5.3. Facility for Whole Proteome Analysis

5.3.1. Scientific and Technological Rationale 156

5.3.2. Facility Description 158

5.3.2.1. Production Targets 158

5.3.3. Technology Development for Controlled Microbial Cultivation and Sample Processing 159

5.3.3.1. Development Needs for Cultivation 161

5.3.3.2. Development Needs for Sample Processing 161

5.3.4. Large-Scale Analytical Molecular Profiling: Crosscutting Development Needs 162

5.3.5. Technology Development for Transcriptome Analysis 162

5.3.5.1. Global mRNA Analysis 162

5.3.5.1.1. Microarray Limitations Requiring R&D 163

5.3.5.2. Small Noncoding RNA Analysis 164

5.3.5.2.1. sRNA-Analysis Development Needs 164

5.3.6. Technology Development for Proteomics 164

5.3.6.1. Methods for Protein Identification 164

5.3.6.2. Methods for Quantitation 165

5.3.6.3. Methods for Detecting Protein Modifications 166

5.3.6.4. Proteomics Development Needs 166

5.3.7. Technology Development for Metabolomics 167

5.3.7.1. Measurement Techniques 167

5.3.7.2. Metabolomics Development Needs 169

5.3.8. Technology Development for Other Molecular Analyses 169

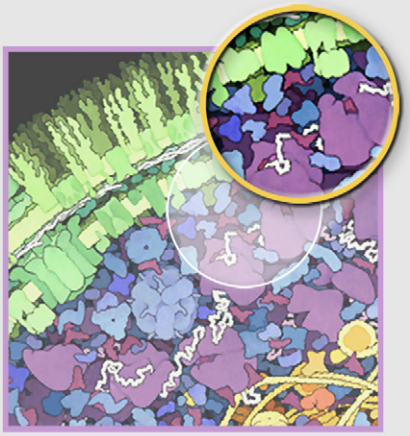
5.3.8.1. Carbohydrate and Lipid Analyses 169

5.3.8.2. Metal Analyses 169

5.3.9. Development of Computational Resources and Capabilities 169

To accelerate GTL research in the key mission areas of energy, environment, and climate, the Department of Energy Office of Science has revised its planned facilities from technology centers to vertically integrated centers focused on mission problems. The centers will have comprehensive suites of capabilities designed specifically for the mission areas described in this roadmap (pp. 101-196). The first centers will focus on bioenergy research, to overcome the biological barriers to the industrial production of biofuels from biomass and on other potential energy sources. For more information, see Missions Overview (pp. 22-40) and Appendix A. Energy Security (pp. 198-214) in this roadmap. A more detailed plan is in Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda, DOE/SC-0095, U.S. Department of Energy Office of Science and Office of Energy Efficiency and Renewable Energy (http://genomicsgtl.energy.gov/biofuels/).

Facility for Whole Proteome Analysis



Identify proteins and other molecules produced by cells in response to environmental cues.

Proteomics Facility

- Measure molecular profiles and their temporal relationships.
- Identify and model key pathways and other processes to gain insights into functions of cellular systems.

The Facility for Whole Proteome Analysis (Proteomics Facility) will be a user facility enabling scientists to analyze microbial responses to environmental cues by determining the dynamic molecular makeup of target organisms in a range of well-defined conditions.

5.3.1. Scientific and Technological Rationale

The information content of the genome is relatively static, but the processes by which families of proteins are produced and molecular machines are assembled for specific purposes are amazingly dynamic, intricate, and adaptive. All proteins encoded in the genome make up an organism's "proteome." Proteins are molecules that carry out the cell's core work; they catalyze biochemical reactions, recognize and bind other molecules, undergo conformational changes that control cellular processes, and serve as important structural elements within cells. The cell does not generate all these proteins at once but rather the particular set required to produce the functionality dictated at that time by environmental cues and the organism's life strategy—a set of proteins that are produced just in time, regulated precisely both spatially and temporally to carry out a specific process or phase of cellular development.

Understanding a microbe's protein-expression profile under various environmental conditions will serve as a basis for identifying individual protein function and will provide the first step toward understanding the complex network of processes conducted by a microbe. Insight into a microbe's expression profile is derived from global analysis of mRNA, protein, and metabolite and other molecular abundance. Characterizing a microbe's expressed protein collection is important in deciphering the function of proteins and molecular machines and the principles and processes by which the genome regulates machine assembly and function and the resultant cellular function. This is not a trivial feat. A microbe typically expresses hundreds of distinct proteins at a time, and the abundance of individual proteins may differ by a factor of a million. Technologies emerging only recently have the potential to measure successfully all proteins across this broad dynamic range; these technologies and others to be further developed will form the facility's core (see Fig. 1. Proteomics Facility Flowchart, p. 157).

Measuring the time dependence of molecular concentrations—RNAs, proteins, and metabolites—is needed to explore the causal link between genome sequence and cellular function (see Fig. 2.

Facility Objectives

- Identify and quantify all proteins, both normal and modified, expressed as a function of time (proteomics).
- Analyze all mRNA and other types of RNA (transcriptomics).
- Analyze all metabolites, the small biochemical products of enzyme-catalyzed reactions (metabolomics).
- Perform other molecular profiling. Lipids, carbohydrates, and enzyme cofactors are examples of other molecular species that can inform investigations of cellular response.
- Carry out modeling and simulation of microbial systems. Test models and inform experimentation, inferring molecular machines, pathways, and regulatory processes.
- Provide samples, data, tools, and models to the community.

