Appendix 3
Systems Biology for Bioenergy Solutions

Defined here are the GTL Knowledgebase (GKB) requirements to aid and accelerate the understanding and engineering of biological systems for biomass conversion to bioenergy.

Background

The development of renewable alternatives to fossil carbon–based transportation fuels has become an urgent national priority. One of the most promising options for near-term, commercial-scale deployment is biofuel from lignocellulosic biomass (wood chips, grasses, cornstalks, and other inedible plant-based materials). Additionally, biological systems offer multiple paths to diverse bioenergy products—biodiesel from algal or plant biolipids, microbial methane production, or algal production of biohydrogen from sunlight and water.

The scientific breakthroughs needed to make lignocellulosic biofuel a cost-effective alternative to petroleum will require coordinated investigations of plant, microbial, and enzyme systems that span many orders of complexity and scale. Although some challenges are common to biological research across all DOE mission areas—such as noise, complexity, size, dynamic nature, lack of standardization, and heterogeneity of biological "omic" datasets—several issues are unique to bioenergy. While both carbon cycle and environmental remediation research are focused on understanding biological systems in their natural environments, bioenergy research seeks to understand and engineer these systems to work in highly controlled, production-oriented environments ranging from 96-well plates to 100,000-L fermentors or 1000-acre fields.

In addition to the core omic datasets resulting from the analysis of biological systems across multiple DOE missions, other key data unique to bioenergy include linked imaging and chemical characterization data for analyzing lignocellulose structures, imaging interactions within natural and constructed microbial communities, chemical analyses of chemical structure and breakdown intermediates of biomass, and a variety of datasets arising from sustainability research that examines the links among carbon, nutrient, and water cycles, as well as the environmental, economic, and societal impacts of bioenergy technologies being developed in the laboratory.

Biofuels: Grand Challenges for Biology

The ultimate goal for fundamental research in bioenergy, including the three DOE Bioenergy Research Centers, is to understand the biological mechanisms underlying biofuel production so well that those mechanisms can be redesigned, improved, and used to develop novel, efficient bioenergy strategies. Research undertaken by the centers and smaller endeavors will create the knowledge underlying three grand challenges at the frontiers of biology:

- Development of next-generation bioenergy crops.
- Discovery and design of enzymes and microbes with novel biomass-degrading capabilities.
- Discovery and design of microbes that will transform the production of fuel from biomass.

The science to solve these challenges is inherently complex, requiring multiple interdisciplinary teams that approach the same problems from different directions to accelerate scientific progress. Research details from DOE's Biomass to Biofuels Workshop can be found in the report, *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda* (U.S. DOE 2006, http://genomicsgtl.energy.gov/biofuels/). The following sections explain some scientific issues related to these challenges.

Development of Next-Generation Bioenergy Crops

Bioenergy crops are grasses, trees, and other plants grown specifically to provide cellulose and other complex cell-wall compounds, such as lignin, that strengthen and support plant structure as feedstocks for bioenergy. Plant cell walls are so complex that several thousand genes, few of which are known or understood, are thought to be involved in their synthesis and maintenance. By understanding the genes and mechanisms that control cell-wall synthesis in plants, scientists could develop new energy crops with altered biomass composition, modified links within and between cell-wall components, or increased accumulation of easily convertible starches and oils in plant tissues.

In addition to altering biomass composition, other energy-crop improvements include increasing biomass productivity per acre, increasing resistance to pests and drought, and decreasing the application of fertilizers and other inputs. Availability of more plant genome sequence information can accelerate the development of DNA markers used to identify and isolate the many genes associated with traits that can improve energy crops and speed identification of desired genetic variants and production of new energy crops. To obtain enough biomass for large-scale production of biofuels, a variety of perennial grasses and woody crops well suited to different geographies and regional climates across the nation will need to be grown.

