Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda

A Research Roadmap Resulting from the Biomass to Biofuels Workshop Sponsored by the U.S. Department of Energy

December 7–9, 2005, Rockville, Maryland

DOE/SC-0095, Publication Date: June 2006
Office of Science, Office of Biological and Environmental Research, Genomics:GTL Program
Office of Energy Efficiency and Renewable Energy, Office of the Biomass Program

DOE Genomics:GTL
GTL Biofuels
Home Page This Document

Chapter PDFs

John Houghton
Office of Science
Office of Biological and Environmental Research
301.903.8288
John.Houghton@science.doe.gov

Sharlene Weatherwax
Office of Science
Office of Biological and Environmental Research
301.903.6165
Sharlene.Weatherwax@science.doe.gov

John Ferrell
Office of Energy Efficiency and Renewable Energy
Office of the Biomass Program
202.586.6745
John.Ferrell@hq.doe.gov

Crosscutting 21st Century Science, Technology, and Infrastructure for a New Generation of Biofuel Research

Opportunities and Challenges

Efficiently and inexpensively producing ethanol or alternative products such as alkanes, fatty acids, and longer-chain alcohols from biomass will require significant advances in our understanding and capabilities in three major areas explored at this workshop: Feedstocks for Biofuels, Deconstructing Feedstocks to Sugars, and Sugar Fermentation to Ethanol. A systems-level approach to understanding and manipulating plants and microorganisms central to processing biomass into liquid fuels depends on obtaining and using detailed chemical and biochemical information on organism states and structures to build functional models that guide rational design and engineering. A systems-level understanding of model plants will facilitate rational improvement of plant cell-wall composition in crops dedicated to conversion into biofuels. New approaches and tools will be necessary to characterize definitively the detailed organizational structures of principal types of plant cellulose and their relative energies and interrelationships with such other structural components as lignins and noncellulosic polysaccharides.

Similarly, systems-level explorations are needed to determine genetic makeup and functional capabilities of such microbial communities as those involved in biomass decomposition and sugar fermentation. Emerging tools of systems biology—together with principles and approaches from metabolic engineering, synthetic biology, directed evolution, and evolutionary engineering—will help to overcome current obstacles to bioprocessing cellulosic feedstocks to ethanol. Increased emphasis must be placed on achieving a predictive understanding of plant and microbial biology including dynamics, regulation, flux, and function, with the ultimate goal of rational design to improve traits of bioprocessing microorganisms and plant feedstocks. For example, a major barrier in efficient use of biomass-derived sugars is the lack of microorganisms that can grow and function optimally in the challenging environment created through biomass pretreatment, hydrolysis, and cellular metabolism. A milestone in surmounting this barrier is to identify and understand molecular mechanisms used by cells to cope with such environmental challenges as high sugar and ethanol concentrations and the presence of inhibitors from biomass pretreatment and hydrolysis. Because resistance mechanisms typically involve complex subsystems and multiple genes, systems biology is needed for rationally engineering microbes to overcome these limitations. Ultimately, the systems biology focus of the Genomics:GTL program

References: p. 180

(GTL) has the potential to enable consolidation of the overall bioconversion process and reduce the number of biorefinery operations.

Powerful new tools and methods, including high-throughput analytical and imaging technologies and data-handling infrastructure, also will be needed to obtain, manage, and integrate information into models. Related tasks involve sequencing model crop-plant genomes, understanding the structure and function of biomass deconstruction enzymes, and modeling metabolic and regulatory networks of microbial systems involved in bioconversion. Some of these technologies, which either exist now or are envisioned in the GTL Roadmap, are planned components of the proposed GTL capability investments; others will need to be enhanced or developed for specific research outlined in this report.

Analytical Tools to Meet the Challenges of Biofuel Research

Systems for effectively converting biomass to liquid fuels will require a wide array of innovative analytical capabilities to facilitate fundamental and applied science. These capabilities encompass new methods for rapid and sensitive analysis of biomass polymers and subunits and high-throughput characterization of plant cell walls and microbial populations catalyzing bioconversion reactions. Analytical methods should provide detailed information on chemical moieties, chemical bonds, and conformation of plant cell-wall polymers; they should rapidly assess the state of microbes having new properties or growing under various defined environmental conditions. As discussed in preceding sections, a capability requirement is analysis of membrane components including lipids, proteins, and carbohydrates. These structures possibly represent major determinants for continued microbial activity in high-ethanol and other extreme environments. Membrane analytics are a challenging problem but an essential capability. New tools are needed to facilitate biocatalyst design and modification for many future processes, and technologies, methods, and computational and informational tools will be established to support core systems biology. Such tools include genomics, transcriptomics, proteomics, glycomics and lignomics, and fluxomics as described below; imaging technologies at various lengths and time scales; structure-characterization techniques; biomass characterization; cultivation; and accompanying modeling and simulation capabilities.

Genomics

Capitalizing on potential biofuel production from cellulosic biomass requires the continued commitment of DOE's Joint Genome Institute to sequence various organisms that contribute to the process. These organisms include crop plants, industrial yeast strains, and microbial communities involved in biomass decomposition and soil productivity. Genetic blueprints provided by DNA sequences will allow the use of systems biology for rational design and process consolidation to increase production of biomass crops and optimize biofuel conversion processes. Support will be

required for sequence assembly, optical mapping, and other techniques to place assembled contigs from a plant, microbe, or community on a physical map. Comparative genome analyses will infer similarities or differences in regard to other organisms, and expressed sequence tag and other libraries will facilitate genetic and molecular analysis of feedstock plants.

Genomic information will enable identification and molecular characterization of the diversity and activity of relevant biomass deconstruction enzymes from such new sources as fungi, bacteria, and uncultured microbes. Genomic analysis will open the door to use of genetic resources present in microbial communities specializing in lignocellulose degradation. Genomic projects are envisioned in the following general areas.

Feedstock Plants

This report outlines the importance of new and improved sources of feedstock biomass for conversion to liquid fuels. Considering project scale, soil and climate differences across the United States, and the desire to maintain biological diversity on agricultural lands, investigators recognize that a variety of biomass crops must be used. Grasses potentially are a prime source of biomass, so determining the DNA sequence of one or more grasses is a high priority. The genetic blueprint will assist in short-term development of grass-breeding strategies and the use of longer-term systems biology methods to realize the plant's potential as a source of biomass. Some plant systems to be studied include *Brachypodium* (slated for sequencing at DOE JGI) and *Populus* (already sequenced).

Soil Microbial Communities

Conversion into a high-value commodity is likely to alter the amount and composition of postharvest plant material and residues returned to the soil. Sustainability of biofuel technologies requires a scientific assessment of production practices and their influence on soil quality, including impacts on microbial communities and the processes they catalyze. Given that the technology involves crops tailored to thrive in individual regions, a number of sites representing various ecosystems must be chosen for soil metagenomic analysis. In the short term, such analyses will contribute to an understanding of effects on soil sustainability. The data also will be critical in preventing unwanted long-term effects on soil sustainability and for assessing more-direct effects of soil microbes on plant growth.

Fermentation and Biomass Decay Communities

Biomass conversion to sugars and biofuels requires optimizing microbial breakdown of structural sugars and fermentation of complex sugar mixtures. A number of microbial communities have evolved over millions of years to maximize and coordinate these capabilities, with several of the better-studied ones associated with ruminants and the hindgut of termites. Thus, a reasonable number of model fermentative communities that can degrade lignocellulose are critical targets for metagenomic analysis. In the short term, new insights into activities present or coordinated among members of these well-established communities will result. This information potentially can be used

to design second-generation biofuel systems composed of microbial communities or consortia considered more robust and diverse in regard to environmental conditions they can withstand and types of substrates they can use.

Increased Production Systems

Adaptive evolution of microorganisms using selective pressure in fermentors or chemostats has enormous potential to select for traits that benefit biofuel production. To understand mutations that give rise to desired traits, determining the molecular basis of such changes is critical; we need to sequence industrialized strains and compare them to their progenitors. In the long term, this information is vital in designing new microbial systems that increase bioconversion efficiency and lower biofuel cost (see sidebar, Proteomic and Genomic Studies of Industrial Yeast Strains and Their Ethanol-Process Traits, p. 126).

Transcriptomics: High-Throughput Expression Analyses

Measuring RNA expression (transcriptomics) provides insight into genes expressed under specific conditions and helps to define the full set of cell processes initiated for coordinated molecular response. This information can be used to elucidate gene regulatory networks and evaluate models of cellular metabolism. Current technologies include microarray-based approaches applicable to the study of plants and homogeneous microbial populations in well-mixed systems. Custom microarrays targeted to new plant varieties or microorganisms are critically needed. In heterogeneous systems such as those in mixed microbial populations, current microarray-based approaches are less useful because they provide only an average gene-expression profile across an entire population. Therefore, single-cell gene-expression methods applicable to diverse cell types are required.

