Appendix 2
Use Case Scenarios of Systems Biology Investigations Utilizing the GTL Knowledgebase

The workshop systems biology group identified a set of high-priority scientific and engineering research examples based on the analysis of priority DOE applications, unmet needs, and feasibility. Each of these use case scenarios was selected to match projected phases of GTL Knowledgebase (GKB) development.

Use Case Scenario 1

Use Case 1 has two objectives:

- Support the capability to rapidly assess the metabolic potential and regulatory features of any culturable or sequenced prokaryote of primary importance or relevance for all DOE GTL focus areas.
- Map parts (genes) and modules (pathways, subsystems, and regulons) that constitute the core of life across thousands of diverse species within Use Cases 1 and 3.

Identification and accurate functional assignment of genes involved in the key cellular processes of any organism with a completely sequenced genome would allow assessment of the organism's metabolic and regulatory capabilities with respect to their applications. This information would provide a foundation for further detailed reconstruction and modeling and allow assessment of the organism's role and interactions within the community. Although many components of the required workflow (including a substantial body of annotated genomes and tools) already exist in the public domain, a considerable effort would be required to automate and scale up the process and, at the same time, maintain and improve coverage, quality, and consistency of annotations.

Understanding and integrating thousands of diverse genomes and the associated nongenomic information—inferences (gene annotations, subsystems and pathways, and regulons) within a framework (tools)—are critical for their assessment and comparative analysis. Although the microbial sequencing projects throughout the world have created a rich, diverse collection of microbial genomes, strong biases are evident in what has been sequenced thus far. Use Case 1 would be an extension of ongoing work seeking to understand related species, as outlined in the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project, which is aimed at systematically filling in sequencing gaps along the bacterial and archaeal branches of the tree of life. Sequencing large numbers of diverse microbial genomes has been a mission focus of DOE for several years. A comprehensive goal would be to sequence a representative of every species discussed in Berghey's manual. The GEBA project is a piloted, modest sequencing effort along this line. This project represents a new paradigm in using the tree of life as a guide to sequencing the target selection. Supporting this level of analysis, which we term "draft reconstruction" or stage 1 analysis, for thousands of diverse genomes—regardless of the current perception of their immediate importance for DOE applications—is crucial for creating several resources:

- Rich pool of diverse metabolic and other features for future applications, not all of which can be foreseen (e.g., new engineering needs and synthetic biology).
- Comparative framework for comprehensive and accurate functional annotation of application-related organisms, including complex eukaryotes.
- Reference set of sequences for metagenomic data analysis.

Issues and Requirements

The analysis under consideration would be coordinated with ongoing efforts in gene sequencing and annotation at the DOE Joint Genome Institute (JGI). The assessment level would include the following critical aspects, which require robust informatics support either not provided or only partially covered by existing efforts.

- Accurate gene-function assignments (annotations) supported by various lines of evidence, including homology-based projection from experimentally characterized genes; delineation of structural domains and conserved motifs; genomic context (conserved operons, regulons, and phylogenetic occurrence); phylogenetic profiling; and functional context (role in a relevant pathway, subsystem, or complex).
- Functional predictions for previously uncharacterized gene families. For example, gene candidates would be proposed to fill in gaps in known pathways. A critical investigation into experimental testing of functional inferences also must be considered high priority.
- A controlled vocabulary (for function definition) and connection with a collection of analyzed reactions and metabolites for consistent propagation of annotations and their further use for reconstruction and modeling.
- Curation of gene annotations and pathways.

The genomic reconstruction of regulons (for a recent review and many examples of references therein, see Rodionov 2007) includes identification and capturing of DNA and RNA regulatory motifs (e.g., promoters, operators, attenuators, and riboswitches), along with respective regulatory factors (e.g., transcription factors, regulated genes, and effectors).

Transcriptomic and proteomic data that would become available for some organisms should be minimally analyzed to extract information about which genes are expressed (and under what conditions), which proteins are produced, and protein features such as localization and post-translational modifications.

Typical uses of a knowledgebase for supporting rapid assessment of metabolic and regulatory features include the following:

- **Target:** Organism (or group of related organisms) with completely sequenced genomes. **Use:** Reconstruct a chosen aspect of metabolism and infer some phenotypic properties for further testing. Example: Reconstruct carbohydrate utilization machinery in *Shewanella* spp., including prediction and verification of novel genes, pathways, and physiological properties.
- **Target:** Group of related organisms with completely sequenced genomes. **Use:** Reconstruct a substantial fraction of metabolic regulons.
- **Target:** Set of desired metabolic and regulatory properties important for a certain class of applications (e.g., in bioenergy). **Use:** Identify a group of optimal candidate species.