Gene-Protein-Metabolite Time Relationships, p. 158). Generally, a microbial cell responds to a stimulus by expressing a range of mRNAs translated into a coordinated set of proteins. Measuring RNA expression (transcriptomics) will provide insight into which genes are expressed under a specific set of conditions and thus the full set of processes that are initiated for the coordinated molecular response. An even-greater challenge will be detection of precursor regulatory proteins or signaling molecules that start the forward progression of a metabolic process. An example is master regulator molecules that simultaneously control the transcription of many genes (see sidebar, Genetic Regulation in Bacteria, p. 67). When activated and functioning, proteins expressed by RNA will yield metabolic products. Each organism has a unique biochemical profile, and measuring the cell's collection of metabolites, "metabolomics," is one of the best and most direct methods for determining the cell's biochemical and physiological status. Each of the molecular species' distinct temporal behaviors and their interrelationships must be understood. In this facility, temporal measurements—snapshots in time—will be made by taking a time series of samples from large-scale cultivations (see Table 1. GTL Data: Thousands of Times Greater than Genome Data, p. 159). The Cellular Systems Facility, by contrast, will nondestructively track processes as they happen within the microbial-community structure.

High-capacity computation is needed to integrate all the data from transcriptomics, proteomics, and metabolomics with additional information obtained from research programs and other GTL facilities. These data will be combined to understand and predict microbial responses to different intracellular and environmental stimuli. Petabytes of data generated from all these different measurements will require a substantial investment in computational tools for reducing and analyzing massive data sets and integrating diverse data types.

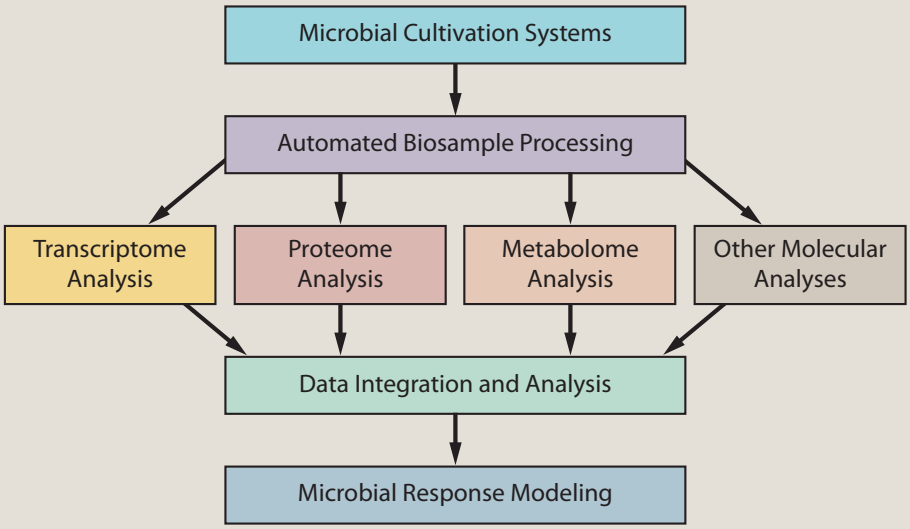


Fig. 1. Facility for Whole Proteome Analysis Core Capabilities.

5.3.2. Facility Description

This user facility will provide capabilities and supporting infrastructure to enable conceptualizing and modeling a cell's molecular response to environmental cues by identifying critical molecular changes resulting from those conditions. The Proteomics Facility, consisting of a 125,000- to 175,000-sq.-ft. building, will house core facilities for controlled growth and analysis of microbial samples. The facility's laboratories will grow microorganisms under controlled conditions; isolate analytes from cells in both cultured and environmental samples; measure changes in genome expression; temporally identify and quantify proteins, metabolites, and other cellular constituents; and integrate and interpret diverse sets of molecular data (see Fig. 1, p. 157). This high-throughput facility will have extensive robotics for efficient sample production and processing with suites of highly integrated analytical instruments for sample analysis.

The facility's computational capabilities will include data-management and -archiving technologies and computing platforms to analyze and track facility experimental data. In addition, computational tools will be established for building and refining models that can predict the behavior of microbial systems. Captured in data, models, and simulation codes, this comprehensive knowledge will be stored in the GTL Knowledgebase to be disseminated to the greater biological community, enabling studies of microbial systems biology.

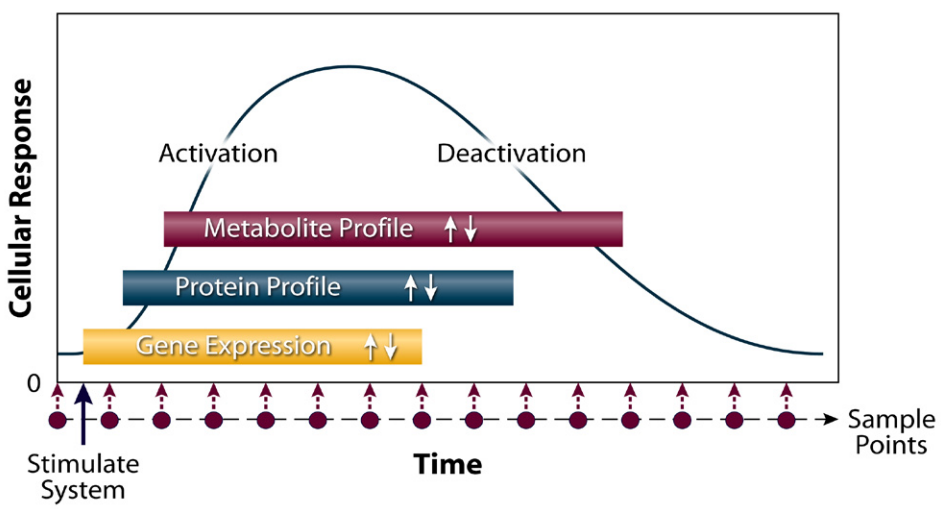


Fig. 2. Gene-Protein-Metabolite Time Relationships. To accurately establish causality between measured gene, protein, and metabolite events, sampling strategies must cover the full characteristic time scales of all three variables. Little is known about the time scale of gene, protein, and metabolite responses to specific biological stimuli or how response durations vary among genes and species. [Figure adapted from J. Nicholson et al. (2002).]

Offices for staff, students, visitors, and administrative support; conference rooms and other common space; and all the equipment necessary to support the proposed facility's mission will be included. The DOE design and acquisition process will include all R&D, design, testing, and evaluation activities necessary to ensure a fully functional facility upon completion.