Discovery and Design of Enzymes and Microbes with Novel Biomass-Degrading Capabilities

Nature uses both enzymes and multienzyme complexes called *cellulosomes* to break down cellulosic biomass. The performance of the limited number of biomass-degrading enzymes and cellulosomes studied thus far is much slower than many other types of enzymes as a result of a combination of the nearly impenetrable architecture of plant cell walls, chemical and physical changes to biomass during pretreatment, and structural features of the enzymes.

Each DOE Bioenergy Research Center is searching diverse biomass-degrading environments, from hot springs to rainforests to the guts of leaf-eating insects, to identify bacteria that specialize in producing enzymes that degrade biological materials.

Molecular-level understanding of how enzymes and cellulosomes degrade biomass is a prerequisite to designing improved enzymes for processing. Each center is integrating different combinations of methodologies to identify, characterize, and rationally optimize important factors influencing the rapid deconstruction of plant materials into sugars and other energy-rich components that can be converted to biofuels. These methods include high-throughput screens for proteins and metabolites, chemical analyses, state-of-the-art imaging technologies, and computational modeling.

Discovery and Design of Microbes That Will Transform the Production of Fuel from Biomass

In addition to cellulose, other carbohydrates (collectively called *hemicelluloses*) in plant cell walls are broken down into fermentable sugars when biomass is pretreated with heat and chemicals. Although cellulose is made of one type of six-carbon sugar (glucose) that is readily converted into ethanol and other products, microbial fermentation of the five- and six-carbon sugar mix from hemicelluloses is less efficient, thus representing a key area for improvement. En route to the fermentation tank, biomass currently is subjected to physical, chemical, and enzymatic processing steps that can create by-products and conditions that might inhibit microbial conversion of sugars into biofuels. Ethanol and other biofuel products also inhibit microbial fermentation at high concentrations. Consequently, developing microbes robust enough to withstand the stresses of industrial processing and tolerate higher ethanol concentrations is another important research area. Consolidated bioprocessing (CBP) is a more distant research target that could dramatically simplify the entire production process.

Consolidated Bioprocessing

The strategy of consolidated bioprocessing combines cellulose deconstruction and sugar fermentation into a single step mediated by a single "multitalented" microbe or stable mixed culture of microbes. CBP requires a redesign of microbial systems far more extensive than conventional genetic engineering approaches involving only the modifications of a few genes associated with microbial production of a single drug or other biochemical product. A successful CBP microbe or specially designed microbial consortium may be required to produce a variety of biomass-degrading enzymes; produce minimal numbers of molecules that inhibit the overall process; ferment both five- and six-carbon sugars; and thrive in industrial reactors with high temperature, low pH, and high concentrations of biofuel products.

Investigations of Bioenergy Systems Utilizing the Knowledgebase and Systems Biology Methods

Extremely complex phenotypes or functional characteristics important to bioenergy production—plant cell-wall biosynthesis, makeup, and structure; biomass degradation; and product tolerance and toxicity for biofuel-producing microbes—result from the protein products of numerous genes working together to control mechanisms at molecular, cellular, and higher levels. For example, within a plant genome are hundreds of genes participating in cell-wall biosynthesis, and the genomes of certain biomass-degrading microbes encode dozens of genes for hydrolyzing specific plant cell-wall polymers under different environmental conditions.Two general approaches are used to link variations in genes, pathways, and cellular mechanisms to particular phenotypes:

- Top-down: from known phenotype to understanding its cellular mechanisms and the bases for phenotype improvements.
- Bottom-up: from genomic characteristics of the sequenced bioethanol-producing microorganisms to specific phenotypes by quantification of molecular functions and their network clustering from omics technologies, as depicted in Fig. A3.1. Comparing Genomics to Phenotypic Characteristics, p. 82.

Fig. A3.1. Comparing Genomics to Phenotypic Characteristics.

Plant feedstocks involve additional layers of complexity at the tissue, organ, and whole-organism level that must be considered when studying phenotypes such as high productivity, partitioning to biomass, drought resistance, growth rate, diameter, morphology, field conditions, nutrient uptake rates, and yield. Analyzing observable plant phenotypes from specific natural variations or genetic modifications requires longer time frames for organism growth and development (months or years rather than hours for microbes) and largely manual interpretation by researchers.