- **Target:** Defined metabolic function (e.g., enzymatic reaction). **Use:** Identify known and putative candidate genes performing this and other related functions in all organisms, including supportive evidence (experimental, homologous, genomic, and functional context).
- **Target:** Selected organism. **Use:** Automatically compute a list of proteins (enzymes and transporters), inferred reactions, and metabolites for use as a first step to building a metabolic model.
- **Target:** List of genes (proteins) in a target organism. **Use:** Provide all associated information and features including functional assignments from various sources and evidence; association with protein families (phylogenetic profiles); multiple alignments and phylogenetic trees for each family; domains, motifs, and structural features (known or predicted); genomic context (operons and regulons); functional context (associated pathways and subsystems); gene expression data (chosen from integrated or uploaded datasets); proteomic data; associated reactions, metabolites; and other types of data connecting to specific genes.
- **Target:** Experimental "omic" data (e.g., gene expression). **Use**: Identify clusters (lists) of functionally coupled genes (e.g., stimulons), retrieve their properties as described above, and support a detailed correlational analysis (e.g., assess correlations between gene expression and pathways or gene expression and protein levels).

Use Case Scenario 2

Use Case 2 has one objective:

- Support the capability to predict and simulate microbial behavior and response to changing environmental or process-related conditions for target sets of prokaryotic species.

Implementation of this objective would require all components employed in Use Case 1, but additional data would be needed for detailed predictive modeling. These capabilities would be acquired and integrated to allow for the development of computational models, so all relevant data and inferences would be provided in a format for modeling. Among required capabilities of this use case would be the framework and tools that support iterative improvement of models through comparison of model predictions and experimental data. Approaches would be needed to identify inconsistencies between model predictions and experimental observations and to automatically generate hypotheses that would resolve inconsistencies through novel components or component interactions. An example case would be to provide guidelines for organism engineering and even de novo design for tasks both fundamental (e.g., proof of understanding and novel model systems) and applied (e.g., metabolic engineering for biofuel production).

Issues and Requirements

I. Detailed Reconstruction and Modeling of Metabolic Networks

Deliverables described above in Use Case 1 would provide a foundation (i.e., "draft reconstruction") from which detailed models for any species or groups of species selected for specific applications can be developed. The workflow in this case would require additional manual effort aimed to fill in gaps in the metabolic reaction networks, reconcile inconsistencies, and account for published legacy data and accumulated additional (omic

and phenotypic) data; iterative gap filling; model simulations; experimental validation and refinement of the model; and final application to address scientific and engineering objectives. Such detailed analysis would be a focused effort applied to dozens rather than thousands of species. It would proceed in concert with the experimental work, including systematic generation of large volumes of postgenomic data.

Typical uses of a knowledgebase to predict and simulate microbial behavior and response to changing environmental or process-related conditions include the following:

- Quantitative assessment of metabolic capabilities (e.g., production yield or metabolic flux distribution in the network) for a selected organism as a function of environmental and growth conditions.
- Support of optimization tasks (e.g., fermentation conditions optimized for production).
- Guidelines for rational engineering of industrial organisms, including assessment of the potential for directed evolution.

Scope and Requirements

This analysis, which already has been successfully prototyped, would require a level and type of informatics support not currently provided by existing public domain resources. It would aim to cover the following:

- Detailed metabolic and genetic network definition (genes, proteins, roles, reactions, and compounds).
- Stoichiometric matrices and sets of constraints for modeling.
- Predictive computational models and modeling tools.
- Inferred fluxes, phenotypes, growth and application-related properties.
- Experimental validation of models.

Example Uses Employing Complex Queries from the Integrated GKB

- **Target:** Organism with a completely sequenced genome, draft reconstruction (Use Case 1), and collection of additional data (biomass composition, medium composition, and growth characteristics). **Use:** Build a detailed and consistent metabolic reconstruction.
- **Target:** Detailed metabolic reconstruction. **Use:** Apply modeling tools (e.g., flux-balance analysis) to test whether the model is consistent with known physiological properties and growth characteristics; refine the model.
- **Target:** Validated metabolic model. **Use:** Address application and optimization tasks. For example, estimate the maximal yield of the desired product or optimize the growth medium.
- **Target:** Validated metabolic model. **Use:** Predict which genes are essential and dispensable under given growth conditions (e.g., as a way to test the model or to support engineering goals).
- **Target:** Validated metabolic models of two or more organisms. **Use:** Compare their metabolic capabilities with respect to a desired application.
- **Target:** Validated metabolic model. **Use:** Suggest re-engineering strategy (e.g., gene elimination, addition, deregulation, or amplification) to improve organism properties with respect to application.

II. Integrative Modeling of Transcriptional Regulatory Networks

Rationale (modified from Bonneau, Baliga, et al. 2007)

Rapid DNA sequencing technology has provided access to a large number of complete genome sequences from diverse and often poorly characterized organisms. Use of this information is expected to help engineer new biotechnological solutions to diverse problems spanning bioenergy and environmental remediation. In principle, re-engineering new processes by selectively combining otherwise distinct biochemical capabilities encoded in different genomes is a reasonable expectation. In reality, however, this will be possible only when we have a sophisticated understanding of how RNAs and proteins encoded in each individual genome dynamically assemble into biological circuits through interactions with the environment. Given that more than 500 genomes already have been sequenced and that little biological information exists for most of these organisms, a classical gene-by-gene approach is inefficient. Furthermore, since every organism is unique, it is impractical to rely on accumulated sets of known interactions from select model systems to construct really detailed models. A data-driven systems approach, on the other hand, is ideally suited to tackle this problem.