5.3.2.1. Production Targets

Table 1, p. 159, illustrates the capacity needed for analyzing a single microbial experiment at various levels of comprehensiveness. This facility's goal would be to perform at least tens of such analyses per year using a phased approach, with the initial potential for that number to grow rapidly. Samples will be derived from experiments in mono- and mixed-population cultures and environmental samples.

5.3.3. Technology Development for Controlled Microbial Cultivation and Sample Processing

Automated, highly instrumented, and controlled systems will be developed for producing microbial cultures under a wide range of conditions to permit the high-throughput analysis of proteins, RNA, and metabolites. With the goal of producing and analyzing thousands of samples from single- and multiple-species cultures, technologies must be improved to provide continuous monitoring and control of culture conditions. To ensure the production of valid, reproducible samples, the Proteomics Facility must be able to grow cultures under well-characterized states, measure hundreds of variables accurately, support cultures at a scale sufficient to obtain adequate amounts of sample for analysis, and grow microbial cells in monoculture as well as in nonstandard conditions such as surfaces for biofilms (see Table 1, this page). These cultivation systems will be supported by advanced computational capabilities that allow simulation of cultivation scenarios and identification of critical experimental parameters. This facility will set the standard for cultivation, which other GTL facilities and research programs will use as starting points for their studies.

Table 1. GTL Data: Thousands of Times Greater than Genome Data

Experiment Templates for a Single Microbe

Class of Experiment	Time Points	Treatments	Conditions	Genetic Variants	Biological Replication	Total Biological Samples	Proteomics Data Volume in Terabytes	Metabolite Data in Terabytes	Transcription Data in Terabytes
Simple	10	1	3	1	3	90	18.0	13.5	0.018
Moderate	25	3	5	1	3	1,125	225.0	168.8	0.225
Upper mid	50	3	5	5	3	11,250	2,250.0	1,687.5	2.25
Complex	20	5	5	20	3	30,000	6,000.0	4,500.0	6
Comprehensive	20	5	5	50	3	75,000	15,000.0	11,250.0	15

Profiling Methods

Proteomics: Looking at a possible 6000 proteins per microbe, assuming ~200 gigabytes per sample

Metabolites: Looking a panel of 500 to 1000 different molecules, assuming ~150 gigabytes per sample

Transcription: 6000 genes and 2 arrays per sample ~100 megabytes

Typically, a single significant scientific question takes the multidimensional analysis of at least 1000 biological samples.

This table shows how quickly GTL experiments will generate terabytes (10^{12} bytes) of proteomic, metabolomic, and transcriptomic data. Global proteomics currently generates ~1.0 terabytes (TB) a day with expected 5- to 10-fold increases per year. Not only massive in volume but also very complex, these data span many levels of scale and dimensionality. For example, in a simple study of a microbial system under a single treatment (such as pH or toxin exposure), three different growth states may be studied, with ten samples taken over the growth of the culture. Replicates of each of these samples will be run as part of quality-assurance protocols. This will result in a total of 90 (3 × 10 × 3) analyses and the generation of more than 18 TB of proteomics data, 13.5 TB of metabolomics data, and 0.018 TB of transcriptomics data. If, however, a more complete set of data is taken to achieve greater temporal fidelity and better understand mechanistic response, the amount of data can grow rapidly. This example of growth in data output demonstrates one of the major data-management challenges of GTL. Strategies and technologies for data compression must be developed that avoid "data decimation," which means knowing all the information that must be extracted from raw data before any is discarded. Current proteomics efforts are employing preliminary technologies for near real-time data reduction.

FACILITIES

Biological systems inherently are inhomogeneous; measurements of the organism's average molecular expression profile for a collection of cells cannot be related with certainty to the expression profile of any particular cell. For example, molecules found in small amounts in ensemble samples 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, techniques such as flow cytometry will be used to separate various cell states and stratify cell cultures into functional classes.

Standardized, statistically sound sampling methods and quality controls are essential to ensure reproducibility and interpretability of advanced analyses. Robotics and liquid-handling systems will be developed and automated for initial isolation of proteins and other molecules from microbes, final sample preparation (e.g., desalting, buffer exchange, and sample concentration), and treatment of samples as required for analysis. Microtechnologies such as microfluidic devices will be developed wherever applicable to improve performance and speed, reduce sample handling and potential sample losses, and reduce use of materials and costs (see Table 2. Controlled Cultivation and Sample Processing Technology Development Roadmap, this page).

Table 2. Controlled Cultivation and Sample Processing Technology Development Roadmap

Technology Objectives	Research, Design, and Development	Demonstration: Pilots, Modular Deployment	Integration and Production Deployment	Facility Outputs
Controlled Cell Growth, Analysis Flexible, highly instrumented and monitored cultivation systems Online metabolite monitoring Sample preparation, characterization, stabilization Sample archiving, tracking User environments Community outreach, education	Define and determine: • Appropriate parameters, culture variability • Workflow processes • Scale factors • Hardware, software, instrumentation Develop: • Reactor and instrumentation, interfaces, sampling methods • Reactor-based growth models, simulations • Searchable sample archive • Isotope labeling • High-throughput cultivation, isolation of community members	Pilot: • High-throughput controlled cell growth, processing • Methods for large sample collection • Online analytical systems for high-throughput metabolite measurements • Experiment and sample database • Automation, standardization, protocols Develop methods: • Commensal cocultures • Extremophiles • Biofilms and structures • Sample receipt and delivery	Establish high-throughput pipeline based on defined products, standards, protocols, costs Scale up parallel processes for multiple organisms Process automated, reproducible samples Scale up user-access protocols for sample receipt, growing, delivery	Coordinated high-quality analyses of microbial samples for nucleic acids, proteins, metabolites, and others as needed Detailed cultivation and sampling parameters Efficient, high-capacity, annotated biosample archives User environment for access, protocols, process