Proteomics

Although carbohydrates are biomass conversion's primary substrate, proteins are workhorse catalysts responsible for constructing biomass-forming polymers, depolymerizing biomass before fermentation, and converting sugars to desired end products such as ethanol. To interrogate these various processes, identifying and quantifying proteins and protein complexes for various plant and microbial systems and subsystems are vital. The genome's information content is relatively static, but proteins produced and molecular machines assembled for specific purposes are dynamic, intricate, and adaptive. All proteins encoded in the genome are collectively termed the "proteome." The cell, however, does not generate all these proteins at once; rather, the particular set produced in response to a specific condition is precisely regulated, both spatially and temporally, to carry out a process or phase of cellular development.

Proteomics can be used to explore a microbe's protein-expression profile under various environmental conditions as the basis for identifying protein function and understanding the complex network of processes facilitated by multiprotein molecular machines. Identifying the suite of proteins

involved in construction and breakdown of cell-wall polymers is important in understanding biomass conversion to biofuels. In addition to analysis of intracellular proteins, all degradation enzymes and other proteins secreted by biomass-degrading microbes need to be characterized because they probably are key to the depolymerization process. Membrane protein systems are particularly important and need to be improved because they control cellulose production, cellulolytic enzyme excretion, and fermentable sugar transport.

To facilitate mass spectroscopy (MS) detection and quantitation of proteins in complex mixtures, improvements in instrumentation and chemistry are needed to enhance protein and peptide ionization, increase sensitivity and mass resolution, quantify protein levels in complex mixtures, and widen the dynamic range of their detection. New information and bioinformatic tools are required to analyze proteomic data.

Metabolomics

Analysis of the cell's metabolite content has lagged behind transcript and protein profiling but is an equally important indicator of cellular physiology. The metabolite profile may be a better indicator of cell physiology because metabolite concentrations (and fluxes) occur sooner in response to changes in the extracellular environment than do gene expression and protein production. Also, metabolites are precursors to the cell's transcripts and proteins and all other cell macromolecules, and regulation of cellular processes may not always be reflected in transcript or protein profiles.

The difficulty in profiling metabolites derives from their structural heterogeneity and short lifetimes inside the cell. Proteins and RNA are each composed of a constrained or limited set of precursors (amino acids in proteins, nucleic acids in RNA). Differences lie in amino acid or nucleic acid sequences, enabling a single separation method to analyze the entire transcript or protein profile at once. Unlike RNAs and proteins, metabolite heterogeneity makes analysis nearly impossible with a single separation technique. Although such techniques (e.g., high-performance liquid chromatography, thin layer chromatography, gas chromatography, and capillary electrophoresis) can separate metabolite groups having common structures, very few reports on separation and analysis of a comprehensive set of cellular metabolites have been issued. Furthermore, metabolites tend to have much shorter (seconds or less) half-lives than do proteins (hours) or RNAs (minutes). Shorter half-lives make rapid metabolism sampling and quenching even more important than analysis of RNAs and proteins. Methods such as nuclear magnetic resonance (NMR) that can effectively integrate and store information about flux through particular pathways and potentially can be applied to living systems. Localized metabolite concentrations such as those within an organelle also need to be measured.

Another complication is that a metabolite's intracellular concentration says little about its importance in cellular physiology. Indeed, some of the most potent cellular-signaling molecules found recently exist at relatively low concentrations inside the cell, and some metabolites produced and

consumed at the highest rates have relatively low intracellular concentrations. In the former case, extremely sensitive methods are necessary to measure these metabolites. In the latter case, the metabolic flux profile is more important than absolute metabolite concentrations (see section, Fluxomics, p. 161).

Metabolomic research will address fuel production, including methods to isolate, extract, and analyze labile metabolites such as those involved in cellular energy metabolism (e.g., ATP, GTP, NADP, and NADPH). Multidimensional separation methods are vital to analyzing plant or microbial-cell metabolites. New analytical techniques are required for continuous metabolite measurements, monitoring processes, and in vivo metabolite analyses.

Glycomics and Lignomics

As parallel concepts to transcriptomics, metabolomics, and proteomics described in the GTL Roadmap, biomass-to-biofuels research will use similar high-throughput and high-content analysis with compounds having low molecular weight and involvement in synthesis and degradation of plant cell-wall polymers. Glycomics (profiling materials related to structural polysaccharides) and lignomics (profiling materials related to lignin) are essential capabilities. New powerful analytical tools will provide information for systems-level understanding of biomass structure and chemistry and its role in biomass conversion to fuels and valuable chemicals.

Information from these analyses will help us understand native and modified pathways for synthesis of cell-wall polymers by tracking precursor consumption, generating and utilizing intermediate structures, and exploring their connection to plant cell-wall chemical composition and physical-chemical structure. A "toolbox" of options for plant breeders and crop scientists eventually will be available to improve feedstock substrates specifically for fuel production. New glycomic and lignomic tools also will be applicable directly to other plant-development needs.

Information obtained from lignomics and glycomics will elucidate substrate modification by tracking structural changes and concentration fluxes in saccharification products that may be linked to harsh pretreatments. Saccharification-product profiling also will guide selection of appropriate enzymatic cocktails by identifying specific chemical bonds and functional groups in solids and larger oligomeric products. The addition of enzymes tailored to known recalcitrant structures could significantly enhance product yields. Information on the nature of recalcitrant structures also can be used to guide plant breeding and pretreatment conditions. These approaches will provide insight into rate-limiting steps that arise in substrate- and enzyme-limited systems.

Accurate and robust glycomic and lignomic analytical tools will play an important role in fermentation research. Many lignin- and carbohydrate-derived compounds are fermentation inhibitors, and other materials may be converted to undesirable side-products. Tracking these materials from

feedstock through products will provide valuable insights into possible improvements in all stages of biomass-to-biofuels conversion.

Several specific challenges persist in the application of high-sensitivity, high-throughput tools for lignin and carbohydrate analysis. Many robotic and automated tools used in conventional metabolomics research cannot accommodate the larger sample sizes required for representative sampling of biomass substrates. Multivariate analysis tools will need to be developed and validated for accurate quantification of complex mixtures of substrates, pretreatment catalysts, conversion enzymes, and microbes. Many analytical methods for supporting a systems approach to biomass conversion will require biomass standards not readily available from commercial sources. Obtaining small-molecule standards will require prep-scale synthesis and isolation of molecules of interest and the ability to modify (e.g., by isotopically labeling) isolated materials using advanced techniques of carbohydrate and natural-product organic synthesis.

Fluxomics

Use of microorganisms for liquid-fuel production from lignocellulosic biomass depends on fluxes of diverse substrates through complex networks of metabolic pathways. Quantifying metabolic fluxes in microorganisms allows identification of rate-limiting steps in a biosynthetic pathway that could be improved by genetic manipulation or by alterations in cultivation conditions. Substrate turnover by an enzyme in a metabolic pathway can be determined by measuring changes in isotopically labeled substrate levels over a specific period of time. Using isotopomer analysis, quantities of labeled substrates and metabolic products derived from them can be measured and compared. Similar approaches can be used to map metabolic sugar fluxes to ethanol and related pathways.

Simultaneously measuring the turnover of all intracellular metabolites often is difficult, if not impossible. Once flux for certain enzymatic reactions is measured in the laboratory, flux through other cellular pathways can be calculated using mathematical models of the entire metabolic network. By modeling a given organism's metabolism, scientists can quantify the effects of genetic manipulations or changes in growth conditions on the cell's entire metabolic network. Inputs to flux-based models are the set of potentially active metabolic reactions and measurements of the steady-state production rates of such metabolites as DNA, RNA, carbohydrates, fatty acids, and proteins. Improved estimates of cell fluxes can be obtained by feeding a labeled carbon source and using measured transformed fluxes as input to the model.

A range of techniques and methods will be needed to determine fluxes in metabolic pathways of microbial cells employed for biomass fermentations. These techniques include stable and radioactive isotope labeling and associated methods for estimating unmeasured fluxes from isotope distribution in metabolites and macromolecules. MS- and NMR-based methods will determine isotope distributions in cellular metabolites and macromolecules.