An important goal of applying systems approaches in biology is to understand how a simple genetic change or environmental perturbation influences the behavior of an organism at the molecular level and, ultimately, its phenotype. High-throughput technologies to interrogate the transcriptome, proteome, protein-protein, and protein-DNA interactions present a powerful toolkit to accomplish this goal (DeRisi, Iyer, and Brown 1997; Eichenberger et al. 2004; Laub et al. 2000; Liu, Zhou, et al. 2003; Masuda and Church 2003). However, each of these individual data types captures an incomplete picture of global cellular dynamics. Therefore, these data need to be integrated appropriately to formulate a model that can quantitatively predict how the environment interacts with cellular networks to effect changes in behavior (Facciotti et al. 2004; Faith et al. 2007; Kirschner 2005; Kitano 2002). Accurate prediction of quantitative behavior—the ultimate test of our understanding of a given system—will enable re-engineering of cellular circuits for specific applications relevant to DOE missions.

Scope and Approach

The goal would be to develop a framework that would enable the type of integrative analysis necessary to accomplish reconstruction of predictive models of cellular behavior. The approach (illustrated by the work of Bonneau, Baliga, et al. 2007) would involve genetically or environmentally perturbing the cells, characterizing their growth and survival phenotype, quantitatively measuring steady-state and dynamic changes in mRNAs, assimilating these changes into a network model that can recapitulate all observations, and experimentally validating hypotheses formulated from the model. This type of approach would require the integrated development and implementation of computational and experimental technologies (see Fig. A2.1. Systems Approach for Predictive Modeling of Cellular Responses, p. 70) and would comprise the following steps:

1. Sequence the genome and assign functions to genes by using comparative genomic approaches (for example, protein sequence and structural similarities).
2. Perturb cells by changing relative concentrations of environmental factors and gene knockouts.

