of effect of transcription factor deletions/overexpressions and/or environmental changes.

Inputs

1. Depending on the specific type of network inference analysis a user has in mind, a different combination of the following data might be necessary; but minimally, these seven types of information cover most of what is available today.
2. Measurements of transcription (with confidence values [if avail.]) in the form of an n x m matrix with n genes and m conditions (microarray or sequencing)
3. Measurements of fitness associated with systematic gene knockouts or over-expression (maybe these last two can be condensed in measures of genome-scale gene function with confidence in the form of an n x m matrix.... This could be generalized as phenotype and also have associated confidence depending on how it's measured.
4. Gene interaction network(s) = [nodes (genes), edges (interactions/type), confidence values or weights for edges]
5. Gene locations on genome—with RNA-Seq this is becoming extremely precise with direct measurement.
6. Genome sequence or individual upstream sequences (for motif detection)

7. A list of predictors (transcription factors, environmental factors including metabolites)
8. Machine-readable descriptions of conditions, specifically time series info with standardized measurements of environmental factors

User will specify an organism and import (or broadcast) the above data items. Many of these data types are stored in existing databases and can be loaded automatically through interoperability with these data sources. Many of these data types (items 2-4) can be obtained automatically given the organism. Item 1 may be obtained automatically from expression databases such as GEO or MicrobesOnline. Item 3 can be obtained from STRING, and some information is also accessible in MicrobesOnline. Items 4-6 can be obtained from NCBI or MicrobesOnline or other databases.

However, these data are not available for all organisms. One addendum to this work is that many of these types of measurements follow a standard experimental workflow. Once a genome of a cultivated organism gets sequenced, it might be useful to develop a minimal set of functional measurements to aid in this.

As these workflows are being developed and have increasingly precise data such as RNA-Seq and can have associated confidence measures that can be carried through the analyses, this is providing a basis for comparing the precision of results between methods and laboratories that would help to improve quality and would benefit existing systems such as GEO if applied consistently.

Apply clustering and network inference
Group the genes into putative regulatory modules whose transcription is correlated over a set of conditions. Select a subset of known transcription factors and environmental factors that best predict the transcription levels of each module. Additional inputs, such as motifs or protein interactions, may be statistically integrated in the clustering step or the network inference step, and shared regulatory motifs can be computed. Several algorithms have been devised for clustering and discovery of regulatory influences; some are available in R and MatLab.

Outputs

- Clusters of putatively coregulated genes or biclusters containing genes putatively coregulated under subsets of conditions
- Cis-regulatory motifs
- Regulatory network mapping: influences of predictors on genes within clusters/biclusters directly or through *and* and *or* operations. Confidence values for edges.

Results can be exported as raw data or presented to the user in a searchable and browsable form. Users may want to start with a set of genes or a metabolic process and ask which factors are its regulators. Or, users may want to take a given regulator and ask what are its targets. Subnetworks can be graphically displayed along with graphical views of expression profiles and regulatory motifs and the gene content of individual clusters. Useful output will also include the

ability to compute and present predictions (and confidence estimates on predictions) of effect of TF deletions/overexpressions and/or environmental changes.

Scope

A user should be able to select an organism, upload, broadcast, or import expression data from public repositories, and submit a request for network inference; metainformation on experiment design should be automatically parsed from public data, or a user should be prompted to upload this information. All other data types can be automatically parsed from public repositories—an advanced user should have privileges to change or override default settings by changing source of information, threshold of significance, etc. A user should be given options for choice of algorithms based on the amount and type of available data; the user should have access to published citations for the algorithms and basic information on workings of the algorithm in non-technical jargon-free language. It should be possible to store a session with the default or user-edited settings so the entire analysis can be recreated.

Data requirements and computational complexity

It seems that these might be important numbers both from the user's perspective and from the planning perspective, but can these be coherently calculated? We can give some perspectives to the user, for instance, to infer a network with causal influences time series data are a must; for better coverage of regulons one needs to probe responses to at least half a dozen or a dozen environmental perturbations with different dosages and over the time scale of the response; to incorporate mechanisms we need to have physical interactions (P-D, P-P), or information on TF-cis-regulatory motif relationships. However, in principle, one could learn a network based on correlations and cis-regulatory motifs with a relatively small dataset (30-50 experiments - see Gardner's CLR algorithm or Bar-Joseph's DREM). Such a network will give a very limited view of transcriptional control but could be deemed extremely valuable for an organism for which absolutely nothing was known previously. Given the diverse variations in use cases, while we could consider very simple to very sophisticated cases, I would argue that we should focus on use-cases of simple to mid-scale complexity. I say this because advanced users with sophisticated needs are likely to have the capability to do it themselves (without a knowledgebase).

We could have minimal requirements imposed on algorithm developers when they submit their work. This would include a listing of requirements (number of experiments, interaction data etc.). It might be instructive to have the following information as well; I am not sure if we can generalize this to other use cases.

- Estimates of number and diversity of experiments necessary for clustering
- Estimates of quality needed—issues of quality, compendium biases, etc.
- Estimates of computational complexity of biclustering/bayes nets/etc.

Notes

Are there other players we'd like to incorporate, like RNA regulatory elements, for instance? Would we want to get more out? Network motifs? How about Lee's fusion of kinetics and GRNs? Does that require additional input or generate additional output?

Certainly, moving beyond the inference of regulatory structure and gross dynamics would require different experiments. Inferring metabolism requires both different sorts of functional assays and genome-scale experiments; inferring signaling pathways has its own troubles (see Aindrila Mukhopadhyay's document); inferring complex regulation like that implemented in control of sporulation requires more detailed microscopic measurement and mechanistic modeling. However, here we have the opportunity for something that could almost become a standard after analysis of any sequenced genome.