Determining the mechanistic underpinnings of important bioenergy phenotypes will require systems biology approaches for the global analysis of completely sequenced organisms, analysis of microbial communities through metagenomics, and more focused investigations to characterize the structures and functions of plant biomass polymers and the enzyme systems that degrade those polymers.

Systems Biology: Enabling a Predictive Understanding for Sequenced Bioenergy Organisms

Applying systems biology approaches to bioenergy research challenges will require core knowledgebase components based on microbial and plant genomes. The completely sequenced genomes of organisms with capabilities relevant to DOE missions would form the foundation of the GTL Knowledgebase.

Microbial Genome Core Component

The microbial knowledgebase component would be a comprehensive repository of systems biology data for sequenced microbes with capabilities relevant to bioenergy production. Some of these microbes could serve as prototypes for platform microbes that are readily engineered to synthesize different fuel compounds. This component would require improved datasets of predicted gene calls and functions from genome sequences, an increased understanding of a minimal set of cellular functions needed to extract energy from substrates and generate desired fuels, and robust understanding of metabolic and regulatory networks that contribute to or detract from any given biofuel pathway. This activity would include quantitative omics (e.g., transcriptomics, proteomics, and metabolomics), transcription factors, flux analyses, and data to trace the flow of carbon, energy, and nutrients through critical pathways. Computational methods are needed that generate accurate predictive models—inferences for how to maximize transformation of renewable resources into fuels (e.g., improvement on conversion rate and conversion yield or the amount of fuel per equivalent substrate consumed).

Plant Genome Core Component

Genomic research can play a major role in developing new crops optimized for biofuel production without decades of agronomic research. Because of the incredible complexity of plant biology at all levels of organization from genome to cell to whole organism, systems biology for plants lags far behind current efforts for microbes. Although the number of sequenced plant genomes continues to grow, the volume of data associated with each plant is much larger than that of a microbial prokaryote.

Screening Plant Genomes for Bioenergy-Related TFs and Binding Sites

To develop plant feedstocks, a more comprehensive knowledge of transcription factors (TF) and the sites they bind throughout a genome is needed to narrow the list of potential genes associated with a specific phenotype such as increased differentiation of plant cells into xylem. Xylem cell walls in plant tissue are the primary source of cellulose used to make cellulosic biofuels, yet many genes linked to xylem differentiation are unknown. By screening the entire genome for new genes regulated by the same transcription factors that control known xylem differentiation genes, researchers can identify a short list of coregulated gene targets to be functionally characterized by experimentation.

Each cell within a plant contains the same genome, but the regulatory controls, subsets of expressed genes and proteins, and collections of metabolites can vary greatly for each cell type and even for subcellular compartments. In addition, multiple growth conditions, different tissues and organs, dynamicism in developmental states, and longer life cycles for plants result in more extended time frames for experimentation and practically endless combinations of variables to explore systematically. The research community is defining the experimental space most relevant to developing model bioenergy crops.

Dealing with biological complexity is a major obstacle, but another challenge is the limited availability of tools for analyzing plant systems data. By building on high-throughput technologies currently available for human and microbial systems, the first steps toward achieving predictive modeling of plant biology for bioenergy applications are within grasp. One initial goal would be to establish a bioenergy plant *phenome* database that identifies and defines the most important plant phenotypes relevant to bioenergy production and links these traits to genomic- and molecular-level functional information.

Improving Functional Annotations for Microbial and Plant Genomes

The availability of complete genome sequences for plants and microbes is the foundation for a broad range of approaches that can be used to characterize genes of unknown function and identify gene products with bioenergy-relevant functions. A genome encodes an organism's complete set of metabolic pathways, yet many key steps in these pathways may involve genes for which no functional information exists. Identifying subsets of genes that are coexpressed and regulated by the same elements (e.g., transcription factors; see sidebar, Screening Plant Genomes for Bioenergy-Related TFs and Binding Sites, this page) is one approach to discovering new genes involved in pathways that control such complex phenotypes as plant cell-wall biosynthesis.