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

5.3.3.1. Development Needs for Cultivation

- **New Technologies for Online Monitoring.** New sensors are needed to measure environmental variables, volatile and soluble metabolites, and microbial physiology to monitor and adjust conditions continually to ensure the quality of cell growth.
- **Culture Heterogeneity.** Heterogeneity is found in even the most "homogeneous" cultures produced in continuously stirred tank reactors (chemostats). Individual cells in the culture are at various stages in growth and cellular-division cycles, and subpopulations can form on reactor surfaces. Different types of culture heterogeneity also are caused by stochastic effects in microbial populations (Elowitz et al. 2002). We are just starting to develop techniques for assessing this variability and determining its impact on downstream analyses of harvested biosamples.
- **Biofilms and Structured Communities.** Emerging techniques support the growth of microbial structured communities in the form of, for example, biofilms and clusters. Even in clonal populations, the formation of structures can result in a distribution of distinct and unique phenotypes in the microniches of biofilms and other structures (see sidebar, Life in a Biofilm, p. 18).
- **Definition of Media Components and Culture Parameters.** Such culture parameters as dissolved oxygen, pH, density, and growth rate are important for interpreting the culture's metabolic responses and for providing another level of quality assurance from one experiment to another. Components of growth media influence microbial metabolism and physiology and should be defined chemically to ensure reproducibility and to account for chemical mass balance, an indicator of how the culture is processing nutrients.
- **Large Culture Volumes.** Current methods for proteomics based on mass spectrometry (MS) require large-scale cultivation for the very large number of samples required. Improvements in downstream analytical technologies, however, could reduce sample volumes and the need for such large cultures.
- **Growth in Nonstandard Conditions.** Ideal culture conditions in the laboratory should reflect community conditions in natural environments. Several microbes that DOE is studying either require extremes of salt, pH, temperature, aerobic or anaerobic conditions, and light or they exhibit certain unique phenotypes in microniches with unknown and difficult-to-characterize physicochemical states. Cultivation technologies that accommodate such a range of metabolic requirements must be considered, improved, and, in some cases, developed.

5.3.3.2. Development Needs for Sample Processing

- **Biosample Stabilization.** Harvested biosamples must reflect accurately the conditions under which they were produced. This requires the development and use of harvesting procedures that rapidly and effectively stabilize samples. For example, samples of intracellular metabolites should be quenched as quickly as possible (within a few hundred milliseconds) to maintain in vivo concentrations.
- **Sampling Time Scales.** Gene, protein, and metabolic events within cells operate on significantly different time scales. The resulting gene expression, protein synthesis, cell signaling, and metabolic responses to an environmental stimulus are related functionally but can last from milliseconds to hours. Inferred causal correlations among these different kinds of molecular events depend on well-defined temporal relationships in sampling. Having technologies and methods in place is important for accurately measuring the time-dependent patterns of change for a variety of molecular responses (see Fig. 2, p. 158).
- **Environmental Samples.** Analysis of real environmental samples will be a critical capability of this facility. As methods are refined and made more robust, examining environmental samples with their increased complexity and lack of controls will become more feasible, with protocols supporting these analyses.

5.3.4. Large-Scale Analytical Molecular Profiling: Crosscutting Development Needs

Several technological factors impact the kinds of measurements that can be made on the molecular inventories of cells: (1) limit of detection (the lowest number of molecules that can be detected), (2) dynamic range (ability to detect a low abundance of a molecular species in the presence of other more-abundant molecules), (3) sample complexity or heterogeneity, and (4) analysis throughput. All these factors must be improved to develop technologies that can make the high-throughput molecular measurements required for GTL research.

The kinds of measurements that GTL needs for systems biology will require great improvement in throughput—not just for individual instruments within an analysis "pipeline," but for the entire system. MS technologies today vary in dynamic range from about 10^3 to 10^6. Although usually adequate for proteomic measurements, this dynamic range is not sufficient for global analysis of metabolites. To explore the full range of metabolites of an individual organism today, researchers must use a time-consuming combination of technologies that makes data comparisons and analyses difficult. Another limitation of current technologies is poor detection of molecules present in low numbers. A cell may have only a few copies of some molecules with important biological effects, making them impossible to detect without substantial concentration steps before analysis.

A comprehensive understanding of microbial response can be achieved only by linking and integrating results from many different kinds of molecular analyses. Every technology and method multiplies the scale and complexity of data and analysis (see Table 1, p. 159). Computational methods for designing and managing experiments and integrating data must be part of plans for developing experimental procedures from the ground up.

Exceptional quality control, from cultivation to experimental analysis and data generation, must be maintained to ensure the most reliable data output. To draw meaningful conclusions from transcriptomic, proteomic, and metabolomic studies, researchers need data generated from protocols that have been highly validated in a process similar to that currently used in gene sequencing. This will require understanding error rates and variability in measurements and defining how many measurement replicates are needed for confident identification of biologically significant changes. Today, months are required to measure the proteome of even a simple microbial system, making replicates of proteome measurements impractical for most individual laboratories.

In addition to these crosscutting challenges to multiple analytical methods, research and development are needed for methods and technologies specific to each type of molecular analysis conducted at this facility, as described below.

5.3.5. Technology Development for Transcriptome Analysis

Large-scale RNA profiling involves quantifying and characterizing the entire assembly of RNA species present in a sample, including all mRNA transcripts (the transcriptome) and other small RNAs not translated into proteins (see Table 3. Transcriptome Analysis Technology Development Roadmap, p. 163).

5.3.5.1. Global mRNA Analysis

Microarrays have become a standard technology for high-throughput gene-expression analysis because they rapidly and broadly measure relative mRNA abundance levels. The mRNA expression patterns revealed by microarrays provide insights into gene function, identify sets of genes expressed under given conditions, and are useful in inferring gene regulatory networks. The most common types of microarrays are slide based and affixed with hundreds of thousands of DNA probes, with each probe representing a different gene. In addition to glass slides, probes can be attached to such other substrates as membranes, beads, and gels. When

the probes bind fluorescently labeled mRNA target sequences from samples, the relative mRNA abundance for each expressed gene can be determined. The more target mRNA sequence available to hybridize with a specific probe, the greater the fluorescence intensity generated from a particular spot on an array.