The Super Imager

The potential is to create compound, multifunctional instruments that individually include many of the following capabilities:

- Mapping of molecular species such as RNA, proteins, machines, and metabolites through the use of fluorescent tags of various kinds
- Multiple excitation and detection wavelengths including both fluorescent and infrared absorption methods
- High-speed 3D imaging
- Nonlinear contrast imaging including second- and third-harmonic generation and coherent Raman scattering
- Lifetime mapping as sensitive probes of local environments
- Rotational correlation mapping for in situ analysis of protein structure and function
- Magnetic resonance imaging with 10-micron–scale analyses of metabolite concentrations and providing data on diffusion properties and local temperatures
- Acoustical imaging at micron-scale resolution of the system's physical parameters
- Atomic force microscopy (AFM) mapping of structures with added information provided by the controlled interaction of light and sharp metallic AFM tips to obtain optical resolutions of ~20 nm, one-tenth the diffraction limit
- High spatial resolution (nanometer scale) using X-ray and electron microscopies, including the use of special DOE facilities or perhaps the development of laboratory-based X-ray sources for imaging

[Sources: *Report on the Imaging Workshop for the Genomes to Life Program,* Office of Science, U.S. Department of Energy, 2002 (www.doegenomestolife.org/technology/imaging/workshop2002/); GTL Roadmap, pp. 182–87]

New analytical techniques are needed to quantify extracellular metabolites and quickly and easily determine biomass composition.

Enzyme Structure and Function

Developing and deploying improved enzymes and multienzyme complexes for biomass deconstruction and conversion will require more understanding and the production of suitable substrates as well as enzymes and their appropriate complexes. They will be used to achieve a mechanistic understanding of cellulose and cell-wall interactions with degrading enzymes.

Defining and Producing Substrates

Despite nearly 100 years of pretreatment research, detailed understanding of cellulosic biomass's fundamental physical and chemical features is still lacking. The plant cell wall—the substrate for degradative enzyme systems—is complex, containing various forms and quantities of cellulose, hemicellulose, and lignin. Chemical and physical pretreatment may alter biomass to make enzymatic feedstock digestion more difficult, so merging pretreatment and deconstruction steps would simplify processing. Critical to understanding any enzymatic reaction is having both a reliable assay and a defined substrate, thus allowing activities to be measured accurately. Constructing standardized substrates suitable for high-throughput assays is required to optimize these enzymes. Synthesis and characterization of model substrates require multiple technologies, first to define the substrates by employing multiple analytical and computational methods and then to synthesize the substrates. High-throughput methods are needed for compositional analysis and characterization of biomass substrates, once a family of basic reference structures has been identified.

Identifying Enzymes and Degradative Systems

Multiple classes of enzymes are required for biomass conversion to achieve maximum sugar yields including hemicellulases and ligninases. The rate-limiting step is making cellulose accessible for subsequent saccharification steps. Surveying and identifying suitable enzymes require many technologies listed in the GTL Roadmap, including metagenomic analyses of model biomass-degrading communities, high-throughput protein expression, generation of affinity tags, reassembly of complexes, activity measurements, and biochemical and biophysical characterizations of enzymes and complexes. Degradative systems fall into two general classes: Individual enzymes working synergistically; and the cellulosome, a large, mega-Da complex normally attached directly to cells and found to date only in anaerobic bacteria. The cellulosome is a LEGO-like system with many interchangeable structural and degradative protein components deployed when specified for particular substrates. This system provides the basis for engineering to create

cellulosomes optimized for particular substrates. The cellulosome can be detached from cells and functions, even under aerobic conditions. With appropriate modifications, the cellulosome has potential for enhanced stability and activity. Individual enzymes and those in cellulosomes might benefit from immobilization technologies designed to increase activity and stability. How membrane proteins such as sugar transporters interface with cellulosomes attached to cell walls is unknown (see sidebar, The Cellulosome, p. 102).

Imaging Technologies

"Thought is impossible without an image." —Aristotle, 325 B.C.

Imaging technologies with wide applications will be critical to many of the research challenges identified in this report. Biomass deconstruction is a crucial step in conversion, yet relatively little is known about the detailed molecular structure of plant cell walls and how they are constructed from various components. To address this limitation, new and improved methods are needed to analyze plant cell-wall composition and structure at the nanometer scale. Organization of polymer components in biomass structures should be analyzed in three dimensions using noninvasive tools. Such new capabilities are anticipated to reveal key molecular processes occurring in real time during the full life cycle of cell-wall formation and maturation and during experiments to transform crop species into feedstocks suitable for bioconversion by optimizing cell-wall makeup. Many requirements generally have been anticipated by GTL and by capability development and planning as documented on pp. 182–87 in the GTL Roadmap (U.S. DOE 2005) and Imaging Report (U.S. DOE 2002).

Imaging needs for each specific research area are summarized below. They include advances in a wide range of imaging technologies using NMR, optical, X-ray, and electron-based methods as well as atomic force and scanning tunneling microscopies. Given differences in resolution, sensitivity, and selectivity, the full impact of these methods will be realized best when used in combinations, either within a single instrument (see sidebar, The Super Imager, p. 162) or parallel applications assembled and correlated through advanced image-management software.

Imaging Needs for Feedstock Research

Within 5 years, methods will be developed and deployed for chemical-specific imaging over a wide range of spatial scales (0.5 nm to 50 μm), with contrast methods and tags enabling many molecular components to be distinguished easily. Imaging of living (or never-dried) materials is perhaps just as

Some Imaging Technologies Relevant to Feedstock Characterization

Atomic Force Microscopy (AFM)

In AFM, a scanning-probe technique allows the direct study of surface using tapping probes, the tips of which project less than a micron. The dynamic behavior of surfaces and molecules often can be observed. A great advantage of scanning force microscopy over most other high-resolution techniques is its ability to operate in a liquid environment. High-resolution AFM images of a cellulose surface recently have been reported. Using AFM to support interpretation of pretreatment and enzyme action on biomass surfaces represents a tremendous opportunity.

Scanning Electron Microscopy (SEM)

SEM really is the backbone of traditional biological surface analysis. New developments in the design of SEM sample chambers and "optics" permit the analysis of samples containing some natural moisture, which is critical for biomass fractions. Also, new strategies for creating replicas of biological samples provide more versatility in analyzing proteins and microbial-cell surfaces.

Transmission Electron Microscopy (TEM)

Cryoelectron microscopy is the leading technique for high-resolution molecular machine structural-biology research, and TEM is the most common platform for this method. The technique makes the following special demands on electron microscopy: Ultrahigh and clean vacuum for contamination-free observation, stable low-drift cryotemperature holders, special functions to provide low-dose imaging conditions, and cooled slow-scan charge coupled display cameras for low-dose digital image recording.

critical as imaging harvested materials that have been dried, either for storage in a processing facility or for structure examination. These imaging capabilities should enable observation of cell-wall deconstruction and construction. New optical technologies and advances at synchrotron light sources will allow some imaging methods to function as high-throughput devices, producing hundreds to thousands of images per minute with a single instrument. Important possible consequences are dramatic reduction of cost and processing time per sample and greater community access to these technologies as part of a user facility. Second, enhanced imaging speed enables real-time observation of biological processes. Availability of real-time data is vital to building and validating models that allow a systems biology approach.

Temporal imaging of dynamic subcellular small molecules, including metabolites, is among advances expected to take longer development times. Improved molecule-specific imaging tags and a means for introducing them into cells without disrupting function are critical needs. The scale of various biomass conversions will require enhanced imaging for obtaining high-resolution, large-volume tomographic images as well as real-time imaging of living systems at high spatial resolution and for use in the field at lower spatial resolution. The critical aspect of enzyme engagement with biomass components, such as cellulase interfacing with cellulose, requires new capabilities for characterizing enzyme binding sites by atomic force microscopy, scanning electron microscopy, transmission electron microscopy, and electron spectroscopy for chemical analysis [see sidebar, Some Imaging Technologies Relevant to Feedstock Characterization, p. 163, and Image Analysis of Bioenergy Plant Cell Surfaces at the OBP Biomass Surface Characterization Lab (BSCL), p. 40]. Similarly, capabilities are needed to characterize interactions between microbial cells and their solid-phase substrates.

Imaging Needs for Microbial Communities in Deconstruction and Conversion of Biomass to Ethanol

Meeting imaging requirements for processing biomass to ethanol will draw on the above capabilities but will focus more on microbial-cell and -community imaging to aid in delineating sub- and extracellular organization of proteins, protein complexes, transporters, and metabolites. Appropriate tags and methodologies will be required for investigating multicellular interactions in mixed microbial populations (e.g., those observed in fermentation) and communities (e.g., those observed in colonization of decaying biomass) with solid substrates such as plant cell walls. Understanding these interactions will require information on amounts, types, and locations of secretion products, reactions they catalyze, and how these reactions and products impact cell viability. Imaging-based approaches are essential to achieve temporal and spatial resolution sufficient for understanding these processes (see Fig. 1. Probing Microbial Communities, p. 165).