Inference and Measurement

Is it possible to describe the situations where it is better to try to infer genetic regulatory network topology, rather than try to measure the regulatory interactions directly? There have been remarkable experimental strides made in determining the sequences to which transcription factors bind (e.g., Hesselberth et al, "Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting," Nature Methods 6, 283 - 289 (2009) doi:10.1038/nmeth.1313).

Workflow 4: Signaling

Summary

Due to the focus on microbes and microbial communities, the workflow pertains to signaling in bacteria. Bacterial signaling forms a subset of the Genetic Regulatory Network discussion (see Workflow 3) and therefore contains overlap in experimental design, analysis and workflow.

Microbial genomes present signaling systems to sense and respond to both external and internal stimuli[1,2]. Signals include numerous factors considered to be stresses, intracellular cues, and environmental changes. In bacteria, two component signal transduction systems, typically comprised of a sensor histidine kinase and a response regulator, provide the primary mechanism of signal sensing and response[3,4]. Signal transduction occurs via phosphotransfer or phophorelay and results in an activated response regulator. The best-studied response mechanisms include either the direct modulation of chemotaxis by the activated response regulator, or in a large number of studies, response regulator modulated differential expression of target genes. New classes of response regulators that modulate function via alternate mechanisms such as c-diGMP cyclase or phosphodiesterase domains have also been described[1,2]. Available sequenced genomes from environmental organisms encode numerous sensor and response regulator proteins containing domains of unknown function indicating that additional mechanisms for effector function have yet to be discovered. Environmental bacteria such as *Geobacter metallireducens*, *Desulfovibrio vulgaris,* and the cyanobactrium *Nostoc spp.* have upward of 60, to more than 150, sensor kinases[5]. The responses regulated by the corresponding two component systems are no doubt at the core of environmental process of key significance. These systems also provide the parts for developing valuable sensory modules to build sophisticated engineered systems (using synthetic biology methods).

Definition of a signal: With regard to the type of research being conducted by the Genomic Science groups, signals can vary widely. In environmentally relevant microbes, a signal could be a change in environmental cue (e.g., the lack/abundance of resources such as carbon source, electron acceptors, electron donors, amino acids, vitamins, etc.); stresses (e.g., salt, pH, heat, cold, metals, toxins, oxygen, a variety of small molecules); or variability in other organisms in the microenvironment. The responses to these signals, including the triggering of altered physiological states (e.g., biofilm formation, sporulation, virulence, swarming, etc.) are all initiated via signal sensing and corresponding response. In microbes that are being engineered for industrial uses (e.g., biofuel production), perturbation from toxins present in carbon feed, intracellular triggers due to imbalance in metabolic intermediates, and accumulation of final (often toxic) products serve as signals.

A vast body of knowledge exists for these systems from individually studied systems. Efforts to compile and integrate information on regulatory modules from such studies have only recently begun to emerge as described in Workflow 3. However, the impact of multiple stimuli on a given organism or comprehensive understanding of all signal sensing for a single organism is still rare. In the few cases where such studies have been undertaken, valuable and interesting phenomenon have been discovered[6,7]. The tremendous increase in sequenced organisms and

corresponding computational[2,5,8] and experimental tools[9,10] now makes it more possible to undertake such efforts.

Metadata: This documents what signals are being studied, under what defined conditions, on which organisms and by whom? What media and growth phases are being used? What methods are being used (sequencing, arrays, analytical)? What analysis tools, *in silico* prediction algorithms, and validation experiments are to be used?

Aspects of studying signaling

- *What are the genomes in question?* For a given organism, there may be more than one genome sequence if there are modified, engineered, evolved, adapted or shuffled versions.
- *Sensing and responding to signals:* Study of two-component, cAMP, and c-diGMP systems, transcriptional factors, global regulators, sigma factors; cell-wide studies (transcriptomics, proteomics, and metabolomic studies), and mapping ligand (signal) binding, phospho-transfer, and other post-translation modification.
- *Information gathered:* Ligand binding and transport, two component phosphorylation, other assays (chemotaxis, binding to cyclic-diGMP, DNA gel shifts), ChIP-chip arrays, ChIP-seq, microrrays, RNA-Seq, mapping post translational modifications (phosphoproteome, methylations etc), mapping protein interactions and localization.

Data types

This will form the core of the database for this topic

- Genome sequences
- Knockout and expression libraries: corresponding phenotypic data (e.g. from omniglogs or other such HT strategies)
- Transcript level data: microarray, RNA-Seq, absolute mRNA quants (e.g. nCOUNTER)
- DNA binding: ChIP-chip, ChIP-seq, microfluidics
- Mass Spec data: Protein levels, Post translational modifications, metabolites
- (data from different types of mass spectrometers)
- Ligand binding mapping: Semi HT
- Regulator-DNA binding: Semi HT
- Regulatory motifs and maps generated using computational methods

Resources

- Common sensory proteins include histidine kinases, methyl-accepting chemotaxis receptors, Ser/Thr/Tyr protein kinases, adenylate and diguanylate cyclases and c-di-GMP phosphodiesterases. A webpage maintained Galperin and coworkers contains a

fairly comprehensive repository of signal transduction systems from over 500 bacteria and archaea[11].

- Predictive tools and databases (e.g. Regtransbase [12], MicrobesOnline[13], and MiST[16]).
- Specific tools for predicting cognate partners of sensor histidine kinases and response regulators such as that developed by Burger and van Nimwegen[8].
- Methods developed for mapping two component phosphotransfer and relay[9], rewiring sensor kinases[14].
- Classification system developed for categorizing bacterial signaling proteins[1,2,15].

Illustration using one concrete problem

BESC is studying a number of Caldicellulosirupter species, which are non-sporulating, anaerobic, gram positive, thermophilic bacteria that can facilitate the direct conversion of cellulosic biomass (e.g. from switchgrass) to ethanol, H2 and other products.

A study of this organism will utilize the features afforded by other knowledgebases: Annotated genomes and their use in generating arrays, predictions for regulatory networks and motifs, predicted two component systems, transcriptional factors (including those that work with transporters), sigma factors, small RNA regulators.