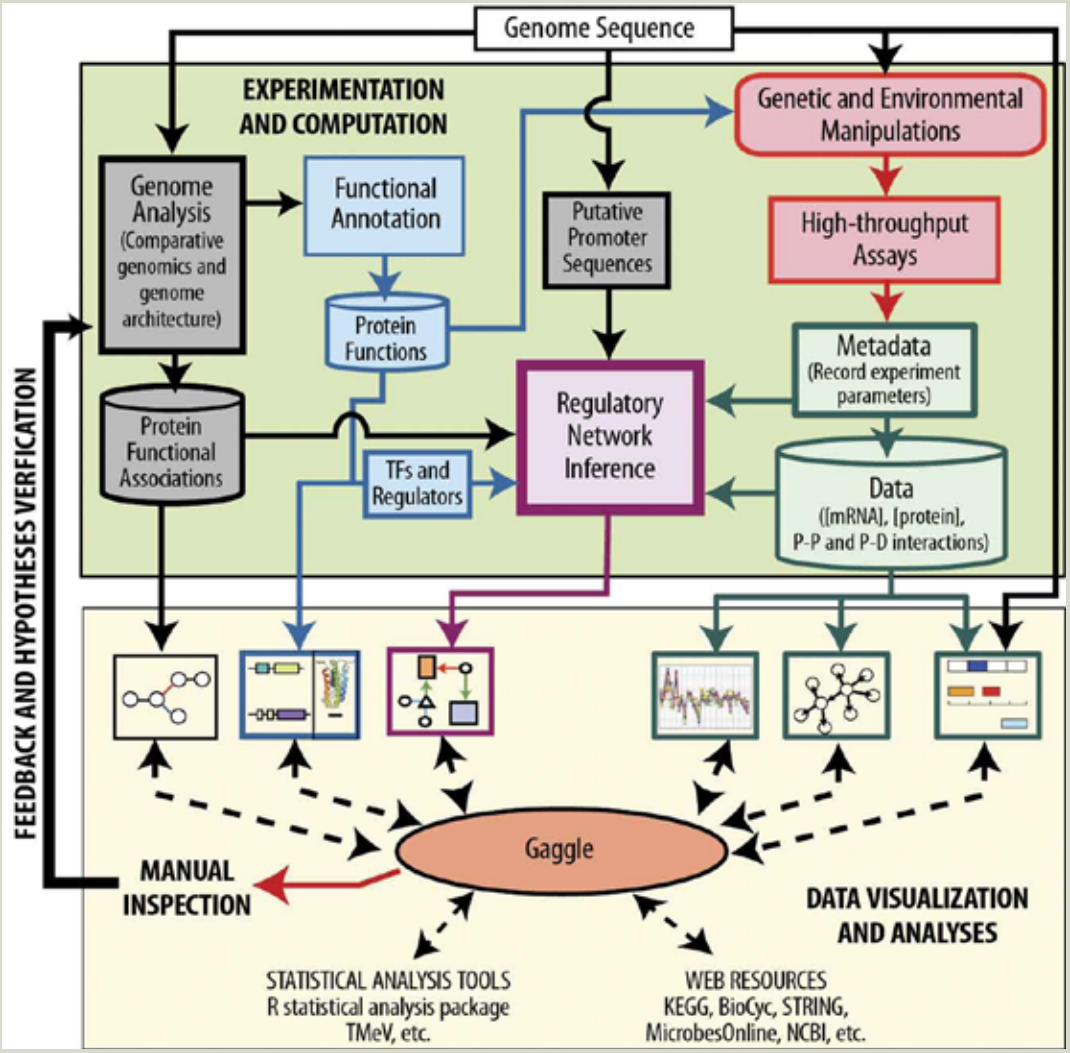

Fig. A2.1. Systems Approach for Predictive Modeling of Cellular Responses. After genome sequencing, two major interconnected and iterative components—experimentation and computation—are followed by data visualization and analyses. Within the first component, major efforts needed include computational genomic analyses for discovering functional associations among proteins (black boxes); putative functional assignment to proteins using sequence- and structure-based methods (blue boxes); and high-throughput microarray, proteomic, and ChIP-chip assays on genetically or environmentally perturbed strains (red boxes). All data from these approaches, along with associated records of experiment design (green boxes), are analyzed with network inference algorithms (purple box). The resulting model is explored with underlying raw data, using software visualization tools within a framework (yellow box) that enables seamless software interoperability and database integration. The interface should be extensible to provide a cost-effective interface to third-party tools and databases. This manual exploration and analysis enable hypothesis formulation and provide feedback for additional iterations of systems analyses. [Source: Adapted with permission from Elsevier. From Bonneau, R., N. Baliga, et al. 2007. "A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell," *Cell* **131**(7), 1354–65 (http://www.sciencedirect.com/science/journal/00928674).]

3. Measure the resulting dynamic and steady-state changes in gene expression, protein-protein interactions, protein-DNA interactions, and protein modifications. Archive these measurements along with digital metadata that capture the genetic and environmental context.

4. Integrate diverse data such as gene expression, evolutionarily conserved associations among proteins, metabolic pathways, and *cis*-regulatory motifs to reduce data complexity and identify subsets of genes that are coregulated in certain environments (biclusters) (Reiss, Baliga, and Bonneau 2006).

5. Using a machine-learning algorithm such as Inferelator, construct a dynamic network model to predict the influence of changes in environmental factors and transcription factors on the expression of coregulated genes (Bonneau et al. 2006).

6. Explore the network in a framework for data integration and software interoperability (Shannon et al. 2006) to formulate and then experimentally test hypotheses to drive additional iterations of steps 2 through 6.

From a practical standpoint, the following are the types of activities associated with the approach described above:

- Extract function information; microarray, proteomic, and metabolomic data; and physical interactions (protein-protein and protein-DNA) from different databases.
- Submit these data to one or more algorithms (written in different computational language environments: R, Matlab) that can infer operational relationships among the genes.
- Interactively visualize, explore, and analyze the inferred network model in the context of underlying raw data to gain biological insight and discover inconsistencies to drive new experiments.
- Record new insights and curate function information to propagate knowledge.

Unmet Technical Needs

The GTL Knowledgebase would need to address the following technical needs during Use Case 2 activities:

- Mapping schema across databases.
- Standardized normalization of data (e.g., quantitation by sequencing versus arrays, two channel versus one channel).
- Standardized statistical models that capture uncertainty.
- Meta-information concerning data collection (e.g., perturbation, growth parameters, and metadata genotype).
- User interface to algorithms to adjust parameters.
- Algorithms that accept standardized data formats and output in standardized formats compatible with visualization software.

Use Case Scenario 3

Use Case 3 has one objective:

- Expand Use Cases 1 and 2 toward key application-related aspects of microeukaryotes (e.g., fungi and algae) and plants.

Implementation of this use case would require a substantial increase in the volume of eukaryotic genomic data, and important issues pertaining to eukaryotes may have requirements that extend in scope and complexity beyond those for prokaryotes. Therefore, the GTL Knowledgebase development strategy for the eukaryote case would combine the following:

- Limiting the initial scope of modeling by key aspects of obvious applied value (e.g., primary and secondary metabolic pathways and selected categories of enzymes); extensively studied model organisms to train the tools; and several most important and tractable target organisms to address actual application issues.
- Providing the foundation (data and tools) for developing new research tools and modeling techniques to expand the initial scope's limit toward the behavior and responses of more complex systems and organisms.

This strategy would allow work to begin immediately on some priority application tasks associated with fungi and plants, in parallel with other developments.