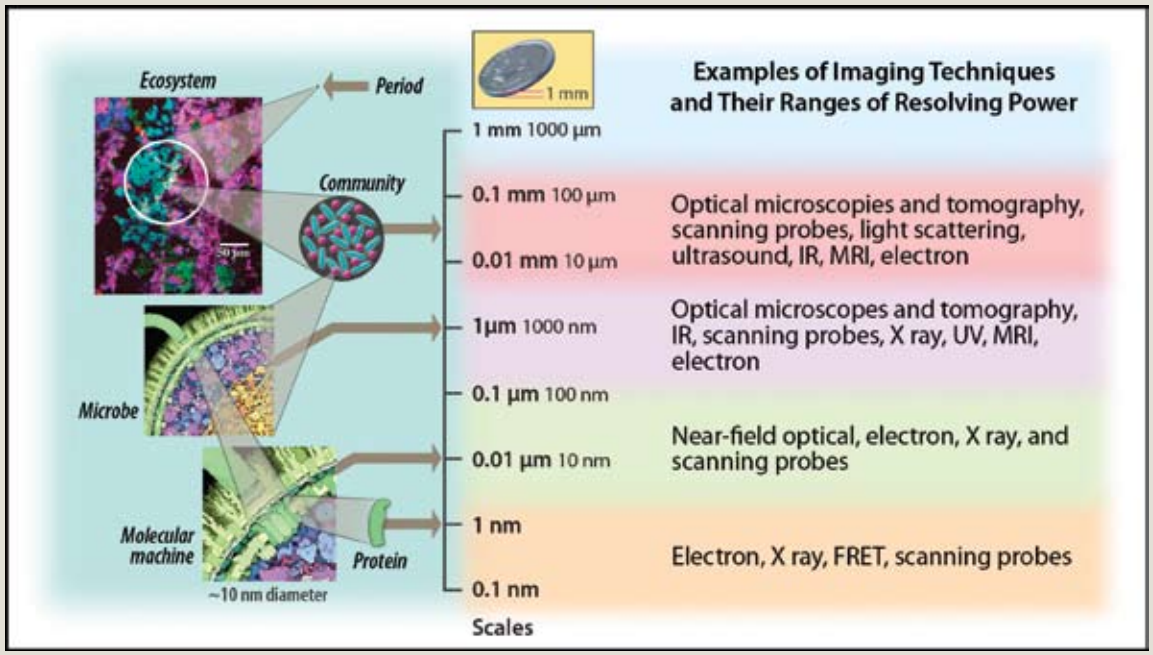

Fig. 1. Probing Microbial Communities. Microbial communities and ecosystems must be probed at the environmental, community, cellular, subcellular, and molecular levels. The environmental structure of a community will be examined to define members and their locations, community dynamics, and structure-function links. Cells will be explored to detect and track both extra- and intercellular states and to determine the dynamics of molecules involved in intercellular communications. Probing must be done at the subcellular level to detect, localize, and track individual molecules. Preferably, measurements will be made in living systems over extended time scales and at the highest resolution. A number of techniques are emerging to address these demanding requirements; a brief listing is on the right side of the figure. [Figure source: GTL Roadmap, p. 176 (http://doegenomestolife.org/roadmap/).]

Microbial Cultivation

The biorefinery environment is complex, and cultivation technologies must be capable of reproducing critical aspects of industrial systems. Efficient conversion of biomass to liquid fuels will require new and innovative approaches for controlling cultivation and simultaneously monitoring cell physiological states and metabolic processes under a range of conditions. Such capabilities are needed to identify gene regulatory and metabolic networks and to develop and evaluate modes of cell metabolism. Obtaining a systems-level understanding of fermentation organisms will require, in some cases, thousands of samples from single- and multiple-species cultures; and technologies for continuously monitoring and controlling culture conditions and interrogating the physiological state of microbial cells. Relatively homogeneous and complex microbial populations will be tested in physically and chemically heterogeneous environments such as those associated with solids (plant biomass). Necessary infrastructure will support cultivation at scales sufficient to obtain adequate amounts of sample for analysis and to grow microbial cells in monocultures and in nonstandard conditions (e.g., in association with solids or biofilms). These cultivation systems will be enhanced by advanced computational capabilities

Laboratory Cultivation Techniques to Simulate Natural Community Structure

To identify the function of genes preferentially expressed by specific populations in a structured microbial community, such as those deconstructing biomass in soils or in a bioreactor, new cultivation techniques are being devised. During the past decade, researchers have developed reactors in which biofilms can be imaged using confocal scanning laser microscopy (CSLM) and other light-microscopic techniques (Wolfaardt et al. 1994). When combined with fluorescent in situ hybridization to distinguish populations of cells in multipopulation biofilms and fluorescent reporters (green fluorescent protein) of functional gene expression, CSLM has been used to demonstrate how gene expression by one population affects gene expression in another proximally located population (Moller et al. 1998).

Pacific Northwest National Laboratory

The mobile pilot-plant fermentor shown here has a 90-L capacity and currently is used to generate large volumes of cells and cell products such as outer-membrane vesicles under highly controlled conditions. Future generations of fermentors will be more highly instrumented with sophisticated imaging and other analytical devices to analyze interactions among cells in microbial communities under an array of conditions.

that allow simulation of cultivation scenarios and identification of critical experimental parameters.

Most industrial bioconversions rely on pure cultures while, in contrast, environmental bioconversions are more commonly catalyzed by mixed populations or communities with "specialists" working together in an apparently integrated and stable fashion. A fundamental question is, Are there stable, self-regulating multiplex solutions for biofuels? In this regard, new techniques are needed to understand and reproduce composition and function in stabilized mixed cultures (see sidebar, Laboratory Cultivation Techniques to Simulate Natural Community Structure, this page).

Biological systems are inherently inhomogeneous; measurements of the organism's average molecular-expression profile for a cell population cannot be related with certainty to the expression profile of any particular cell. For example, molecules found in small amounts in samples with mixed microbial species may be expressed either at low levels in most cells or at higher levels in only a small fraction of cells. Consequently, as a refinement, new analytical and imaging techniques will be necessary to interrogate individual cell physiological states in heterogeneous culture systems. For experimentation in all aspects of work with omic tools, production of high-quality reproducible samples is paramount.

Data Infrastructure

Progress toward efficient and economic processes for converting biomass to biofuels will benefit from a research-data network able to manage, preserve, query, and efficiently disseminate large amounts of experimental and analytical data, as outlined in the GTL Roadmap. Because many research activities are data intensive and will engage a large number of investigators from national laboratories, academia, and the private sector, these activities will benefit from a distributed data environment.