A given experiment would entail growth of Caldicellulosirupter on ground plant material and monitor production of waste products including ethanol, H2, acetate etc. Genetic engineering could create the production of alternative end products in different proportions.

A range of factors (signals) would be examined in this context. Beneficial factors include C source, cell density, etc. Harmful factors include exposure to O2, cold shock, inhibitors from lignocellulosic biomass, acetate, non-optimal pH, salt, and the accumulation of other final products.

Current studies include: Log phase growth using cellobiose and switchgrass (pretreated), stationary phase growth in switchgrass (pretreated), log phase growth under ethanol stress with either cellobiose or switchgrass.

A systematic examination of any of the above factors could be conducted using the following:

1. Transcript level measurements: arrays, RNA-Seq, other targeted measurements.
2. ChIP-chip, ChIP-seq
3. Analytical assays:
    a. HT: Mass spec based analysis (protein levels, protein complexes, PTMs)
    b. MT: ligand-docking, transport,
    c. LT: Mapping HK-RR phosphotransfer, RR-DNA gel shifs
    d. LLT: Imaging for morphological changes or cellular localization of complexes.
4. Study of knockout or expression strain libraries (transposon, targeted, site specific). Corresponding phenotypic data and iterative (1), (2) and (3)

5. Collecting and integrating the above data types into previous or initial regulatory network prediction (see Workflow 3: Inference of Gene Regulatory Networks).

HT: High throughput; **MT**: Medium throughput; **LT**: Low throughput; **LLT**: Low Low throughput; **PMT**: Post translational modifications.

References

[1.] Galperin, M. Y., Diversity of structure and function of response regulator output domains, *Curr Opin Microbiol* 13 (2), 150-9, 2010.

[2.] Galperin, M. Y., Higdon, R., and Kolker, E., Interplay of heritage and habitat in the distribution of bacterial signal transduction systems, *Mol Biosyst* 6 (4), 721-8, 2010.

[3.] Stock, A. M., Robinson, V. L., and Goudreau, P. N., Two-component signal transduction, *Annu Rev Biochem* 69, 183-215, 2000.

[4.] Gao, R. and Stock, A. M., Biological Insights from Structures of Two-Component Proteins, *Annu Rev Microbiol*, 2009.

[5.] Alm, E., Huang, K., and Arkin, A., The evolution of two-component systems in bacteria reveals different strategies for niche adaptation, *PLoS Comput Biol* 2 (11), e143, 2006.

For histidine kinases in various genomes see www.microbesonline.org/cgi-bin/hpk/browse.cgi

[6.] Tagkopoulos, I., Liu, Y. C., and Tavazoie, S., Predictive behavior within microbial genetic networks, *Science* 320 (5881), 1313-7, 2008.

[7.] Skerker, J. M., Prasol, M. S., Perchuk, B. S., Biondi, E. G., and Laub, M. T., Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis, *PLoS Biol* 3 (10), e334, 2005.

[8.] Burger, L. and van Nimwegen, E., Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method, *Mol Syst Biol* 4, 165, 2008.

[9.] Laub, M. T., Biondi, E. G., and Skerker, J. M., Phosphotransfer profiling: systematic mapping of two-component signal transduction pathways and phosphorelays, *Methods Enzymol* 423, 531-48, 2007.

[10.] Zhou, L., Lei, X. H., Bochner, B. R., and Wanner, B. L., Phenotype microarray analysis of *Escherichia coli* K-12 mutants with deletions of all two-component systems, *J Bacteriol* 185 (16), 4956-72, 2003.

[11.] Galperin, M. Y., Higdon, R., and Kolker, E., www.ncbi.nlm.nih.gov/Complete_Genomes/SignalCensus.html, 2010.

[12.] Kazakov, A. E., Cipriano, M. J., Novichkov, P. S., Minovitsky, S., Vinogradov, D. V., Arkin, A., Mironov, A. A., Gelfand, M. S., and Dubchak, I., RegTransBase--a database of regulatory sequences and interactions in a wide range of prokaryotic genomes, *Nucleic Acids Res* 35 (Database issue), D407-12, 2007.

[13.] Dehal, P. S., Joachimiak, M. P., Price, M. N., Bates, J. T., Baumohl, J. K., Chivian, D., Friedland, G. D., Huang, K. H., Keller, K., Novichkov, P. S., Dubchak, I. L., Alm, E. J., and Arkin, A. P., MicrobesOnline: an integrated portal for comparative and functional genomics, *Nucleic Acids Res* 38 (Database issue), D396-400, 2009.

[14.] Salis, H., Tamsir, A., and Voigt, C., Engineering bacterial signals and sensors, *Contrib Microbiol* 16, 194-225, 2009.

[15.] Mascher, T., Helmann, J. D., and Unden, G., Stimulus Perception in Bacterial Signal-Transducing Histidine Kinases, *Microbiol. Mol. Biol. Rev.* 70 (4), 910-938, 2006.

[16]. Ulrich, L.E., Zhulin, IB. MiST: a microbial signal transduction database. *Nucleic Acids Res.* 35:D386-390, 2007

Workflow 5: Structural Biology

Summary

The nucleic acid sequence of a protein gene encodes an amino acid sequence that typically folds to generate a specific three-dimensional shape. This structure is often vital for the protein's function. In enzymes, the structure serves to keep key catalytic residues in a unique geometry, poised to act on substrate molecules. As such, the relationship between primary sequence and tertiary shape is central to our understanding of molecular biology. There is a wealth of information that is ripe for analysis in the context of the Knowledgebase. For example: the relationship between sequence and fold (for proteins and nucleic acids), the assembly of single molecules to form larger complexes, and the evolutionary relationships within and between protein families. Rapid progress can be made in providing functionality to researchers via the Knowledgebase. Initial workflows can focus on visualizing the linkage between sequence and structure (see first and second workflows below) and dissection and visualization of cellular compartments (see third workflow below). These will provide users with powerful tools to probe sequence/structure relationships, which otherwise are limited to experts.