- **Target (Near term):** Mission-relevant and model plants and microeukaryotes—parts inventory and draft metabolic reconstructions. Some fungi and microalgae: detailed reconstruction of core metabolism, key biochemical pathways (i.e., lipid biosynthesis), selected functional enzyme categories (i.e., cellulases), and regulatory circuits.
- **Target (Long term):** Reconstruction of secondary metabolism and identification of potential natural products and secondary metabolites. Development of whole-organism metabolic models. Understanding of molecular bases of plant-microbe and microeukaryote-bacteria interactions (symbiosis and mutualism); coupled organism models.

Use Case Scenario 4

Use Case 4 has one objective:

- Expand Use Cases 1 and 2 toward progressively more complex communities and ecosystems.

As Genomics:GTL science moves toward research involving complex organism interactions and eventually communities, the GTL Knowledgebase would need to evolve capabilities, many of which are difficult to predict at this early stage. Not only different modeling styles and approaches but different types of data would be applied as complexity increases in this objective. The following are some first steps that could be performed within the GKB to enable this area of science.

- **Target:** Well-defined mixed cultures of two or several model prokaryotic species. **Data and modeling needs:** Complete individual genomes, individual and combined metabolic and regulatory reconstructions, and models; obtain omic data for individual organisms and for a mixed culture; assess cross-species signaling, communication, and physical interactions; and obtain imaging data and physiological measurements.
- **Target:** Eukaryote-prokaryote interactions (such as plant-microbes). **Data and modeling needs:** Repeat as above with the addition of at least partial reconstruction of plant metabolism.
- **Target:** Phylogenetically complex microbial communities. **Data and modeling needs:** Obtain metagenomic data (including gene expression, proteomic, and metabolite); inferred metabolic potential; and metabolic coupling. Physiological measurements, including the assessment of carbon and nutrient flow (isotope labeling) in support of metabolic modeling, are particularly important.
- **Target:** Ecosystems and environments. **Data and modeling needs:** Repeat as above but at a substantially larger scale, including the analysis of multiple samples from different locations and measurements of various environmental parameters in conjunction with biological measurements.

Potential impacts of this objective range from fundamental understanding and the ability to simulate and predict global processes associated with carbon cycling and climate changes to the rational control of these global processes. At this stage, the main limiting factor appears to be the lack of modeling techniques adequately reflecting even simple mixed cultures.

Issues and Requirements

Metabolic properties of complex communities would need to be assessed by using metagenomic and related approaches. Significant scientific opportunities exist in an open research area in which many concepts and approaches to data analysis and visualization are yet to be developed. Among specific challenges are the following:

- Data often intrinsically incomplete and unstable.
- Bottom-up approach—importance of reference genomes to infer parts and modules.
- Binning of shotgun sequences to reconstruct phylogenetic groupings.
- Inferences from community composition (e.g., 16S RNA).
- Novel type of probabilistic and intentionally ambiguous assignments (e.g., *either this or this but not that*).
- Estimation of metabolic potential (as opposed to detailed reconstruction).
- Assessment of expressed metagenomes (RNA and proteins) and meta-metabolomes (metabolites).
- Importance of metadata describing samples and features of the environments from which they were collected.
- Remote capture and collection of spatially and temporally heterogeneous environmental and biological data to assess metabolic and biogeochemical processes.

Modeling a Fully Defined Microbial Community

Microbial community modeling is another open research area. Even if we had all the underlying data, we would not know how to model even the simplest and best-defined community (several species, for example). New approaches to modeling microbial interactions and the functions of communities should be developed and tested using simple (reduced) or reconstructed communities. Investigations of cocultures in bioreactors, where factors that limit or shift populations can be tightly controlled and responses monitored, would be a good starting point. Single-cell analyses (i.e., transcriptome, proteome, and investigation of metabolic interactions) also would contribute to focused model development on a system that is not overly complex.

Given the challenge of applying metagenomics to complex communities, nongenomic or targeted genomic approaches should be considered (i.e., process-level models that are genome informed). One example would be a top-down approach (e.g., experimental assessment of carbon flux, other biochemical measurements, and respective modeling techniques).

A long-term challenge will be to bridge the gap between a bottom-up approach, which would develop a metabolic model of an individual organism by using detailed genomic and metagenomic information, and the top-down approach, which would involve measuring and modeling coarse-grained processes such as nutrient and metabolic fluxes at macroscopic scales.

Illustration of Use Case Scenario 1: Integrated Approach to Reconstruction of Metabolic and Transcriptional Regulatory Networks in Bacteria

Comparative genomic and experimental techniques are combined to support the mapping of metabolic and regulatory networks in hundreds (soon to be thousands) of diverse species with completely sequenced genomes. Dissection of cellular processes to compact subsystems allows researchers to accurately project functions of individual genes (enzymes, transporters, and regulators) and elementary modules (pathways) from model organisms to those of others. The growing collection of curated subsystems in the SEED genomic integration (http://theseed.uchicago.edu/FIG/index.cgi; Overbeek et al. 2005) provides a framework supporting the inference of metabolic scenarios (DeJongh et al. 2007) and regulatory circuits (Rodionov 2007).