Data from global microarray analysis must be validated with lower-throughput, more-conventional methods such as Northern blot hybridization, as well as real-time polymerase chain reaction that can be used to benchmark these facility results for comparison to researchers' lab measurements.

5.3.5.1.1. Microarray Limitations Requiring R&D

- **Global Quantitative Expression.** Relative abundance of mRNA can be measured, but quantitation is poor.
- **Interpretations of Microarray Results.** Unexpected formation of secondary mRNA structure, cross hybridization, or other factors could produce artificially low expression levels for particular genes. In addition, gene function and regulation based entirely on mRNA expression data may miss functionally related genes not expressed together or may incorrectly predict functional relationships between genes that just happen to be coexpressed. Gene expression is a piece of the systems biology puzzle that also requires proteomic and metabolomic analyses to obtain a comprehensive understanding of gene function and genome regulation.
- **Sensitivity.** The lower limit of detection for current microarray technologies is 10^4 copies of a target molecule, which is not sufficient for many applications. Low-abundance cellular mRNA cannot be detected.
- **Time Resolution.** Today's techniques lack sufficient time resolution to measure constantly changing mRNA levels.

Table 3. Transcriptome Analysis Technology Development Roadmap

Technology Objectives	Research, Design, Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Products
High-Throughput Gene-Expression Profiling Sample processing Data processing Quantitation QA/QC	Define: • Workflow processes • Improved detection limits, reproducibility, dynamic range • Hardware, software • Lab automation, robotics • QC instrumentation, processes Develop: • Multipurpose, multiorganism array platform • In vivo testing platforms • Expression database • Commercial array applications	Expression pipeline optimization, scaleup: • Improved standards, protocols, costs Pilot: • Array processing pipeline • Expression database • In vivo testing pipeline	Establish high-throughput pipeline based on defined requirements, standards, protocols, costs, and adopted industry standards: • Array processing pipeline • Expression-experiment database • In vivo expression-testing pipeline	High-quality, comprehensive expression data linked to experiment archive and culture and sampling data

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

- **Sufficient Replicates.** Running statistically sound numbers of replicate microarray experiments can significantly decrease false-positive results and increase the statistical significance of all ensuing and coordinated experimental results.

5.3.5.2. Small Noncoding RNA Analysis

We have only begun to realize the importance of noncoding small RNA molecules (sRNAs, <350 nucleotides) in many different cellular activities. Many sRNAs are known to regulate bacterial response to environmental changes. Regulatory sRNAs can inhibit transcription or translation or even bind an expressed protein and render it inactive. Other types of sRNAs with elaborate 3D structures have catalytic or structural functions within protein-RNA machines (Majdalani, Vanderpool, and Gottesman 2005).

5.3.5.2.1. sRNA-Analysis Development Needs

- **Finding sRNA Genes.** Even with the availability of complete genomes and computational tools for sequence analysis, finding genes that code for functional sRNAs rather than proteins presents a new computational challenge. Because there are so many different types of sRNAs (with many yet to be discovered) and no genetic code to aid the prediction of sRNA transcripts, more-reliable approaches to sRNA gene discovery require further development. For example, traditional methods such as BLAST and FASTA for comparing the sequences of proteins or protein-coding genes are not as useful for sRNA sequence comparisons.
- **Detecting and Quantifying sRNAs.** Still in its infancy, sRNA analysis cannot tell us how many sRNA genes we should expect to find in a microbial genome. Without reliable sRNA sequence information, experimental screening for sRNAs is difficult. Methods must be developed to isolate various sRNAs and distinguish functional RNA molecules from nonfunctional RNA by-products of cellular activities.

5.3.6. Technology Development for Proteomics

Proteome analyses at the facility will focus on identifying and quantifying both normal and modified proteins expressed by a microbe at a particular time. The most widely used proteomics technologies today include a separation technique such as gel electrophoresis and liquid chromatography combined with detection by mass spectrometry. MS will be used to measure molecular masses and quantify both the intact proteins and peptides produced by enzymatic protein digestion (see Molecular Machines Facility, Table 4. Performance Factors for Different Mass Analyzers, p. 148). Identification of expressed proteins will require both moderate-resolution "workhorse" instruments such as quadrupole and linear ion traps as well as high-performance mass spectrometers capable of high mass accuracy, including Fourier transform ion cyclotron resonance (FTICR) and quadrupole time-of-flight (Q-TOF) mass spectrometers. Data output from these instruments will require extensive dedicated computational resources for data collection, storage, interpretation, and analysis.

Currently, few laboratories are capable of carrying out large-scale proteomics experiments. Specialized technologies needed for proteome analysis are still evolving, and no standards exist for representing proteomics data, making comparisons of results among laboratories difficult. The Proteomics Facility will be a venue for the scientific community to validate these techniques and develop cross-referenced standards. It also will be in the forefront of research into completely new techniques that have capabilities going beyond those currently available (see Table 4. Proteomics Technology Development Roadmap, p. 166). Current techniques are described in the following sections.

5.3.6.1. Methods for Protein Identification

One of two general classes of MS-based approaches for measuring the proteome, gel-based methods use two-dimensional electrophoresis (2DE) to separate complex protein mixtures by net charge and molecular mass. Proteins separated on the gel are extracted and enzymatically digested to produce peptides that can be identified with MS, typically by matrix-assisted laser desorption ionization (MALDI) combined with a TOF

instrument. Recent developments in 2DE separations under nondenaturing conditions have shown that this process yields proteins that retain structural conformations, thus preserving enzymatic activity that holds the possibility of detecting other functional characteristics.