Three structural biology workflows were submitted:

1. Locating and visualizing an enzyme active site
    a. Goal: Assign, then visualize the amino acid residues in a protein sequence involved in enzymatic activity
2. Determine and visualize the oligomeric state of molecular complexes
    a. Goal: Determine and then visualize the oligomeric state of a protein complex
3. Locating and visualizing a cellular compartment
    a. Goal: Locate (segment) and visualize one or more cellular compartments in a microbe.

For each research goal, the Inputs, Analysis process, Outputs, Tools, and Knowledgebase context were provided.

1. Locating and visualizing an enzyme active site

Goal: To assign and then visualize the amino acid residues in a protein sequence involved in enzymatic activity.

Inputs

- Sequence of a protein
- High resolution protein structure (from X-ray crystallography or NMR) or high fidelity homology model
- One or more related sequences/structures with known active site residues

Process

- Perform alignment (most likely using multiple proteins) of sequence with unknown active site residues (may also include multiple family members) against known residues
- In cases of high sequence similarity, the active site residues in the unknown can be identified by sequence conservation
- In cases of remote similarity, more complex models (e.g. hidden Markov, sequence motifs, combined sequence/structure alignment) may need to be generated to infer the likely equivalent residues in the unknown
- Predictions of active site residues can be validated against any prior biochemical data and/or phyologenetic information

Outputs

- Protein sequence with active site residues highlighted
- Visual representation in standard molecular viewing software with active site residues highlighted

Tools required

- Parsing protein structure and sequence
- Single and multiple sequence alignment
- Combined sequence/structure alignment
- Sequence display
- 3D structure display

Knowledgebase context

- Provides linkage to and automatic retrieval of related structures in the Protein Data Bank
- Performs complex sequence and sequence/structure analysis without detailed user learning
- Cross validates against other experimental data within the Knowledgebase and in other outside resources
- Displays results in easy to understand visual forms and for download and subsequent analysis

2. Determining and visualizing the oligomeric state of molecular complex

Goal: To determine and then visualize the oligomeric state of a protein complex.

Inputs

- Sequence of a protein
- One or more related sequences/structures with known oligomeric state
- Optionally experimental data to define oligomeric state, such as small angle X-ray scattering (SAXS)
- Optionally high resolution protein structure (from X-ray crystallography or NMR) or homology model

Process

- Perform alignment (most likely using multiple proteins) of sequence with unknown oligomeric state (may also include multiple family members) against sequences of known state
- In cases of high sequence similarity, the likely oligomeric state can be identified from the nearest similar sequence
- In cases of remote similarity, more complex models (e.g. combined sequence/structure alignment) may need to be used to determine if structural features involved in oligomerization interfaces are likely to be conserved
- Predictions of oligomeric state can be validated against any prior experimental data (e.g. SAXS), biochemical data and/or phyologenetic information

Outputs

- Three-dimensional model of oligomer
- Protein sequence with residues involved in oligomerization highlighted
- Visual representation in standard molecular viewing software with interface residues highlighted

Tools required

- Parsing protein structure and sequence
- Single and multiple sequence alignment
- Combined sequence/structure alignment
- SAXS data analysis
    - Calculation of standard distributions
    - Comparison of distributions to those calculated from 3D models
    - Searching of known structures for similar SAXS curves

- Protein structure writing
- Sequence display
- 3D structure display

Knowledgebase context

- Provides linkage to and automatic retrieval of related structures in the Protein Data Bank
- Performs complex sequence and sequence/structure analysis without detailed user learning
- Cross validates against other experimental data within the Knowledgebase and in other outside resources
- Displays results in easy to understand visual forms and for download and subsequent analysis

3. Locating and visualizing a cellular compartment

Goal: To locate (segment) and visualize one or more cellular compartments (e.g. mitochondria) in a microbe.

Inputs

- Three-dimensional reconstruction of one of microbes of interest (e.g. from EM-tomography or soft X-ray tomography)
- Characteristics describing the compartment of interest (e.g. shape, density, proximity to other features), or a human-generated training set
- Optionally a visual label indentifying the compartment of interest

Process

- Read 3D data
- Perform pattern matching analysis to identify likely compartments on the basis of input data
- Segment volume data to assign the identity of compartments (note that for some data, it is possible to *a priori* segment on the basis of density, but the problem of identifying compartments still remains)
- Calculate statistics (e.g. volume of cell occupied by compartment, standard deviations between samples)
- Cross validate against any other relevant biochemical data

Outputs

- Statistics of compartments segmented
- Visual representation in volume rendering viewing software with compartments highlighted

Tools required

- Parsing large 3D volume datasets
- Pattern matching algorithms to identify compartments
- 3D volumetric data display

Knowledgebase context

- Performs segmentation analysis without detailed user learning
- Cross validates against other experimental data within the Knowledgebase and in other outside resources
- Displays results in easy to understand visual forms and for download and subsequent analysis

Workflow 6: Imaging Bioinformatics

Summary

One of the major advantages of phenotypic characterization through microscopy is the ability to visualize cellular organization, morphology and ultrastructure, and localization. More importantly, microscopic imaging allows cell-by-cell measurements, revealing a cellular heterogeneity that is often lost when using OMIC data only. For example, *Desulfovibrio vulgaris* (Dv) is known to form micro-colonies at certain stages of development due to cell-cell communication, a complex mechanism that remains largely unknown. Such population phenotypes can then be interrogated at multiple scales through multiplexed imaging probes to identify changes in structure, morphology, and localization on a cell-by-cell basis. These morphometric features can then be linked to omic data to query molecular predictors of a specific phenotypic subset. The main challenge in managing image-based data is identifying a quantitative view for each assay, which can be integrated with omic data. These quantitative views are often represented as vectors and relationships between vectors. Figure 1 is an example of a typical workflow in *Imaging Bioinformatics*.