Comparative subsystem analysis reveals a remarkable picture of conservations and variations in the central machinery of life. Most notably, the observed metabolic diversity appears to emerge, by a combinatorial explosion, from a limited set of operational subsystem variants (stable gene patterns) implemented in different species (Ye et al. 2005).

Application of genome context analysis techniques (e.g., delineation of conserved operons and regulons in prokaryotes) allows researchers to predict previously uncharacterized genes and pathways (Osterman 2006; Osterman and Overbeek 2003; Osterman and Begley 2007). Focused validation experiments are used to test the most critical functional predictions. Among the recently published examples applying this approach are prediction and verification of the novel N-acetyglucosamine utilization pathway in *Shewanella oneidensis* (Yang et al. 2006); genomic discovery and reconstitution of a complete biosynthetic pathway for the osmolyte di-myo-inositol-phosphate in *Thermotoga maritima* (Rodionov et al. 2007); reconstruction of catabolic pathways driven by the novel glycerate-2-kinase in *T. maritima* and other bacteria (Yang et al. 2008); genomic reconstruction and validation of nicotinamide adenine dinucleotide (NAD) metabolism in *Synechocystis* sp. (Gerdes et al. 2006); genomic reconstruction and experimental validation of novel regulons controlling NAD biosynthesis in diverse bacteria including *S. oneidensis*, *Synechocystis* sp., and *T. maritima* (Rodionov et al. 2008a; Rodionov et al. 2008b); bioinformatic prediction and experimental characterization of the novel class of multicomponent vitamin transporters (Rodionov et al. 2009); and novel lactate utilization machinery predicted from *Shewanella* studies and experimentally validated (Pinchuk, Rodionov, Yang, et al. 2009).

Case Study 1: Genomic Encyclopedia of Carbohydrate Utilization

The integrated approach was applied to systematically map mono- and disaccharide catabolic pathways in 15 species of *Shewanella* with completely sequenced genomes. This project is a component of the *Shewanella* Federation studies aimed at better understanding the ecophysiology and speciation of this important genus.

A bioinformatic approach established in this study is briefly illustrated in Panel A of Fig. A2.2. Genomic Reconstruction and Experimental Testing of Sugar Utilization Pathways in *Shewanella* spp., p. 75. A set of ~480 protein families (FIGfams)—including transporters, transcriptional regulators, and several classes of metabolic enzymes (e.g., sugar kinases, oxidoreductases, epimerases, and aldolases) from a