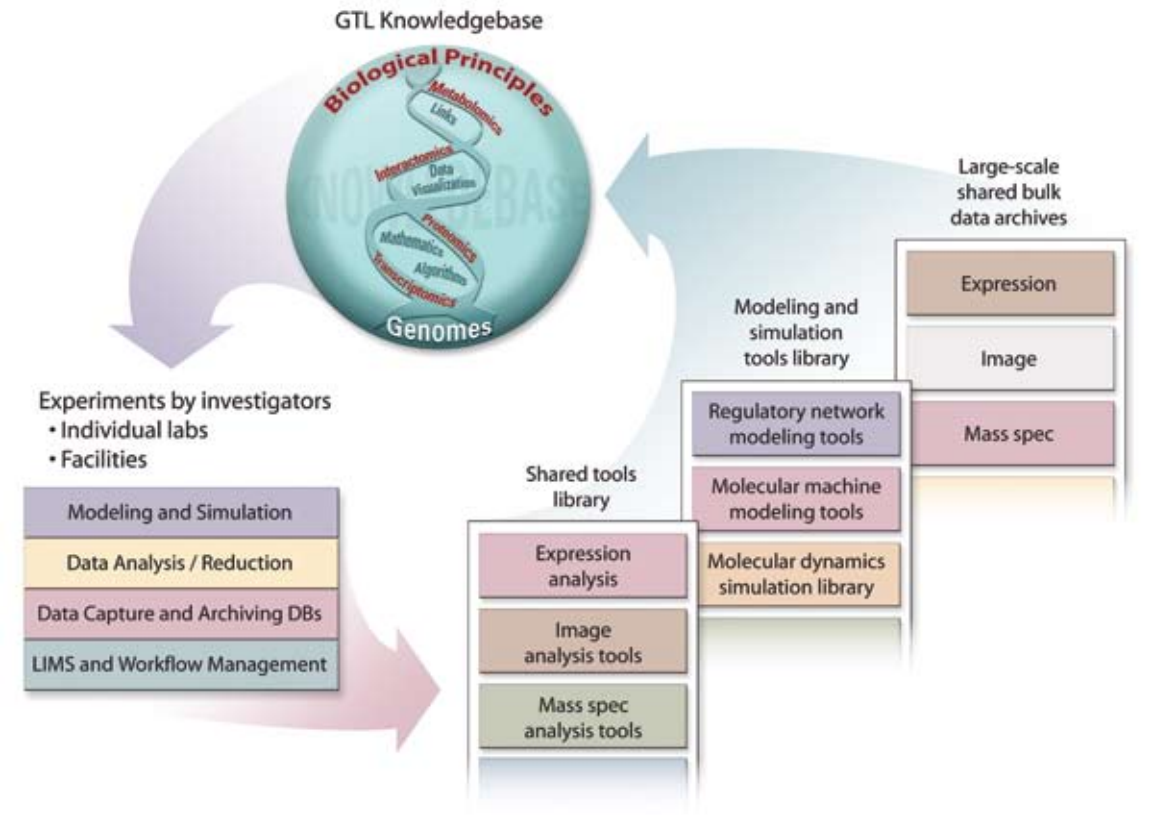

Fig. 2. GTL Integrated Computational Environment for Biology: Using and Experimentally Annotating GTL's Dynamic Knowledgebase. At the heart of this infrastructure is a dynamic, comprehensive knowledgebase with DNA sequence code as its foundation. Offering scientists access to an array of resources, the knowledgebase will assimilate a vast range of microbial and plant data and knowledge as they are produced. [Figure source: GTL Roadmap, p. 83 (http://doegenomestolife.org/roadmap/).]

Consisting of federated databases and repositories, the data environment will comprise descriptive, quantitative, and visual information, method libraries, query and data-mining tools, and a communication network. This environment provides controlled sharing and management and communication of biochemical, genetic, and other types of biological data and information. When available through a single portal, such a cyberinfrastucture can have a dramatic impact in efficient use of new knowledge. Additional benefits from efficient data sharing include fostering collaboration among projects, facilitating standards development, and accelerating implementation of new enabling technologies. GTL also will need a centralized research data network for model plants and specific biomass crop plants.

Data-infrastructure requirements of biomass-to-biofuels research align extensively with the Integrated Computational Environment for Biology described in the GTL Roadmap (see Fig. 2. GTL Integrated Computational Environment for Biology, this page), reproduced as a programmatic subset of the GTL information infrastructure. Biomass-to-biofuels research will require multiple databases, imaging archives, and LIMS in

each of its major research areas: Feedstock for Biofuels, Deconstructing Feedstocks to Sugars, and Sugar Fermentation to Ethanol.

For example, research on deconstructing feedstocks will necessitate (1) LIMS for tracking and documentation of samples for enzymatic pretreatment experiments and high-throughput compositional analysis of cell-wall material; enzyme databases for characterizing such key enzymes as glycohydrolases, esterases, and ligninases; metagenomic sequence databases for characterizing lignocellulolytic and other microbial communities; imaging repositories for a vast array of microscopies including optical, X ray, electron, and atomic force; and archives for molecular dynamic simulations of the molecular machinery that breaks down plant cell-wall components.

Feedstock and fermentation research goals have comparable data-infrastructure needs, with descriptive and quantitative data types differing mainly in experimental and analytical methods employed. Proteomic, metabolomic, transcriptomic, and plant and microbial genomic data will require corresponding databases and integration within the GTL Knowledgebase.

A variety of computational tools will be needed to support biomass-to-biofuels research and such enabling technologies as proteomics and imaging. For example, computational tools are necessary for streamlining the conversion of isotopomer flux-tracing data into flux distributions for subsequent analysis; the process currently is limited by the computational sophistication of required methods rather than by the ability to generate raw data. High-throughput, automated image acquisition, storage, processing, and analysis will be required for examining data on molecular machines in vivo. These tools also will aid in localizing and validating complexes, dynamics, docking, intercellular communication, extracellular matrix, and metabolite distribution.

Computational Modeling

Computational technologies will be critical to the overall platform that will enable success in developing effective technologies for converting biomass to biofuels. Significant amounts of data generated at a number of investigative levels must be made accessible to investigators at different institutions. Data sets will need to be analyzed and used to develop and evaluate computer models that will contribute to guiding overall decision making and program design. These decisions could involve the best lignocellulosic composition of biomass for a given process and how to create it as well as designing enzymes or engineering microbes for optimal ethanol production.

Computational modeling technologies and tools will be needed to address these challenges (see Fig. 3. From Genome Data to Full Cell Simulation, p. 169). Some key needs and examples are discussed below.

Modeling: Genome Sequence Analysis

A wide range of organisms across many different kingdoms of life—including plants, fungi, and bacteria—are relevant to biomass conversion to biofuels. With the power of modern genome sequencing and capabilities at

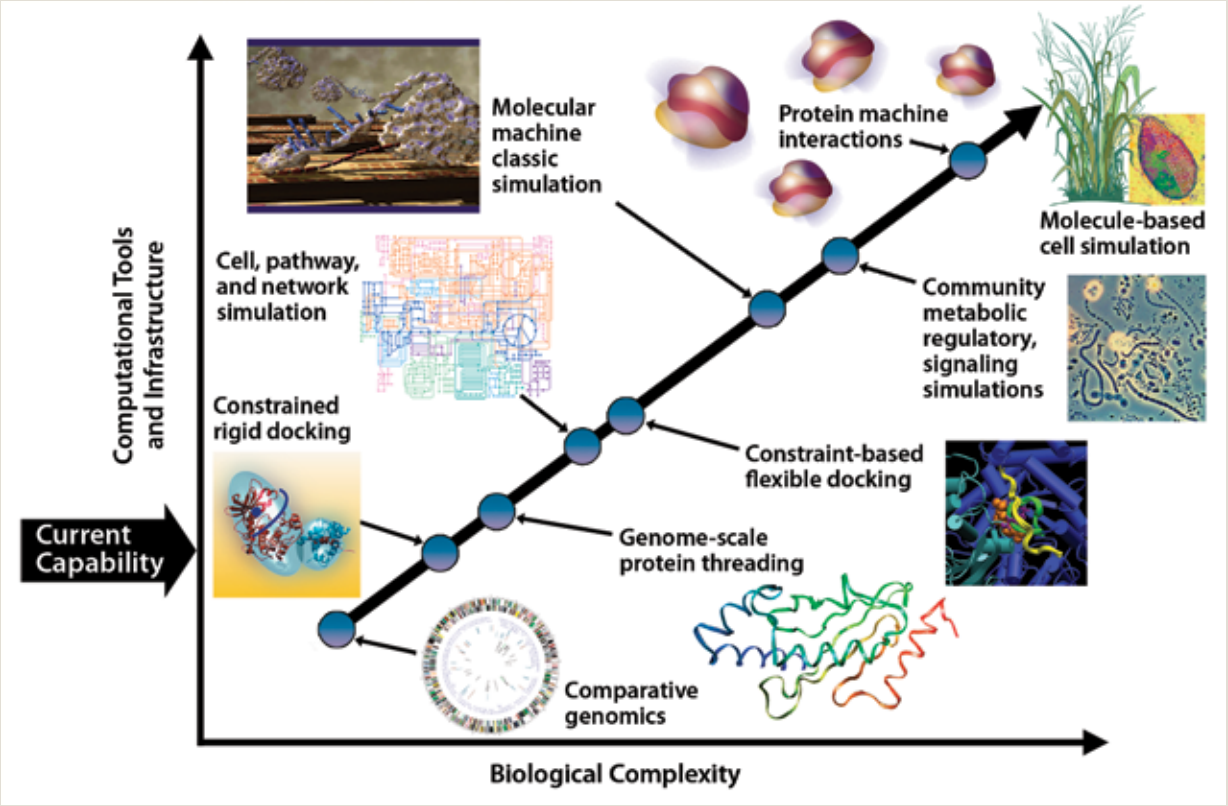

Fig. 3. From Genome Data to Full Cell Simulation. This concept diagram schematically illustrates a path from basic genome data to a more detailed understanding of complex molecular and cellular systems. New computational analysis, modeling, and simulation capabilities are needed to meet this goal. The points on the plot are very approximate, depending on the specifics of problem abstraction and computational representation. Research is under way to create mathematics, algorithms, and computer architectures for understanding each level of biological complexity. [Figure source: GTL Roadmap, p. 89 (www.doegenomestolife.org/roadmap/).]