Integrating Different Biological Datasets from Genome-Wide Analyses

The ability to integrate and compare orthogonal datasets would be central to scientific discovery from the GTL Knowledgebase. Relevant tools could either be built into the GKB or be external tools enabled by the GKB (e.g., MicrobesOnline,

http://www.microbesonline.org). The myriad unanticipated ways of combining data types point to the value of allowing for the latter approach—strong support for a tool-development community external to the knowledgebase team itself.

Of primary importance in systems biology research is improving the signal-to-noise ratio. Integration of multiple orthogonal datasets can accomplish this objective. As an example, one might remove spurious components of comparative gene neighbor–based regulon predictions by combining gene expression or protein-level data, protein-interaction data, and transcription factor binding-site motif detection. An additional goal of systems biology is to discover novel and unanticipated relationships. An example might be adding genes to a subsystem or pathway by combining a comparative phylogenetic footprint, expression profiles, and knockout assays of function. Numerous other ways could be used to integrate orthogonal datasets, and the value of such approaches will continue to grow as we devise new combinations.

The following are specific challenges in bioenergy research that might be addressed by integrating orthogonal data.

- Determining the necessary and sufficient subset of genes for elevated ethanol tolerance by combining expression analysis with comparative gene-content analysis of ethanol-resistant microbes.
- Determining the impact of stress conditions on metabolic pathways involved with biofuel synthesis by combining expression and protein-level data with data on metabolites and metabolic flux.
- Engineering novel metabolic pathways for biofuel synthesis from sugars or removal of pathways wasteful or detrimental to producing the desired end product, requiring measurements of the impact on modified pathways by combining expression and protein-level data on metabolites and metabolic flux.

Characterizing and Modeling Microbial Communities

In addition to characterizing omic data from completely sequenced organisms, the GTL Knowledgebase also would need to handle data from the metagenomic analyses of microbial communities. Metagenomics combined with environmental transcriptomics and functional assays for lignocellulose degradation would permit identification of the novel glycosyl hydrolases, transferases, and other important proteins involved.

GTL Knowledgebase Components Unique to Bioenergy Research

In order to identify the genetic basis for the recalcitrance of biomass to deconstruction by enzymes and microbes, omic data must be integrated, compared, and correlated with heterogeneous data. The data should come from a broad range of analyses to identify the most efficient biomass-degrading enzymes and characterize the composition and structural features of plant cell walls.

Linking Imaging Data to Other Experimental Data

Imaging biological systems over a wide range of spatial and temporal scales is an essential component of GTL bioenergy research because this capability provides a method for linking genomic and molecular information to complex biological functions (e.g.,

Integrating Existing Resource Data into GKB Component on Carbohydrate-Active Enzymes

To create a GTL Knowledgebase component focused on carbohydrate-active enzymes, a new integrated database for existing and newly discovered GHs and GTs will pull data from a variety of bioinformatics resources. Among those are the following:

- GenBank for genomic sequences (http://www.ncbi.nlm.nih.gov/Genbank/)
- MicrobesOnline (http://www.microbesonline.org) for comparative genomic and phylogenetic analysis
- Robetta structure prediction server (http://robetta.org) for structural model building
- RCSB Protein Data Bank (PDB, http://www.rcsb.org/pdb/) for structure data
- Gene Expression Omnibus (GEO) for microarray data (http://www.ncbi.nlm.nih.gov/geo)
- BRENDA (http://www.brenda-enzymes.info/), CAZy (http://www.cazy.org), and the UniProt Knowledgebase (http://www.uniprot.org) for guides that may be used more directly as repositories for enzyme functional parameters

The GDB carbohydrate-active enzyme component also will capture experimental conditions, protocols, and results to develop a cross-referenced database indexed to enzyme sequence and, where possible, high-resolution crystallographic structures.