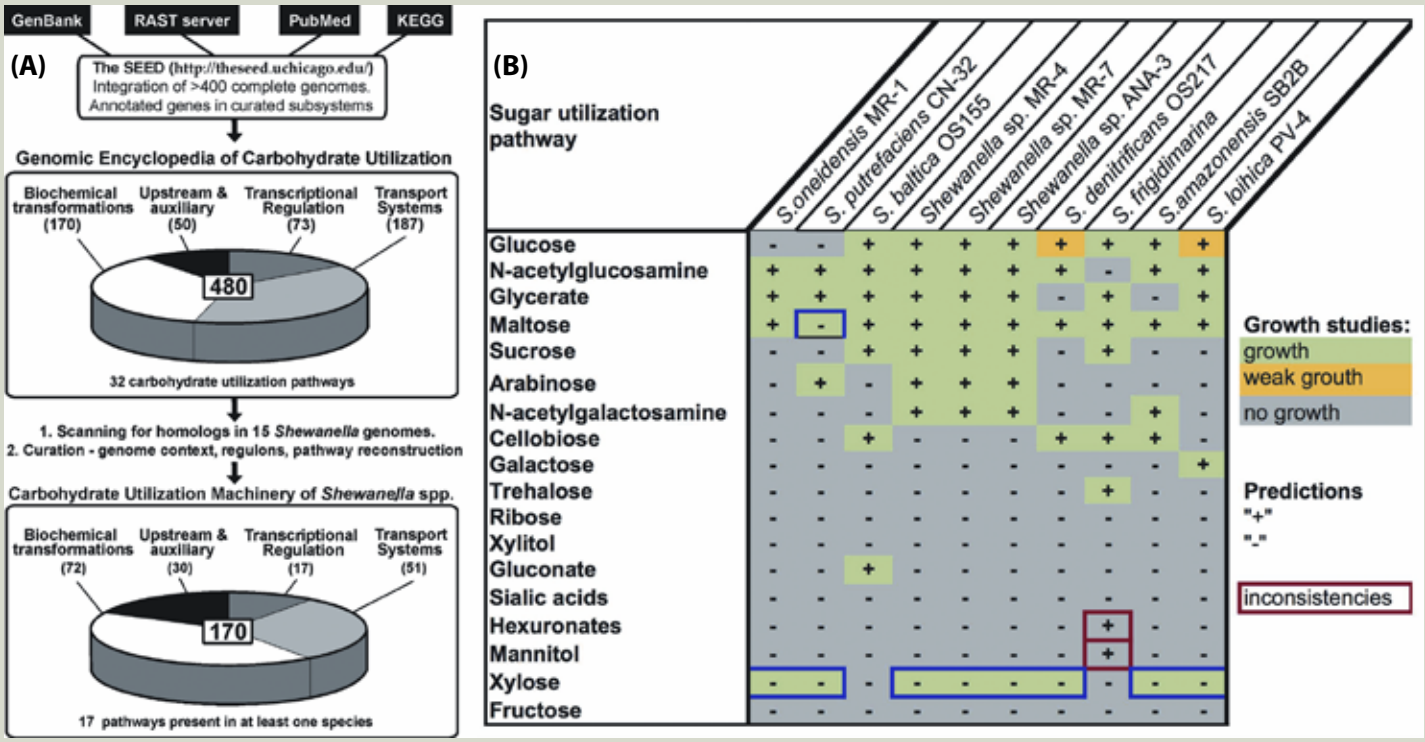

Fig. A2.2. Genomic Reconstruction and Experimental Testing of Sugar Utilization Pathways in *Shewanella* spp. (A) A subsystems-based approach. **(B)** Growth studies illustrated for a subset of 10 *Shewanella* spp. on 18 individual sugars (shown by color) in comparison with bioinformatic predictions (shown by + or –). A few observed inconsistencies are outlined by boxes. For example, many strains were able to grow on xylose, whereas no genes corresponding to any known version of xylose utilization pathways could be identified, suggesting a topic for further research. Some of these results were reported in Fredrickson et al. (2008) and Yang et al. (2006).

collection of carbohydrate utilization subsystems in the SEED—was used to scan for homologs in 15 *Shewanella* genomes. Identified candidate genes were the subject of further genome context analysis for their accurate functional assignment and reconstruction of respective pathways. This analysis detected substantial variations in a sugar diet among different *Shewanella* species, reflecting various aspects of their ecophysiology and evolutionary history. More striking, however, are the differences revealed by comparison with a classical model system of *Escherichia coli.* These differences are manifested at various levels, from the presence or absence of certain sugar catabolic pathways to a dramatically different organization of transcription regulatory networks in the central carbon metabolism. The results of this analysis included prediction of several novel variants of carbohydrate utilization pathways (e.g., for N-acetyglucosamine, sucrose, cellobiose, arabinose, and glycerate) and tentative functional assignments for previously uncharacterized gene families (e.g., within GlcNAc operon SO3503-3507 in *S. oneidensis*, sucrose operon Sfri3988-3991 in *S. frigidimarina*, and cellobiose operon Sbal0541-0545 in *S. baltica*). Overall, of the 170 identified protein families involved in sugar utilization pathways in various strains of *Shewanella*, 60 families may be considered novel because their specific functional assignments were previously unknown. Representatives of eight predicted families in four pathways were verified experimentally by genetic complementation and biochemical characterization of purified recombinant enzymes. The most important validation of inferred pathways, however, was obtained by experimental testing of predicted phenotypes by growth studies. A high level of consistency between predicted and observed phenotypes (the ability to grow on a panel of individual sugars)

is illustrated by the table in Panel B of Fig. A2.2, this page. In addition to the specific knowledge of carbohydrate catabolism in the *Shewanella* genus, this study led to a substantial expansion of the current version of the *Genomic Encyclopedia of Carbohydrate Utilization*. A systematic iterative application of this approach to multiple taxonomic groups of bacteria will further enhance this knowledgebase, providing adequate support for efficient analyses of newly sequenced genomes as well as of emerging metagenomic data.

Case Study 2: Genomic Reconstruction of Metabolic Regulons in *Shewanella* spp.

An integrative comparative genomic (knowledge-driven) approach was used to infer transcriptional regulatory networks (TRN) in 13 species of *Shewanella* spp. To accomplish this goal, we combined the identification of transcription factors (TF), TF-binding sites (TFBS), and cross-genome comparison of regulons with the analysis of the genomic and functional context inferred by metabolic reconstruction (see Panels A and B of Fig. A2.3. Genomic Reconstruction of Transcriptional Regulons in *Shewanella* spp., p. 77). The reconstructed TRNs for the key pathways involved in central metabolism, production of energy and biomass, metal ion homeostasis, and stress response provide a framework for the interpretation of gene expression data. This analysis also helped to improve functional annotations and identify previously uncharacterized genes in metabolic pathways. Overall, this approach allowed us to identify candidate TFBSs for ~80 TFs from the *Shewanella* group. For ~30 described regulons, TFs were conserved between *Shewanella* and *E. coli*, whereas most others were characterized for the first time. Among many other observations, this analysis revealed a substantial rewiring in the TRN of the central metabolism between *Shewanella* and a classical model system of *E. coli* K12 (see Panel C of Fig. A2.3, p. 77).