DOE's Joint Genome Institute, obtaining genome sequences for essentially any organism is possible. Research will use sequence-analysis tools surpassing traditional genome annotation. The complexity of biomass conversion to biofuels will necessitate the evaluation and integration of sequence data from multiple organisms. The need to establish functional roles for a multitude of proteins involved in mediating plant cell-wall biology and defining metabolic and regulatory pathways will require robust bioinformatic capabilities. Within each genome sequence is a wealth of information that can be mined and accessed to improve our knowledge about these organisms. In some cases, traditional bioinformatic approaches may be sufficient. Many areas, however, still need research advances.

Sequence-based analysis methods and tools are vital in many areas. Creation of an improved-quality standardized pipeline for genome-sequence annotation, including definition of open reading frames and functional assignment of genes, is a critical need for essentially all sequenced organisms that are subjects of GTL mission-directed research. These annotations need to be updated continuously because advancements can be rapid in many areas and should be applied to single organisms, including plants, and metagenomes of soil ecosystems and fermentative consortia. Specific

to biomass-to-biofuels research are computational methods for improving analysis of metagenomic data to identify new enzymes related to lignocellulose degradation and fermentation. Directed evolution has the potential to rapidly evolve new and useful microbial strains for fermentation reactions. Comparative genomics of evolved microbes following resequencing will identify specific changes in the genome. Comparative analyses of multiple plant genomes will generate family clusters of gene components mediating cell-wall biosynthesis (i.e., enzymes, regulatory factors, and processing proteins). A more generic requirement is enhanced annotation related to recognition of interacting proteins, regulatory signals, and structures.

Modeling: Molecular

A number of critical biochemical processes, occurring at the molecular scale, govern the overall conversion of biomass to biofuels. These processes include, for example, mobilization of building blocks for plant cell-wall synthesis, machinery for cell-wall formation in plants and their biophysical characteristics, degradation of lignocellulosic materials by enzymes, and transport of sugars into microbes for further metabolism and conversion to desired end products. In combination with structural information, mechanistic models of plant-derived polymeric substrates and their enzymatic conversion are needed to identify "bottlenecks" in bioconversion of lignocellulosic biomass. Gaining a detailed understanding of these mechanisms will help in designing superior technologies to address key barriers. As a component of this research, new computer programs and codes are needed for ultralarge biological models of more than a million atoms to enable the analysis of complex molecular machines such as the cellulosome.

Molecular-scale models are critical to a number of biomass-to-biofuels aspects. For example, biophysical models of plant cell-wall composition are needed to delineate pathways for synthesis and transport of key cell-wall components. Our knowledge of fundamental polysaccharide-water interactions within lignin carbohydrate complexes will benefit greatly from molecular dynamics simulations, as will our understanding of water and chemical transport along or through the plant cell wall. Dynamic molecular models will aid in determining cell-wall enzymatic degradation and interactions among these enzymes and their substrates. Specifically, molecular models of such cell-wall–degrading enzymes as hemicellulases and ligninases will help to establish structure-function relationships for substrate interactions and guide the development of novel enzymes. Similarly, dynamic modeling of large molecular machines and their substrates on nano- to millisecond measures will provide insights into the structure and function of key protein complexes such as the cellulosome. Molecular modeling at a scale exceeding that for individual enzymes will enable understanding of critical events occurring at the interface between cellulose and an adhered cell. On the microbial side, models are needed to predict membrane composition and changes in response to stress induced by high concentrations of fermentation end products.

Modeling: Pathways and Networks

Molecular machines that carry out individual transformations and processes important in biomass-to-biofuels conversions often operate within much larger systems of interacting proteins and metabolic pathways in cells. Understanding, controlling, and manipulating the overall phenotypical state of these systems in either plant or microbial cells will be necessary to overcome identified technical barriers. Computational models of these cellular systems and pathways will facilitate integrative analysis of experimental omic data sets while also providing the basis for predictive simulations that can generate testable hypotheses. To increase the efficiency of fermentation reactions, improved metabolic and regulatory models are needed to understand mechanisms that control glycolytic flux and its impact on cellular metabolism. Because careful alterations of growth, energy, and redox conditions often are required to optimize fermentation reactions, robust models focused on analysis and regulation of cellular energetics are necessary. Regulatory and metabolic modeling at community and individual levels is essential for identifying and understanding gene regulatory networks.

A number of specific biomass-to-biofuels topics will require systems-level models and simulation methods. Genome-scale models of key industrial microbes are needed to understand metabolism details and allow rational design for improved biofuel production, including ways to identify novel biotransformation routes. Dynamic pathway models incorporating key enzyme kinetics for ethanol-producing pathways and other relevant cellular subsystems are a specific need. Methods also are required to assess physicochemical limitations of cellular systems and enzyme components to determine maximum-achievable metabolism rates (e.g., identify rate-limiting steps, both kinetic and diffusion limited). From the plant perspective, detailed pathway models of cell-wall biosynthesis in the context of overall plant metabolism are needed to generate testable hypotheses for controlling cell-wall composition. These models should incorporate information about catalytic activity, gene expression, mechanisms of control, and interspecies variations.

Generic modeling capabilities, many of which are required in the broader GTL program, include integrated modeling of metabolic pathways and regulatory networks and simulation of their functional capabilities. In support, automated techniques are needed to integrate data sets and rapidly reconstruct metabolic pathways and gene regulatory networks. Methods also will be required to incorporate the next level of complexity into systems-level cell models by adding information from cellular-component imaging. Other methods will account for the impact of protein and enzyme spatial localization within integrated models of cellular systems.

Modeling: Biorefinery Process

The long-term vision for biomass-to-biofuels research involves all cellular- and molecular-based procedures within an integrated biorefinery. To determine design requirements and alternatives, the overall process should be modeled computationally. Modeling is important, as is obtaining

experimental validation of large-scale designs that will enable implementation of genomic and fundamental science research. In addition, physical properties and material handling including feed, mass, and fluidic transport parameters need to be determined. Without these parameters, designing bioprocessing facilities—including bioreactors, heat exchange, filtration, bioseparations, centrifugation, pumps, valves, and other chemical-engineering unit operations—is difficult and not comprehensive. This capability encompasses economic models and those that assess the net carbon balance impact of different processing schemes and the advantages and disadvantages of various coproduct processes. Models do exist for some of these processes, but they are not readily available to the community. To allow researchers to assess the feasibility of different design concepts, results and methodologies should be readily accessible and subject to continuous feedback-based improvements.

Process modeling will be very useful in evaluating the feasibility of consolidated bioprocessing. Considering the goal to provide the science underpinning one or more commercial processes for producing liquid fuels from biomass, a proper evaluation of most-probable future scenarios is needed. For example, one scenario envisioned for more efficiency and economy is simultaneous saccharification and fermentation (SSF). Saccharification enzyme mixes currently are optimal at 50°C and at a pH between 4 and 6, so as indicated by other study groups, "*none of the current ethanol-producing microorganisms is suitable.*" Process modeling will evaluate the importance and sensitivity of SSF vs a two-step process. Modeling is needed to demonstrate rational design of integrated biomass processing by predicting and then verifying overall hydrolysis yields for native and modified biomass species using different pretreatment chemistries, temperatures, and specific enzymes.

In summary, research is needed to develop process models for assessing mass balances and economic models for bioprocess engineering and to disseminate them for community benefit (see chapter, Bioprocess Systems Engineering and Economic Analysis, p. 181).

Capability Suites for Bioenergy Research and Facility Infrastructure

To address mission science needs in energy and environment, the DOE Office of Science has proposed establishing vertically integrated research centers that will draw upon the range of advanced, high-throughput technologies and information-management computing described in the GTL Roadmap. The first of these centers will focus on bioenergy. This section provides an overview of existing and developing GTL capabilities that will be required to accelerate biomass-to-biofuels research. The GTL program has established pilots of most key technologies required to perform systems biology that might be incorporated into research centers. Biological capabilities need to be investigated at many scales (see Fig. 9. Understanding Biological Capabilities at All Scales Needed to Support Systems

Biology Investigations of Cellulosic Biomass, p. 21). Over the past decade, genomic sequencing has been established as a highly efficient production component at the DOE Joint Genome Institute. Such capabilities provide the genomic foundation for research described in this report.

DOE Joint Genome Institute

DOE JGI's mission is to provide integrated high-throughput sequencing and computational analyses to enable genomic-scale, systems-based scientific approaches to DOE challenges in energy and environment. JGI capabilities will have an immediate impact on biomass-to-biofuels science by providing high-quality genome sequences of relevant crop plants and microorganisms, including naturally occurring and engineered microorganisms used for biomass conversion and soil microbial communities as they relate to soil quality and sustainability.