Datasets from Sustainability Research

Sustainability research to analyze the potential economic, environmental, and social consequences of different pathways to large-scale biofuel production will play a critical role in determining the viability of new bioenergy technologies arising from systems biology research.

changes in microbial community behavior, enzyme-biomass interactions, and plant cell-wall structural information). Improving biomass conversion efficiency is key to biofuel production and requires the integration of multiple technologies to correlate changes in chemical properties with observed changes in cell-wall structure during the degradation process (see sidebar, Plant Cell-Wall Characterization and Visualization, beginning on p. 86, for a description of diverse data types generated by different techniques used to analyze biomass structure and chemical composition at multiple scales of space and time).

A major bioinformatics challenge for the GTL Knowledgebase would be developing efficient strategies for analyzing, storing, and sharing the vast amount of image data generated by the characterization of biomass and enzyme structures. Storage of all raw data within the GKB would be impractical, so another requirement would be tools for reducing noise and retaining only the most relevant, high quality data needed for subsequent analyses and visualization. To enhance interpretation and evaluation of image data, methods for associating experimental conditions with images and extracting and annotating the most biologically meaningful image features also would be needed.

Tools for Rapid Identification of Important Bioenergy Functions from Large-Scale Datasets

One of the most pressing needs in advancing production of lignocellulosic biofuels is the development of a comprehensive and robust knowledgebase of carbohydrate-active enzymes such as glycoside hydrolase (GH) and glycosyl transferase (GT) enzymes (see sidebar, Integrating Existing Resource Data into GKB Component on Carbohydrate-Active Enzymes, this page). This GKB component would combine the evolutionary history of GHs and GTs with structural, thermodynamic, and kinetic characterization to produce a more robust data and training set and comprehensive resource for the GH and GT research communities. This activity would include the development of

normalized functional assays for single and multiplexed GH and GT enzymes, as well as substrates for targeted reaction schemes and products.

Combining this work with comparative genomic and metagenomic studies would go a long way toward realizing the goal of establishing and predicting GH and GT libraries that would permit engineering of custom activities. Also important would be determining supporting genes within such systems as well as the combinations of enzyme subfamilies within a single organism or in combination. Unquestionably, accomplishing this goal will occur only by integrating the above data types to determine the rules governing which sequence begets a specific structure. Key data to be

Plant Cell-Wall Characterization and Visualization

Progress in understanding and manipulating the many steps involved in biofuel production will require a broad range of information regarding the chemical composition and molecular ultrastructure of the cell walls that make up the bulk of biomass (see figure, Switchgrass—Fluorescence Microscopy, facing page). Ready access to this molecular phenotype information will allow the genetic basis to be determined, providing a much deeper level of understanding than bulk analysis (e.g., mass and energy transfer and efficiency assessments).

Rapid development of plant cultivars whose biomass is more readily or efficiently deconstructed to simple sugars requires in-depth knowledge of cell-wall biosynthesis. Genetic analysis alone cannot provide this knowledge because the physical, chemical, and biological consequences of genetic manipulation must be identified. Quantitative data will be much more useful than qualitative data in this context. For example, quantitative analytical data describing the cell wall's physical and chemical features can be used to map these features to plant genetics (e.g., by quantitative trait loci analysis). Such data also are invaluable for forward and reverse genetic experiments. Furthermore, identifying molecular mechanisms that lead to increased efficiency of pretreatment chemistry, enzymatic deconstruction, and biological deconstruction requires quantitatively accurate analysis of chemistry and ultrastructure of cell walls in biomass before and after chemical or biological processing. Such analysis will, for example, provide information regarding the amount, molecular accessibility, and chemical and enzymatic susceptibility of various polymeric constituents of the biomass and how these parameters change as a function of genetic manipulation, development, and environment.