- Increasingly, proteomic techniques use liquid-chromatography (LC) separations coupled with electrospray ionization (ESI) MS for the characterization of the separated peptides or proteins. Intact proteins or peptides generated from enzymatic digestion of proteins are analyzed by direct accurate mass measurement or by tandem mass spectrometry (MS/MS), or some combination of these approaches. MS/MS analysis can provide characteristic spectra that can be searched against databases (or theoretical MS/MS spectra) to identify proteins.
- An alternate approach takes advantage of high mass accuracy of FTICR mass spectrometers to identify proteins, substantially eliminating the need for MS/MS analysis. This approach uses accurate mass and time (AMT) tags for peptides or proteins derived from the combined use of LC separation properties and the accurately determined molecular mass of a peptide or protein. Such measurements allow a certain peptide or protein to be identified among all possible predicted peptides or proteins from a genomic sequence. A database of verified AMT tags for an organism is generated using "shotgun" LC-MS/MS methods for peptide identification as described above. Once this initial investment is made (currently less than a week of work for a single microbe), use of AMT tags can achieve much faster, more quantitative, and more sensitive analyses. These methods will be augmented by new data-directed MS approaches that allow species displaying "interesting" changes in abundances (e.g., between culture conditions), but for which no AMT tag initially exists, to be targeted for identification by advanced MS/MS methodologies (as well as generation of an AMT tag for the species). The combined result will be capabilities to broadly and rapidly characterize proteomes (Lipton et al. 2002) (see Molecular Machines Facility, Table 4, p. 148).

5.3.6.2. Methods for Quantitation

The facility will require that all proteome analyses be quantitative and that the data generated have associated levels of uncertainty so that, for example, changes in protein abundances as a result of a cellular perturbation may be determined confidently. Although MS-based techniques are excellent for protein identification, protein-quantification methods are still under development, and the most-effective approaches are not yet clear.

Challenges for quantitation using MS are related to variations in peptide or protein ionization efficiencies, possible ionization-suppression effects, and other experimental factors affecting reproducibility. Recent research has suggested that quantitative results are achievable in conjunction with LC separations by using very low flow rates with ESI. Although significant effort is needed to develop methods for routine automated measurements, the use of spiked (calibrant) peptides or proteins also provides a basis for absolute quantitation in proteome measurements. Combined with appropriate normalization methods, direct-comparison analyses to understand proteome variation after a cellular perturbation appear to be possible in the future.

In addition, highly precise quantitative measurements are feasible by analyzing mixtures of a proteome labeled with a stable isotope and an unlabeled proteome. These approaches, which introduce a stable-isotope label as an amino acid nutrient in the culture, have the advantage that high-efficiency labeling can be obtained without significant impact on the biological system. Capabilities are envisioned for absolute-abundance measurements and stable-isotope labeling for high-precision analyses that will be beneficial and complementary. In many cases, the facility will apply both methods of quantitation simultaneously to provide precise information for comparison of two different proteomes as well as intercomparison of changes across large numbers of experimental studies.

In addition to limitations in ionization, several other issues must be resolved to achieve better MS-based quantitation: Incomplete digestion of proteins into peptides, losses during sample preparation and separations, incomplete incorporation of labels into samples, and difficulties with quantifying extremely small or large proteins.

Table 4. Proteomics Technology Development Roadmap

Technology Objectives	Research, Design, Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Products
High-Throughput Protein Profiling Sample processing MS for global proteomics Other analysis techniques Data processing and analysis QA/QC	Define: • Workflow processes • Lab automation and robotics • QC instrumentation and processes • Improved detection limits, reproducibility and dynamic range • Hardware, software, instrumentation Develop methods: • Peptide identification and quantitation • Identification of protein modifications • Analysis of intact proteins, including membrane associated proteins	Whole-proteomics pipeline: • Optimization and scaleup • Improved standards, protocols, costs • Pilot of global proteomics database development • Determination of global state of modification of cellular proteins Evaluate and implement: • Hardware advances • Software advances • Instrumentation advances	Establish high-throughput pipeline based on defined standards, protocols, costs	High-quality, comprehensive proteome data linked to experiment archive and culture and sample data

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

5.3.6.3. Methods for Detecting Protein Modifications

Covalent protein modifications (e.g., phosphorylation or alkylation) and other modifications (e.g., mutations and truncations) can affect protein activity, stability, localization, and binding. The majority of cellular proteins are, in fact, modified by one or more chemical processes into their functional form. MS techniques can be used to detect and identify modified peptides. For example, when a phosphate group, lipid, carbohydrate, or other modifier is added to a protein, the modified amino acid's molecular mass changes. Any technique based on mass analysis of peptides, however, can miss modifications on peptides that are not detected. This "bottom-up" analysis recently has been complemented by a "top-down" analysis scheme in which intact proteins are analyzed by ESI FTICR MS. This top-down approach has provided greater detail on both the types and sites of these modifications. Improvements in the ability to effectively ionize a wider range of intact proteins are needed, however.

5.3.6.4. Proteomics Development Needs

- **Analyzing Intact Proteins.** Although today's MS techniques are well suited for analyzing peptides produced by enzymatic digestion of proteins, improved capabilities for the MS analysis of intact proteins are needed, especially higher molecular-weight proteins and membrane-associated proteins. In both cases, ionization is a major limitation.
- **Improving Separation Methods.** The proteome's complex, heterogeneous nature requires separation of peptides or proteins before analysis. Improved separation technologies are needed to provide higher-speed, yet higher-performance, separations. A longer-term solution may include improved MS-based approaches

that use selective ionization and ion mass selection (e.g., MS/MS, gas-phase reactions) to minimize the need for high-performance separations.