Identification of genes that control cell-wall composition in biomass species and development of tools such as gene chips depend on the availability of genome sequences for biomass crop species. Since micro-RNAs are expected to play a role in controlling expression in many relevant genes, DNA sequencing must be comprehensive enough to identify all micro-RNAs in these species.

Implementating effective biomass conversion will require JGI to rapidly and cost-effectively determine the sequences of a number of organisms, consortia, and metagenomes. In addition to biomass crop plants, JGI could provide genome sequence for new species of white rot fungi and brown rot fungi, actinomycetes, and such other biofuel-relevant organisms as natural consortia and communities involved in biomass decomposition. Full metagenomic sequencing will elucidate hemicellulose and lignin breakdown observed in mixed microbial populations and will allow harnessing of their collective processes.

Production and Characterization of Proteins and Molecular Tags

Functional Capabilities

These capabilities would enable the high-throughput expression of proteins mediating cell-wall biosynthesis (laccases, peroxidases, and glycosyl transferases) in *Arabidopsis* and other model and feedstock plants. Most enzymes of interest in cell-wall biosynthesis, such as glycosyl transferases, are thought to be membrane associated and, therefore, difficult to purify and characterize by conventional methods. Heterologously expressed proteins would be characterized biochemically and used as a resource for first-pass identification of function and the generation of antibodies and tags. To improve enzymes for lignocellulose deconstruction, the protein production capability would be a resource for native and recombinant forms of enzymes and modified enzymes from directed evolution and from rational-design, site-directed mutagenesis approaches. For pilot and validation studies, significant quantities of protein would be produced. Proteins, natural and modified, could be used to rapidly reconstitute

machines such as cellulases and cellulosomes for improved performance. Antibody resources would have multiple applications, including pinpointing cellular localization of target proteins; protein complex identification; quantification of natural functional diversity; and quantitation of molecular outcomes from breeding experiments. In addition, this resource would generate tags and probes for imaging cell-wall polymers and progress of enzymes that degrade these polymers.

Assays for many cell-wall synthesis and deconstruction enzymes will require new approaches to screen for activity. In particular, they will involve access to different polysaccharide substrates and capabilities. In combination with proteomic capabilities, the resource would provide valuable information about proteins and mechanisms through which biomass is converted to biofuels. This knowledge will enable engineered-organism combinations to optimize biofuel production.

These capabilities will be critical for generating and characterizing new proteins and biomarkers for imaging. For example, a collection of fluorescently labeled proteins for such specific model plants as *Arabidopsis* or *Brachypodium* would be particularly valuable.

Instrumentation and Methods

From genome sequences, protein production and characterization methods will express proteins and generate reagents for interrogating cell function. Specifically, the goal will be to create capabilities to produce on demand all proteins encoded in any genome; create molecular tags that allow each protein to be identified, located, and manipulated in living cells; gain insights into function; and perform biophysical and biochemical protein characterizations. Using high-throughput in vitro and in vivo techniques (i.e., cellular and cell free) will lower protein-production costs to levels that will allow comprehensive analysis of cellular proteins. These methodologies will be applied to the production and characterization of modified and native proteins. Products and analysis capabilities will be made available to scientists and technologists.

"Affinity reagents" will be generated in parallel with protein production, using many of the same technologies. Tagged proteins or nucleic acids will permit detection and tracking of individual proteins in living systems, including complex molecular assemblages; intracellular position of all proteins and their spatial dynamics; if secreted, extracellular localization and interaction with other community members; and techniques for manipulating protein activity in the environment.

Potential core instrumentation:

- Gene-synthesis and manipulation techniques.
- High-throughput microtechnologies for protein-production screening.
- Robotic systems for protein and affinity-reagent production and characterization.

- Computing for data capture and management, genomic comparative analyses, control of high-throughput systems and robotics, and production-strategy determination.

Characterization and Imaging of Molecular Machines

Functional Capabilities

These capabilities will enable identification and characterization of protein complexes responsible for essential functions of energy-relevant phenotypes, including cell-wall biosynthesis. Understanding the function of these cell-wall synthesizing complexes and their interactions with metabolic pathways that produce sugar nucleotides in the cytosol will be important for understanding polysaccharide biosynthesis. Identifying key protein-protein interactions (interactomics) involving such biomass-deconstruction enzymes as polysaccharidases, hemicellulases, ligninases, and esterases, along with interactions with their substrates, also is a priority. Because polysaccharides probably are synthesized by a multienzyme complex in the Golgi, that also will be identified and analyzed.

Advanced capabilities are needed to understand the structure and function of cellulase and cellulosome molecular machines and improve designer cellulosomes by linking essential enzymes to the desired substrate and eliminating superfluous proteins and reactions. These advanced capabilities include analysis and computation for probing interactions among cells, their associated secreted enzyme complexes, and cellulose; and particularly for understanding the biology occurring at the cell surface–cellulose interface.

Instrumentation and Methods

Capabilities will identify and characterize molecular assemblies and interaction networks (see Fig. 4. Visualizing Interaction Networks, this page). Resources will isolate and analyze molecular machines from microbial cells and plants; image and localize molecular machines in cells; and generate dynamic models and simulations of the structure, function, assembly, and disassembly of these complexes. High-throughput imaging and characterization technologies will identify molecular machine components, characterize their interactions, validate their occurrence, determine their locations within the cell, and allow researchers to analyze thousands

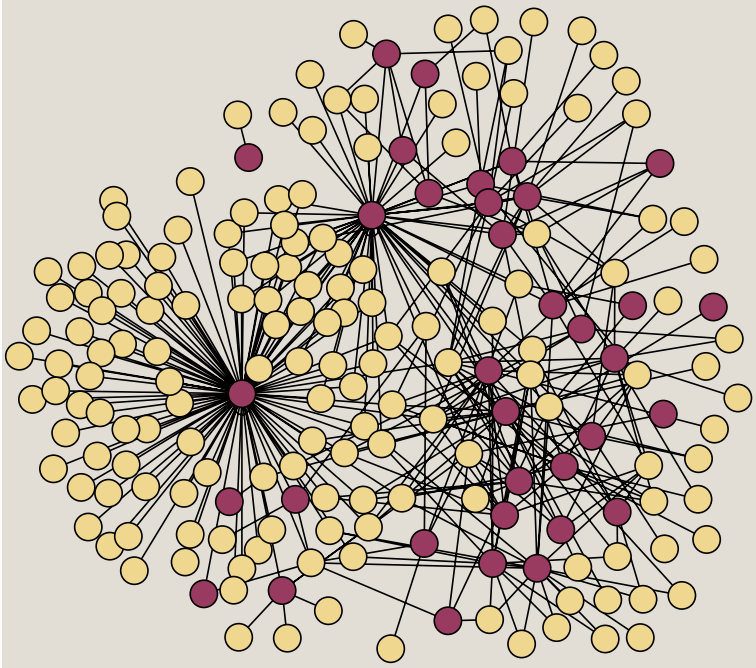

Fig. 4. Visualizing Interaction Networks. Graphical maps display protein interaction data in an accessible form. These visualizations summarize data from multiple experiments and also allow quick determinations of proteins that might be core constituents of a particular protein complex and those that might play roles in bridging interactions among different complexes. The figure above, generated using Cytoscape, summarizes protein interactions in complexes isolated by affinity approaches from *Shewanella oneidensis*. Nodes (yellow or red circles) represent proteins. [Source: GTL Center for Molecular and Cellular Systems at Oak Ridge National Laboratory and Pacific Northwest National Laboratory; previously published in GTL Roadmap, p. 68 (doegenomestolife.org/roadmap/).]

of molecular machines that perform essential functions inside a cell. The capability for completely understanding individual molecular machines will be key in determining how cellular molecular processes work on a whole-systems basis, how each machine is assembled in 3D, and how it is positioned in the cell with respect to other cellular components.

Potential core instrumentation:

- Robotic culturing technologies to induce target molecular machines in microbial systems and support robotic techniques for molecular-complex isolation.
- State-of-the-art mass spectroscopy and other techniques specially configured for identification and characterization of protein complexes.
- Various advanced microscopies for intracomplex imaging and structure determination.
- Imaging techniques for both spatial and temporal intra- and intercellular localization of molecular complexes.
- Computing and information systems for modeling and simulation of molecular interactions that lead to complex structure and function.