A wide range of analytical data will become available in the near future, including the following:

- **Wet chemistry analyses.** These investigations include carbohydrate content, monosaccharide composition, lignin content, and monolignol composition.
- **Spectroscopic analyses such as mass spectrometry (MS), infrared spectroscopy, and nuclear magnetic resonance (NMR) spectroscopy.** These techniques provide specific chemical and structural information, which often takes the form of "spectral features" that can be mapped to specific "structural features." Numerous examples of this type of data include pyrolysis molecular beam MS (PMBMS), which provides lignin and carbohydrate compositional data; solid-state NMR analysis of biomass for information on the crystallinity and allomorph composition of cellulose in the biomass; and high-resolution solution-state NMR analysis of solubilized biomass for data on the composition and detailed primary structural features of lignin and hemicellulose components of the biomass.
- **Surface chemistry analysis.** Novel spectroscopic and imaging methods are being developed to identify and quantitate specific molecules present at solvent-accessible surfaces. For example, MALDI-TOF microscopy provides information regarding the distribution of small molecules or molecular fragments at the sample surface.
- **Molecular mass analysis of polymeric and oligomeric components.** This analysis can be performed using various combinations of chromatographic and light-scattering technologies, as well as by MS methods.

considered would be which set of conditions, in the presence of supporting proteins and other factors, possesses the desired specificity and reaction kinetics for a given sugar biopolymer moiety. Proteomic and genomic tools are available and considered mature technologies. Bioinformatic tools that record structure-function relationships are making advances, but computational modeling tools to improve enzyme performance (e.g., sequence and structure based) are still extremely variable in terms of reliably identifying mutations that modify specificity and improve performance and stability. To provide a reliable tool for these computational efforts, researchers would need to integrate more biochemical experimental data with this sequence and structural information.

- **High-resolution imaging and diffraction methods.** These methods provide information regarding the molecular ultrastructure of polymeric components of biomass. This information includes the organization of cellulose microfibrils and spatial distribution of hemicelluloses and lignin in biomass.
- **Miscellaneous measurements.** These include physical properties such as porosity, density, compressibility, and heat capacity.

The huge amount of raw data recorded using these plant cell-wall characterization techniques can be interpreted by only a few experts. Therefore, the end user will require processed data to extract the most relevant parameters. For example, solid-state NMR spectra include significant information regarding cellulose crystallinity, but the end user probably is interested in a concise specification of the parameters that describe crystallinity. Because of the possibility of processing errors in such tabular data, robust provenance information must be included. For example, the biological source (e.g., genetic background of the plant cultivar) and physical and chemical processing history of the sample analyzed to produce the raw data should be readily available.

Ding et al. unpublished results.

Switchgrass—Fluorescence Microscopy. Fluorescence signals primarily come from chlorophyll, lignin, carotenes, and xanthophylls in plants, each with a different wavelength (color); lignin fluorescence is blue-greenish. Cell lignification is determined by using different filter sets. [Source: National Renewable Energy Laboratory and the DOE BioEnergy Science Center.]

Explicit storage of all raw data is impractical because of data-storage capacities and increase in the amount of data irrelevant to most knowledgebase users (i.e., noise). Only high-quality raw data should be included, for example, to be used for algorithm training.

A broad range of analytical and visualization tools will be required to effectively utilize this diverse cell-wall structural data. Evaluation of microimaging data will require visualization tools to render, zoom, and rotate images and to identify, demarcate, and quantitate relevant structural elements represented in these images.

In addition, a robust object model is required for the abstraction of structural features. That is, information about the same structural features (e.g., cellulose crystallinity) may be obtained using different analytical techniques. Nevertheless, the same formal representation of *cellulose crystallinity* should be used in structural feature tables obtained by different methods. A truly useful object model would be concise enough to ensure efficient querying, yet expressive enough to include all salient features required for meaningful analysis and comparison of data. Fundamental relationships such as partonomy (e.g., "a cellWall hasPart cellulose") and connectivity (e.g., "a sideChain isConnectedTo a polysaccharideBackbone") would provide deeper context for data retrieval and evaluation.