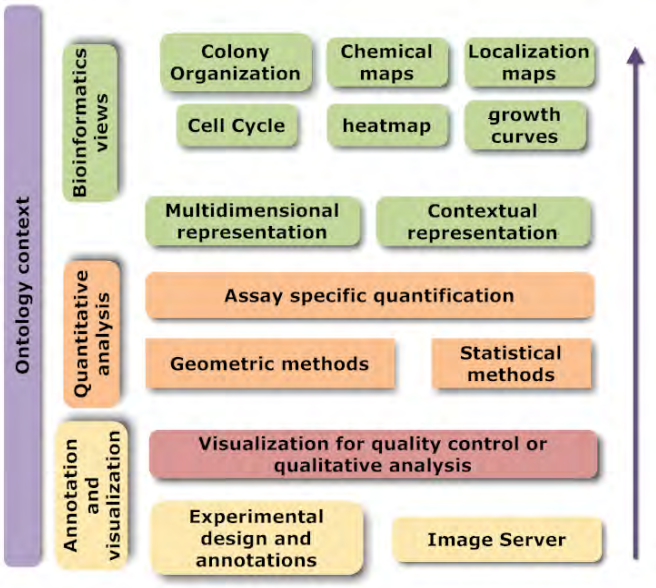


Figure 1. A typical workflow for image-based applications represents potential layers for image-based registration, quantitative analysis, and bioinformatics representation. Integration with omic data should take place through an ontology layer that provides a common set of vocabulary and ontology for exchanging bioinformatics views and their corresponding quantitative results.

Input Data

The input data consists of four types of information: (i) experimental design variables, (ii) imaging system parameters, (iii) raw image files, and (iv) queries used to target specific endpoints. (i) Experimental design refers to the model system, stress conditions, harvest time, imaging assay (e.g., labeling), etc. The main challenge has been reducing the number of user interactions needed to specify experimental design variables, since one rarely enters metadata at the granularity level that is often needed. There are no standards for capturing experimental variables; however, the microarray community has defined a complete protocol that can be leveraged. (ii) Most modern microscopes capture instrument setup information (e.g., optical path, illumination source) and store it as a header (e.g., in the form of a TIFF header) with raw data. Nevertheless, the Open Microscopy Environment (OME) has defined a schema for

specifying instrument configurations, and some vendors plan to support the proposed schema. (iii) Raw data are usually stored in a binary form, and the format varies between different vendors. However, LOCI and OME developed a transformer that can read any image format, parse it, and store it in a five-dimensional format. The end result is a homogenized representation of a diverse file format. (iv) The endpoints or biological queries have to provide a series of templates for guiding quantitative analyses. One can design a taxonomy that allows users to select from multiple templates.

Quantitative Analysis

Analytical requirements for image-based data are quite heterogeneous, and any computational pipeline must be extensible for new application software programs. However, common computational modules can be defined, integrated, and enhanced for a specific application. Nevertheless, there has to be a balance between excessive generalization versus specificity, as too much generalization adds to the complexity of a system and thus increases the learning curve required to use it efficiently. In general, image-based data analysis needs to incorporate a model to recover objects of interest in a robust fashion. Such a model can be expressed either geometrically or statistically. In some cases, model-free methods can be used, at a low level, to aggregate rich tokens for higher-level analysis. Once the images have been quantified, information can be composed and aggregated to form bioinformatics views (see "Output Data" below for more information). With respect to image analysis, the ITK image library provides a rich set of software and an extensible framework for adding new applications. However, it requires expertise in advanced software engineering, which may not be readily available at every institution.

Output Data

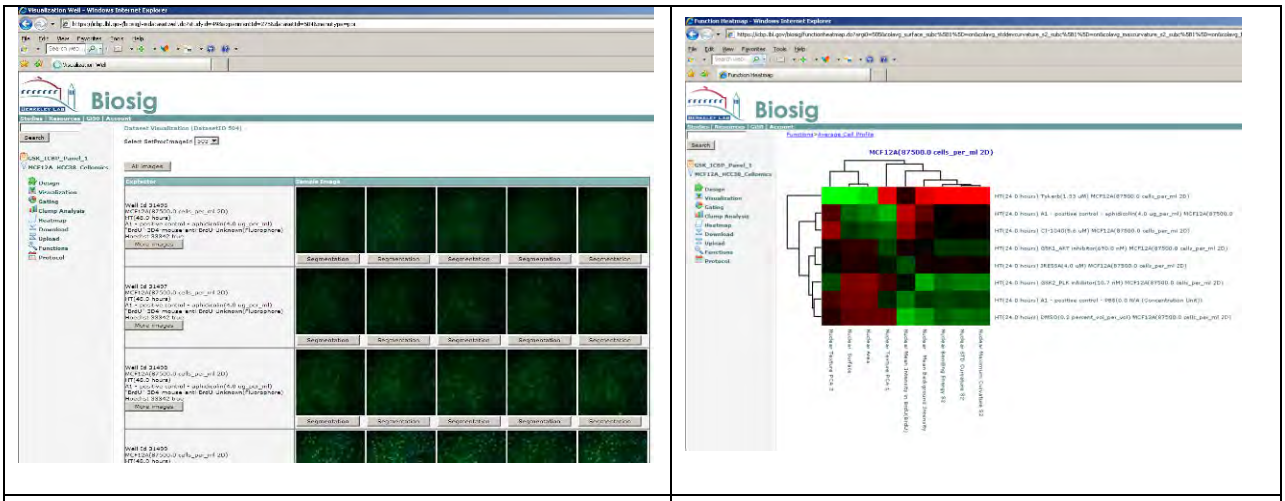
One of the characteristics of image-based assays is that a large number of data are often transformed into a very small amount of data. This is referred to as "bioinformatics views," which are often constructed by downloading computed information, and then processed further by using one of many statistical or data analysis stand-alone software packages. However, it is possible to integrate some basic capabilities into the bioinformatics platform. Examples include a dose-response curve, a growth curve, and co-localization frequencies. One of the advantages of imaging is that it maps cellular localization (or co-localization), chemical composition, and morphometric properties.