- **Improving Dynamic Range.** High-throughput MS-based analysis at the Proteomics Facility will require at least a tenfold improvement in dynamic range over today's best performance.
- **Measuring Protein Turnover Rates.** The ability to introduce stable-isotope labels (e.g., in cultures) opens the doors to global measurements of protein-turnover rates, based on the partial incorporation of stable-isotope labels observed in the isotopic distributions for peptides or proteins measured with mass spectrometers in proteome studies. These measurements reflect the rates at which proteins are being produced, destroyed, or modified; they can be expected to be complex (i.e., vary with protein subcellular localization) and provide valuable data not otherwise obtainable on important aspects of the biological systems.
- **Developing New Ionization Methods.** Ionization methods and the mechanisms underlying their variability are not well understood. New or improved methods are needed for greater ionization efficiency to extend current detection limits and more-uniform ionization to improve quantitative capabilities.
- **Developing Computing Tools and Data Standards.** Such tools are needed to handle data-analysis bottlenecks. Although commercial software packages for data interpretation are quite advanced, additional improvements are needed for automatic analysis of large volumes of data and incorporation of data into larger data structures and the GTL Knowledgebase.

5.3.7. Technology Development for Metabolomics

Metabolites are the small molecular products (molecular weight <500 Da) of enzyme-catalyzed reactions. Metabolite levels are determined by protein activities, so a comprehensive understanding of microbial systems is not possible without measuring and modeling these small molecules and integrating the information with data from proteomics and other large-scale molecular analyses.

5.3.7.1. Measurement Techniques

The high chemical heterogeneity of metabolites requires that technologies be combined to fully explore the entire metabolome of even an individual organism. This heterogeneity, however, also means that metabolome components are much more varied in nature than proteome components and therefore potentially much easier to measure (see Table 5. Global Metabolite Analysis Technology Intercomparison, p. 168). A variety of separation and MS techniques and nuclear magnetic resonance (NMR) commonly are used to measure the metabolome.

- **MS and Chromatographic Separations.** Multiple forms of MS analyzers, including TOF, quadrupole and linear ion traps, and FTICR, can be combined with different separation technologies that have a variety of advantages and disadvantages. While thin-layer chromatography and gel electrophoresis have been combined successfully with MS, the two most common approaches include gas chromatography (GC) MS and LCMS.
  - **Gas Chromatography MS.** Gas chromatography can provide high-resolution separations of many chemical compounds, and MS is a very sensitive method for detecting and quantifying most small organic compounds. For quantitative measurements, an isotopically labeled analogue of the target molecule is required for optimum measurement accuracy. A major drawback is that most metabolites are polar and thus not volatile enough to be analyzed by GC methods. These polar compounds therefore must be derivitized into less polar, more volatile forms before GCMS analysis. This approach is used widely, but the chemical-derivativization steps can decrease sample throughput and introduce sample loss.
  - **Liquid Chromatography MS.** Also used in proteomics analyses, LCMS circumvents the need for derivitization required by GCMS. Like GCMS, LCMS is highly sensitive and capable of detecting

attomoles of target compounds. LCMS, however, generally provides lower-resolution separations than GCMS, which can limit its applicability in metabolite analyses involving more than 1000 species. Recent progress in higher chromatographic separations using "ultraperformance" liquid chromatography shows the potential to provide increased chromatographic resolving power (more GC-like peak resolution) that will permit enhanced detection and quantitation capabilities with shorter run times. LC can be interfaced with a variety of mass analyzers, providing detailed information on metabolite identification at very low detection limits. As with GCMS, isotopically labeled standards are required for quantitative measurements with very high accuracy. These assays can be run on such widely available instruments as quadrupole or linear ion traps. In addition, higher-performance MS instrumentation such as FTICR can be used to obtain high mass accuracy as an aid to identify metabolites.

- **Nuclear Magnetic Resonance.** One of NMR's advantages is its noninvasive, nondestructive nature that can be used to generate metabolic profiles. By analyzing samples in a liquid state, NMR can be adapted for automation and robotic liquid handling. An important NMR limitation is sensitivity, but several methods being studied have the potential to overcome this limitation. For example, recent research has shown that angular momentum of hyperpolarizable gases like xenon can increase dramatically the number of detectable spins. This has the potential to improve NMR sensitivity by a factor of 20,000. Interfacing

Table 5. Global Metabolite Analysis Technology Intercomparison

	GC-MS	LC-MS	NMR
Strengths	Highly sensitive detection of small, nonpolar organic compounds Robust Highly reproducible Well-developed databases Well-established techniques for quantitative measurements Use of high-performance mass analyzers, such as FTICR, to provide accurate mass measurement and minimize the need for separations	Highly sensitive detection High throughput Minimized need to derivitize molecules prior to analysis Potential for single-cell analysis Use of high-performance mass analyzers, such as FTICR, to provide accurate mass measurement and minimize the need for separations	Structural information provided Nondestructive Direct analysis of liquids Highly reproducible Automatable Dynamic range similar to MS
Weaknesses	Derivitizing less volatile metabolites lowering throughput and introducing potential for sample loss Difficult to discover new compounds	Poor analytical reproducibility in multivariate setting Ion suppression and matrix effects Lower resolving power than GC, leading to poor separation of molecules in complex matrices	Sensitivity Resolution Limited application to complex mixtures
Development Needs	Robustness Improved chromatographic resolving power Improved dynamic range Metabolite databases Computational tools for predicting metabolites		Robustness Dynamic range Cryogenic probes Microprobes and nanoprobes Robust interfaces with chromatography

The table above compares and contrasts strengths, weaknesses, and development needs of technologies for use in a high-throughput production environment.

NMR with chromatographic methods such as LC can resolve molecular species that usually are overlapped in the spectra, thus improving detection and structural assignments.

- **Metabolic Flux Analysis (MFA).** MFA is used to quantify all the fluxes in a microorganism's central metabolism. To measure metabolic fluxes, a ^{13}C-labeled substrate is taken up by a biological system and distributed throughout its metabolic network. NMR and MS technologies then can measure labeled intracellular metabolite pools. Intracellular fluxes are calculated from extracellular and intracellular metabolic measurements. Currently, MFA can be applied only to a highly controlled, constantly monitored system in a stationary metabolic state. MFA's main benefit is the generation of a flux map to identify targets for genetic modifications and formulate hypotheses about cellular-energy metabolism.