Acknowledgments. Manuscripts describing Case Studies 1 and 2 are in preparation. Some of these results were presented at the DOE GTL 2008 meeting and briefly described in Fredrickson et al. (2008). These research results are from the work of the *Shewanella* Federation project (J. Fredrickson, principal investigator), supported by a DOE GTL grant. The examples in these sections were provided by A. Osterman and D. Rodionov of the Burnham Institute, La Jolla, California.

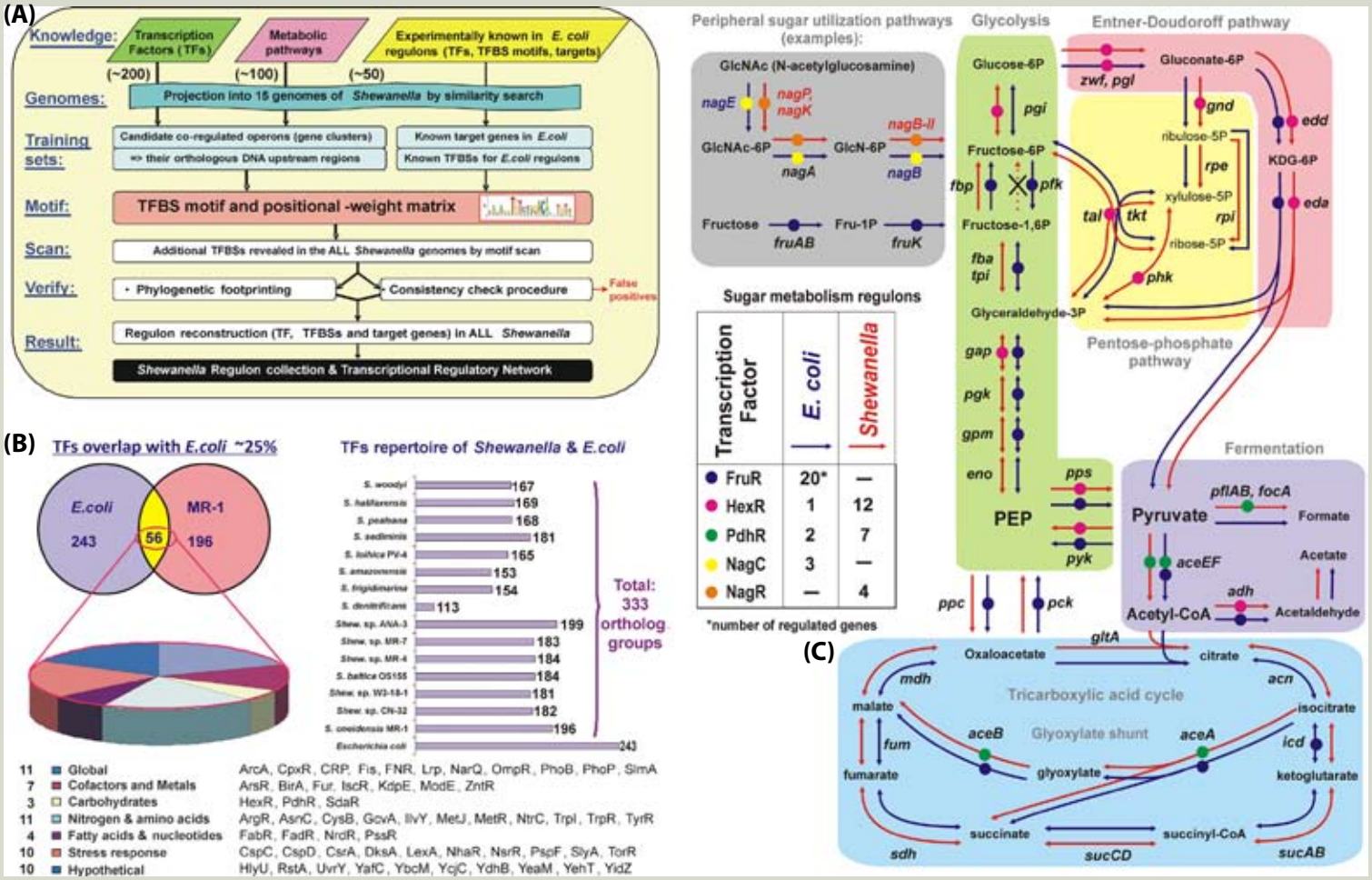

Fig. A2.3. Genomic Reconstruction of Transcriptional Regulons in *Shewanella* spp. (A) A knowledge-driven approach. **(B)** Survey of transcription factor repertoire in *Shewanella* spp. **(C)** Rewiring of regulatory network for central carbohydrate metabolism in comparison of *Shewanella* spp. and *E. coli* K12. For additional information, see Fredrickson et al. (2008). [Source: C portion of figure adapted by permission from Macmillan Publishers Ltd. From Fredrickson, J., et al. 2008. "Towards Environmental Systems Biology of *Shewanella*," *Nature Reviews Microbiology* **6**(8), 592–603 (http://www.nature.com/nrmicro/).]