Current State of the Art

BioSig (ribo.lbl.gov:8080/biosig/home.do) is an example of an imaging bioinformatics system, which is being used for mammalian systems. BioSig builds on OME for image harmonization, leverages MIAMI (www.mged.org) standards for specifying experimental design variables, and has defined a number of tagged templates for assay-specific quantitative analysis. It also supports a schema for multidimensional profiling of cell-based assays for high-content screening, as shown in Figure 2.

The Next Generation

Current imaging bioinformatics platforms lack (i) an ontology and controlled vocabulary for microorganisms of interest to DOE, (ii) an integrated pipeline for bioinformatics and image analysis, (iii) an interface for integrating omic data, and (iv) the necessary analysis tools for mapping at multiple scales of different imaging modalities. The latter is quite important since it enables chemical mapping (e.g., Raman microscopy), localization mapping (e.g., electron or optical microscopy), and mass spectrometry imaging (e.g., MALDI imaging). Furthermore, having created these maps at multiple scales, one is also interested in correlative analysis between these imaging modalities for the model systems of interest under specific environment conditions. A potential correlative query would be how the chemical composition of the plant cell wall, visualized and quantified with Raman, is altered as a result of increase in biomass that is imaged with electron microscopy. In short, the next generation of breakthroughs in quantitative image analysis and imaging bioinformatics resides at the interface of different imaging modalities, and their integration with omic data.



(a) (b)

Figure 2. Imaging bioinformatics views. (a) Thumbnail visualization enables comparison of biological replicates (columns) for each set of experimental variables (rows). Each thumbnail is a hyperlink to a full- resolution version, where quantitative results can be overlaid on top of it. (b) Images in (a) are processed, and each cell is represented by multidimensional features. The user can select a subset of computed features, put them in a particular order, and view them through a heatmap. As a result, multiple phenotypic representations can be viewed simultaneously and compared in the context of experimental variables.

Section III: Strawman Knowledgebase Architecture

The preliminary diagram below was developed as a result of discussions held in conjunction with this workshop. Though this schematic will be refined in upcoming discussions, it is included in this report to indicate how the workflows (research protocols) relate to the ultimate system architecture. The workflows being developed by experimentalists to satisfy scientific objectives are critical to the development of many Knowledgebase architecture layers, such as data repositories (red), computing workflow management, and output visualization design.

The workflows provide information on data sources and types that must be accommodated by the Knowledgebase architecture. In-depth discussions will result in refinement of the workflows by the research and computing communities.

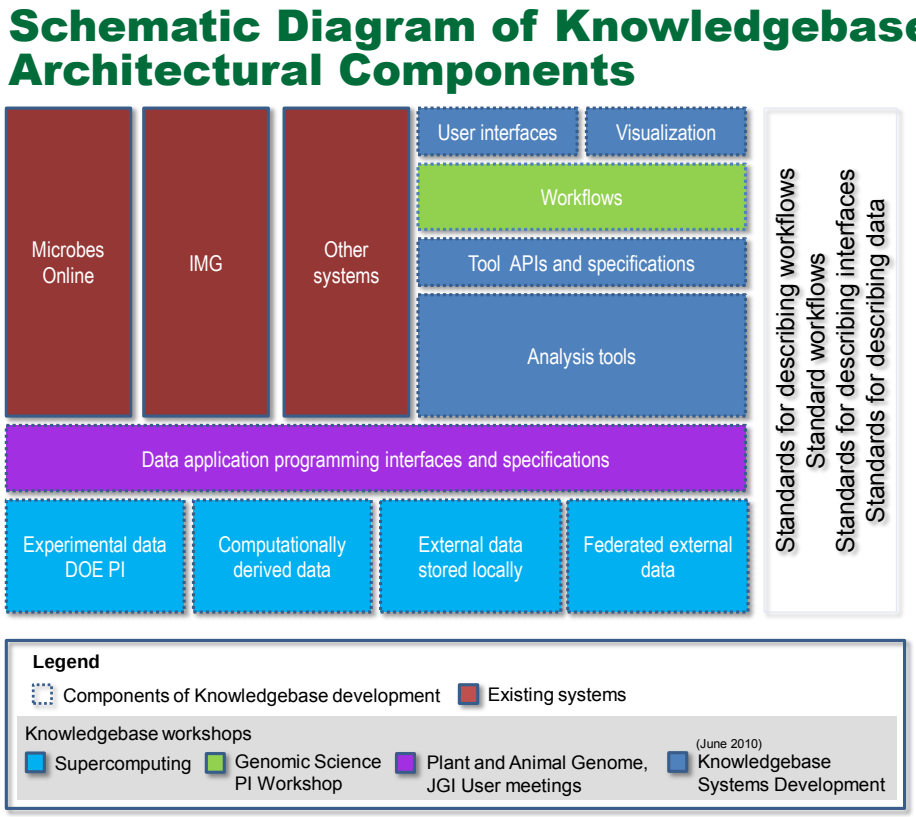


Fig. 3.1. Example Schematic of Knowledgebase Architectural Components.

Data

The data layer represented in the bottom of Fig. 2 illustrates several data components that will be important to achieving the goals of the Knowledgebase project. These data components will utilize several technologies. Relational database technology such as Oracle or MySQL will be used to manage data that is well structured and suited to relational technologies. Examples include the storage of account information, user configurations, and certain data and tool-related metadata. More recently developed technologies for representing data, such as Semantic Web, will be used for biological data with complex data model characteristics.