Analysis of Genome Expression: The Omics

Functional Capabilities

An organism selectively produces portions of its proteome in response to specific environmental or intracellular cues. Studying its constantly changing protein expression thus leads to better understanding of how and why an organism turns genome portions "on" and "off." Identifying, quantifying, and measuring changes in global collections of proteins, RNA, metabolites, and other biologically significant molecules mediated by proteins—including lipids, carbohydrates, and enzyme cofactors—are important to this understanding. The ability to measure time dependence of RNA, protein, and metabolite concentrations will reveal the causal link between genome sequence and cellular function (see Fig. 5. Gene-Protein-Metabolite Time Relationships, this page). High-throughput omic tools (transcriptomics, proteomics, interactomics, metabolomics, glycomics, lignomics, and fluxomics) will be critical for systems-level investigation of plant cell-wall

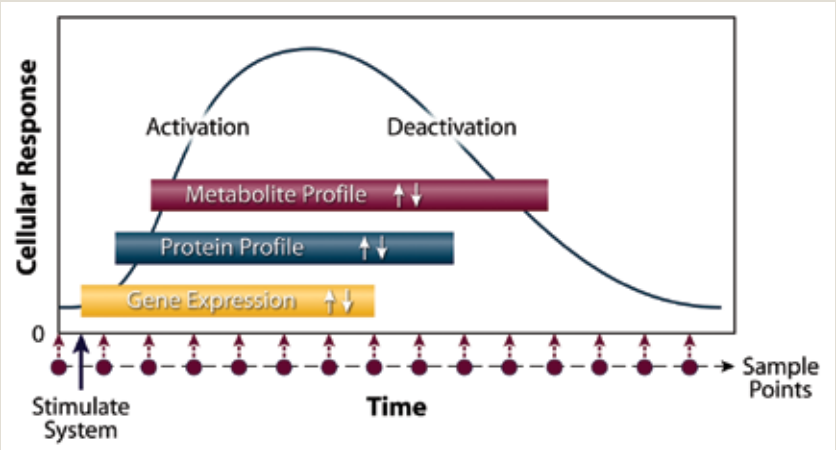

Fig. 5. Gene-Protein-Metabolite Time Relationships. To accurately establish causality among measured gene, protein, and metabolite events, sampling strategies must cover the full characteristic time scales of all three variables. Little is known about the time scale of gene, protein, and metabolite responses to specific biological stimuli or how response durations vary in genes and species. [Figure adapted from J. Nicholson et al., "Metabonomics: A Platform for Studying Drug Toxicity and Gene Function," *Nat. Rev.* 1, 153-61 (2002); previously published in GTL Roadmap, p. 158 (doegenomestolife.org/roadmap).]

construction and processes whereby the cell wall is deconstructed and converted to ethanol. Multiple analytical capabilities will be seamlessly integrated with modeling to achieve the needed level of understanding.

Use of proteomics will allow comprehensive analysis of plant protein complexes, factors controlling their creation and function, and ensuing processes, beginning with the model plant *Arabidopsis.* The thousand or more proteins involved in cell-wall synthesis or modification are highly interacting or located in complexes. Knowing which proteins are in complexes before attempting to develop surrogate expression systems for enzyme characterization will be essential. Many such enzymes are membrane associated, so innovative methods for characterizing membrane complexes must be developed. These capabilities are vital to documenting the molecular makeup of living cell types in vascular and other tissues, including ray parenchyma and phloem.

MS and NMR would be used to develop very sensitive, high-throughput assays for glycosyl transferase activity. Synthesis, purification, and characterization of glycans needed as acceptors would be important.

A "glycochip" or related method for assaying enzymes involved in polysaccharide synthesis and modification also would be needed. High-throughput MS capabilities would be well suited for developing the analytical aspects of a glycochip. These resources also would benefit analysis of microbial systems used for plant cell-wall deconstruction and sugar fermentation. For example, such capabilities could quantify response of white rot fungi to changing culture conditions and provide a systems-level understanding for improving the energetics and carbon-allocation efficiency for cellulase and cellulosome production.

Optimizing microbial cultures for bioconversion will require tools that allow quantitative cellular characterization at the systems level, including those for global transcript, protein, and metabolite profiling in conjunction with metabolic and regulatory modeling. HTP methods to identify binding sites of global regulatory proteins will be required for models of global gene regulation. Additional HTP tools will monitor key metabolites that define cell redox and energy states [e.g., ATP, GTP, NAD(P)H, and NAD(P)] for fermentation optimization.

Instrumentation and Methods

Consolidated omics will be capable of gaining insight into microbial functions by examining samples to identify (1) all proteins and other molecules created by a plant or individual or community of microbes under controlled conditions and (2) key pathways and other processes. Integrating these diverse global data sets, computational models will predict microbial functions and responses and infer the nature and makeup of metabolic and regulatory processes and structures.

Potential core instrumentation:

- Large farms of chemostats to prepare samples from highly monitored and controlled microbial systems under a wide variety of conditions.

- Numerous specialized mass and NMR spectrometers and other instruments capable of analyzing the molecular makeup of ensemble samples from thousands of diverse molecular species.
- High-performance computing and information capabilities for modeling and simulating plant or microbial-system functions under different scenarios. These studies will inform the design of systems-level experiments and help to infer molecular processes from ensuing data.

Analysis and Modeling of Cellular Systems

Functional Capabilities

Biomass complexity, coupled with intricate processes involved in conversion to ethanol, requires experimental capabilities and models for analyzing the biomass-enzyme-microbe system, specifically cell-wall synthesis and microbial decomposition of biomass. These capabilities also can contribute to determining the response of complex soil microbial communities to various cropping regimes.

On the basis of similarity with genes of known function, advanced computation will be used for comprehensive analysis of genes from model and biomass crop plants implicated in polysaccharide synthesis or modification. Acquisition and analysis of HTP gene-expression data from model and crop plants would be powerful in assigning probable function to genes implicated in cell-wall synthesis. As enzyme function emerges, a systems-level plant-cell model would incorporate biophysical and structural properties and knowledge about pertinent proteins. Models of this type would greatly facilitate the rational development of feedstock species based on "design principles," in which wall chemical composition is optimized as feedstock while plant productivity is simultaneously maximized.

These resources should include capabilities for generating models that enable improved cellulase production from near-term hosts such as *Trichoderma reesei* and proceed to bacterial and yeast systems that will require longer times for research and development. Other goals are to improve production of specific preferred components and foster a more global understanding of hyperproduction at industrial scales in industrial strains.

An integrated model of fermentative microorganisms and the ability to test it experimentally will support new concepts to produce liquid fuels and understand the producers' underlying metabolism and responses to stress in producing biofuels such as ethanol. Models and new rounds of modification and experimentation then will be used to alleviate stress responses and increase carbohydrate degradation, minimize cell growth and side products of synthesis, and maximize fuel production.

Metabolic flux analysis (as described earlier) is essential in improving understanding and manipulation of cellular metabolism. Possible analytical needs include appropriate NMR and MS instrumentation; stable isotopic labeling for key molecular moieties involved in metabolic processes; and

synthesis of process intermediates. Intensive mathematical and computational power is required to achieve the final goal of flux estimation.

Instrumentation and Methods

Technologies and methodologies for analyzing cellular systems will be the capstone for ultimate analytical capabilities and knowledge synthesis to enable a predictive understanding of cell, organism, tissue, and community function critical for systems biology. Imaging methods will monitor proteins, machines, and other molecules spatially and temporally as they perform their critical functions in living cells and communities. Within their structures, microbial communities contain numerous microniches that elicit unique phenotypical and physiological responses from individual microbial species. The ability to analyze these niches and the microbial inhabitants within is crucial to our ultimate goal. This grand biological challenge must be addressed before scientists can predict the behavior of microbes and take advantage of their functional capabilities. Modeling resources will enable description of biological interactions with the physicochemical environment and predict how the system will evolve structurally and functionally.

Potential core instrumentation:

- Highly instrumented cultivation technologies to prepare structured microbial communities for simulating natural conditions in a highly controlled environment.
- Instruments integrating numerous analytical-imaging techniques to spatially and temporally determine, in a nondestructive way, relevant molecular makeup and dynamics of the community environment, community, and microbes that comprise it.
- Computing and information capabilities to model and simulate complex microbial systems, design experiments, and incorporate data.

Cited References

Moller, G. M., et al. 1998. "In Situ Gene Expression in Mixed-Culture Biofilms: Evidence of Metabolic Interactions Between Community Members," *Appl. Environ. Microbiol.* **64**, 721–32.

U.S. DOE. 2005. *Genomics:GTL Roadmap: Systems Biology for Energy and Environment*, U.S. Department of Energy Office of Science (doegenomestolife.org/roadmap/).

U.S. DOE. 2002. *Report on the Imaging Workshop for the Genomes to Life Program*, April 16–18, 2002, Office of Science, U.S. DOE (doegenomestolife.org/technology/imaging/workshop2002).

Wolfaardt, G. M., et al. 1994. "Multicellular Organization in a Degradative Biofilm Community," *Appl. Environ. Microbiol.* **60**, 434–46.