An important component of the data layer is data available from other sites that are remotely accessed and used singularly or in a federated manner. Federating external data sources will make heavy use of web services technologies. Web services are newer technologies allowing interoperability between software systems located at distinct sites. Federation will allow us to leave stable data at remote sites (i.e., NCBI Taxonomy) when a façade (wrapper, adaptor, bridge, etc.) can be constructed around the access routines provided by the remote site. The façade will serve to standardize access to data provided by multiple, distinct remote data sources.

Experimental data derived from DOE-funded work that is not available in other data sources in a suitable format will be structured and shared appropriately as part of the data layer. This data generally is thought of as the results of experiments funded by DOE. The data should not be limited to DOE-funded work; if others outside DOE wish to contribute, all the better.

The data layer also will contain data that exists remotely but is aggregated locally. Local aggregation can enhance data usefulness by putting the data in a modified format that corrects for missing metadata, incompatible formatting, or because internal computation integrates additional data. Pathway data, genome data, transcriptome data, and regulatory network data all stored in a suitable form for mash-ups are examples of data that likely will be found in the data layer component that represents locally aggregated external data. Another example of why external data is aggregated locally is because there will be external published data of use and specific data derived from DOE-funded work that is not available in the public domain. Other examples can be driven simply by the fact that computations such as similarity searching require local data sources for performance reasons.

Computationally derived data should represent another component of the data layer. Computations often can produce entirely new datasets rather than just adding value to existing ones. These computations may operate on existing datasets but generally produce a new type of data. For example, a computation on RNA sequencing–based gene expression data might produce a histogram of coverage statistics. This histogram is a new data type linked to the RNA sequencing data through descriptive metadata technology.

Analysis

The analysis component of the architecture will allow for development of both libraries and interfaces that promote the integration of analytical tools into the recognizable Knowledgebase. This component also will provide the facilities needed by the community to develop new algorithms and applications enabled by Knowledgebase infrastructure and data

access layers. Goals of the Knowledgebase project are to build and promote an environment where analysis tools are primarily derived from open contributions.

In direct support of the data layer, semantic-enabled search algorithms will allow metadata—information about the data stored in the data layer—to be more than just an attached publication or protocol list. Metadata is extremely valuable when making new scientific discoveries. Representing, integrating, searching, and performing logic on metadata can be challenging enough when the object being measured is sequence or crystal structure. Complex and conditional data derived from functional measurement of molecules, cells, and communities will make this an exceptional challenge. The rapid development of new technologies for making such measurements further increases the need to track the key information about experimental and analytical protocols for producing data and the processing applied to data before it is stored in accessible formats. A key goal for this effort will be developing tools that attach such information to data as easily as possible, identify the most important pieces of this data for scientific purposes and searching, capture experimental design and goals, and allow queries of this information.

Existing and New Systems and Projects not Developed by the Knowledgebase

Existing systems such as MicrobesOnline, the collection of IMG systems, the RAST systems, and others are expected to continue and benefit from the centralized or virtually centralized (federated) data stores and from direct programmatic access to the open methods developed as part of the Knowledgebase. We also anticipate that new systems will emerge.

It is expected that existing system developers can and will create application programming interfaces (APIs) to their systems and publish the specifications of those interfaces as part of Knowledgebase API specifications. These API specifications are analogous to the Sun Java Docs for the Java APIs. These interfaces may be used by other existing system developers or by the Knowledgebase development community.

As new systems emerge, embracing and nurturing them will be important. Guiding such projects so that they become important components of the Knowledgebase also will be necessary.

Workflows

Scientific workflows can help scientists, analysts, and computer programmers create, execute, and share experimental and analytical processes. These workflows can be captured as free text use cases or more formally represented using workflow languages. Regardless of whether a workflow is captured in a structured or unstructured manner, an important part of the Knowledgebase system architecture will be a graphical user interface that is available to the community so that anyone can access existing workflows and develop new ones.

User Interactions

The user experience will primarily take place through what is known as a horizontal web portal. These portals deliver an integrated front end to what is commonly thought of as several independent websites that allow users to easily search, visualize, and run analytical software on Knowledgebase information. Standard browsers, plugins, and web portal technology will enhance the user experience when command line or other existing user interfaces are not

suitable. This will allow members of the research community to have a customizable entrance to the Knowledgebase.

User interactions can be thought of in two parts—user interfaces and visualization—because these two components use markedly different technologies. For an effective Knowledgebase user experience, we will need to focus on both the user interfaces and the more challenging aspects of scientific data visualization.

Relying heavily on web technologies such as HTML, Javascript, and their derivatives, user interfaces allow a user to navigate the system, configure analysis environments, input data into the system, and retrieve results from it. These user interfaces can leverage the latest advancements in social networking to provide tools to the community for shared annotation and quality assessment and provide forums that easily reference Knowledgebase information.

Visualization (referred to as scientific visualization in some communities) relies heavily on graphics packages. The goal is to present data in a form that is useful for scientific discovery. There may be no interaction required when a visual representation of data is generated and presented.

Standards

Standards will be instrumental in achieving many aspects of the Knowledgebase project. These aspects range from scientific to engineering. From a scientific perspective, describing biological data and the relationships between data in a standardized way is critical to advancing our ability to interpret it. In relation to engineering, standards will enable healthy, continued evolution and growth of the system.

Several objectives will provide efficient and necessary utilization of standards. These objectives include embracing community standards when they are adequate, engaging in the community-development process of a particular standard when there is an existing standard that might be considered inadequate in its current form, and helping the community by initiating standards development where gaps exist.

Although standards for describing data and workflows will be critical, other types of standards will be important as well. Having community standards for data sharing is just one example of what will have to be supported in the Knowledgebase project. Another such example is developing standard workflows and benchmarking data that can be used by the community to facilitate a higher level of exchange among scientists.

In support of an open environment, standards for describing analytical tools, software libraries, data schemas, and other technical artifacts used to build the Knowledgebase will be essential for broad acceptance and use. Software tools and libraries implemented in the Java programming language benefit from a Java community–accepted standard on how to describe APIs. Requiring the use of these standards in code libraries will result in a solid documentation base that is needed for general acceptance and further use of the library by the community. Other programming languages such as Perl have similar standards.

Section IV: Workshop Summary and Conclusions

Workshop participants discussed the need for some level of individual research privacy, which could be achieved with user accounts. Data and code could be held in private, and analyses conducted in a nonpublic environment. The Knowledgebase also will need to allow users to track version history and provenance so that new analyses can be usefully compared with previous ones. Other important capabilities workshop participants discussed include:

- Curation not only of data, but also of models and representation of scientific concepts
- Comparison and analysis of methods and results over time
- Simulation, including the ability to modify and improve models
- Predictions based on simulation and analysis to form new hypotheses
- Comparison of predictions and results to guide experimental design

Only a few researchers today have comprehensive access to such computational capabilities, yet these tools are necessary to conduct research that will lead to important scientific innovations in energy and environment.

Also envisioned for the Knowledgebase are high standards for usability, understandability, discovery, and contribution. System design should be intuitive so that researchers can use it with minimal training. Knowledgebase components also need to be understandable. Although able to use a given software package, many people often do not understand the process by which the software derived its results (e.g., BLAST). Understandability implies that there is a good foundational basis for knowing that results returned to a user are based on robust scientific knowledge or assumptions. If results are not understandable, system features should allow the user to drill down to acquire information about how results were obtained. The Knowledgebase also should promote an environment of discovery, leading to new rounds of experiments or lines of research. Finally, engaging the entire research community in Knowledgebase contribution is critical. Any system being used by scientists ultimately should be measured on how well it accomplishes these concepts, advances research, and supports the scientific method.

Future Considerations for Workflow Definitions. Here we see a range of styles and level of content in the workflows. For the future final report of the Knowledgebase R&D Project, we will need to settle on a style. The Structural Biology workflow is very terse when compared with the others, but it is also very clear. In developing a standard for future workflows, this should be considered. An important question to raise: Do these workflows provide sufficient detail to allow requirements to be established that can drive the Knowledgebase Implementation Plan, and if not, how much more detail is needed?

Appendix 1: Agenda

DOE Genomic Science Microbial Systems Biology Knowledgebase Workshop

Crystal City, Virginia
Tuesday, February 9, and Wednesday, February 10, 2010

February 9

2:00 – 2:30 p.m.	Robert Cottingham, Oak Ridge National Laboratory "Microbial Systems Biology Knowledgebase: Scientific Objectives and Current Prospects"
	Focus on examples of scientific objectives, benefits, and outcomes
2:30 – 3:00 p.m.	Discussion
3:00 – 3:30 p.m.	Robert Kelly, North Carolina State University "Near-Term Prospects for Functional Microbial Genomics: Moving Beyond the Monoculture Paradigm"
	One organism to two organisms, adding complexity
3:30 – 4:00 p.m.	Discussion
4:00 – 4:30 p.m.	Adam Arkin, Lawrence Berkeley National Laboratory "From Pathways to Populations and Back Again: Long-Term Prospects for the Microbial Systems Biology Knowledgebase"
	Much larger complexity of systems, data, models, and impacts
4:30 – 5:00 p.m.	Discussion
5:30 p.m.	Adjourn

February 10

1:00 – 3:00 p.m.	Impromptu follow-up session focusing on workflows

Appendix 2: Participants and Observers

Adam Arkin (LBNL)
Nitin Baliga (Institute for Systems Biology)
Jill Banfield (University of California, Berkeley)
Chris Bare (Institute for Systems Biology)
Ben Bowen (LBNL)
Tom Brettin (ORNL)
William Cannon (PNNL)
James Cole (Michigan State University)
Robert Cottingham (ORNL)
Brian Davison (ORNL)
Mitch Doktycz (ORNL)
Ronan Fleming (University of Iceland)
Cheri Foust (ORNL)
Hector Garcia Martin (LBNL/JBEI)
George Garrity (Michigan State University)
Adam Godzik (Burnham Institute)
Susan Gregurick (DOE)
Loren Hauser (ORNL)
Alyssa Henning (Cornell University)
Kimberly Keller (University of Missouri)
Robert Kelly (North Carolina State University)
Julia Krushkal (University of Tennessee, Memphis)
Libbie Linton (Utah State University)
Yukari Maezato (University of Nebraska)
Betty Mansfield (ORNL)
Victor Markowitz (LBNL)
Lee Ann McCue (PNNL)
Folker Meyer (ANL)
Jonathan Millen (University of Rochester)
Aindrila Mukhopadhyay (LBNL)
Kristen Munch (NREL)
Ambarish Nag (NREL)
Chongle Pan (ORNL)
Nicolai Panikov (Northeastern University)
Morey Parang (ORNL)
Charles Parker (Names for Life, LLC)
Bahram Parvin (University of California)
Amanda Petrus (University of Connecticut)
Madeleine Pincu (University of California, Irvine)
David Pletcher (JBEI/LBNL)
Iris Porat (ORNL)
David Reiss (Institute for Systems Biology)
Dmitry Rodionov (Burnham Institute)
Blake Simmons (LBNL)
Steve Singer (LBNL)
Marvin Stodolsky (DOE)
Ines Thiele (University of Iceland)
Judy Wall (University of Missouri)
Sharlene Weatherwax (DOE)
Steven Wiley (PNNL)
Jian Yin (PNNL)

Acronyms

ANL	Argonne National Laboratory
DOE	U.S. Department of Energy
JBEI	Joint BioEnergy Institute
LBNL	Lawrence Berkeley National Laboratory
NREL	National Renewable Energy Laboratory
ORNL	Oak Ridge National Laboratory
PNNL	Pacific Northwest National Laboratory