5.3.7.2. Metabolomics Development Needs

- **Defining Metabolic Data Standards.** Currently, methods are not standard for formatting, storing, and representing metabolic data.
- **Developing Standardized, Comprehensive Databases of Metabolites.** Although many of the most common metabolites are catalogued and commercially available, the most biologically interesting molecules are unknowns produced by metabolic reactions unique to specific organisms or organism interactions.
- **Developing Methods for Studying Multimetabolite Transport Processes.** Transporters regulate metabolic concentrations just as much as enzymes in some cases.

Table 5, p. 168, compares and contrasts the strengths, weaknesses, and development needs of technologies discussed above. Table 6. Metabolite Profiling Technology Development Roadmap, p. 170, outlines steps in preparing the appropriate mix of these technologies for a high-throughput production environment.

5.3.8. Technology Development for Other Molecular Analyses

5.3.8.1. Carbohydrate and Lipid Analyses

Macromolecules such as lipids and carbohydrates make up cell surface and structural components, impact the function of proteins through covalent modifications, and, as substrates and products of enzyme activities, serve as key indicators of active metabolic pathways. Organic and metallic cofactors, present in many molecular machines, play essential roles in protein folding, structure stabilization, and function. Some current technologies used to analyze these molecules include LC, MS, and NMR. Methods for lipid analysis are mature, but new technologies for carbohydrate analysis are needed. A major obstacle will be to distinguish among many different chemical entities with similar properties and isomers.

5.3.8.2. Metal Analyses

Metal ions are present in many molecular machines relevant to DOE missions. Technologies are needed for measuring metal abundance, coordination state, levels of metalloproteins, and metal trafficking in cells and communities. Current metal-analysis technologies include optical emission and absorption, inductively coupled plasma (ICP) MS, X-ray spectroscopy, electrochemistry, and others. They are relatively mature compared with other global analyses but may need further development to meet the facility's specific needs.

5.3.9. Development of Computational Resources and Capabilities

Computing will be an integral part of all activity within this facility: Managing workflow, controlling instruments, tracking samples, capturing bulk data and metadata from many different measurements, analyzing and integrating diverse data sets, and building predictive models of microbial response. Databases and tools will be created to give the scientific community free access to all data and models produced by the facility (see Table 7. Computing Roadmap, p. 171).

Table 6. Metabolite Profiling Technology Development Roadmap

Technology Objectives	Research, Design, Development	Demonstration: Pilots, Modular Deployment	Integration, Production Deployment	Products
High-Throughput Metabolite Profiling LC/MS and NMR methods for metabolite discovery Sample processing Data processing Analysis, quantitation, QA/QC	Define requirements: • Workflow processes • Detection limits, reproducibility, dynamic range • Lab automation, robotics, QC • Robust LC-NMR techniques • Hardware, software, instrumentation Develop methods: • Identification and quantitation • QA/QC with metrics • Sample processing	Establish pilot metabolite-profiling pipeline Optimize, scale up Develop improved standards, protocols, costs	Establish high-throughput pipeline based on defined requirements, standards, protocols, costs	High-quality, comprehensive, metabolite-profiling data linked to experiment archive and culture and sampling data

To develop and incorporate the necessary technologies and methods into a high-throughput production environment, a phased process will be followed as described in this roadmap. The process includes research, design, and development; modular and pilot-scale deployment; and final integration and scaleup into operational procedures.

- State-of-the-art systems for tracking and maintaining accurate metadata for all experimental samples (e.g., culturing details, sample-processing methods used).
- High-performance computational tools and codes for efficiently collecting, analyzing, and interpreting highly diverse data sets (e.g., MS data for proteins and metabolites, microarrays, and 2DE gel images). Tool capabilities, including data clustering, expression analysis, and genome annotation, would be linked closely to advances in computing infrastructure being proposed by DOE.
- Databases, biochemical libraries, and software for interpreting spectra and identifying peptides and metabolites. Mass spectra for most metabolites are not in standard libraries. Organism-specific metabolic databases are needed.
- Computational tools for abstracting network and pathway information from expression data and genome annotation. These tools will be used for building mathematical models that represent subcellular systems responsible for protein expression and proteome state (including modified proteins) as a function of conditions. Simulation would be employed to evaluate the state of knowledge contained in these models and validate the accuracy of experimental parameters.
- Database development for expression measurements, metabolome measurement, and networks and pathway systems, models, and simulation codes that may exceed petabytes.

Table 7. Computing Roadmap: Facility for Whole Proteome Analysis

Topic	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment
LIMS and Workflow Management Participate in GTL cross-facility LIMS working group Develop technologies and methods to: • Manage massive dataflow • Process and integrate data • Manage workflow • Conduct QA/QC • Deploy collaborative tools for shared access to data and processes	Archival storage systems Prototype bulk data capture and retrieval systems Prototype inter- and intralab limited LIMS Shared LIMS and workflow technology for each analytical capability	LIMS for each analytical pipeline Data archives for each analytical pipeline Inter- and intralab LIMS	Establish LIMS for each analytical pipeline workflow Products: Output data products Cross-facility access and tracking Information management systems and automation Efficient, analytically rigorous pipelines Components and integration to GTL process
Bioinformatics Participate in GTL cross-facility working group for data representation and standards Provide user environments, community access, database development Integrate data-analysis methods Develop large-scale integrated experiment designs, analysis pipelines	Workflow processes and database needs Evaluation of technical solutions Large-scale storage and retrieval solutions Entire workflow processes and methods for experimentation and analysis Algorithms Quality control and assessment measures	Statistically designed experiments Multidimensional data-analysis and integration tools for large-scale experimentation Multilevel databases for bulk and derived data for each profiling method Analysis pipeline for derived data Community-access systems Cross-facility data-sharing processes and analysis methods Archival, computing, and network capacity to match demand	Bulk data archives for key data sets Process to link archives to production activities Local facility data archive Cross-facility data-sharing processes and analysis methods Mature bulk data archives, analysis piles Scaleup of archival activities, computing, and network capacity to match demand Products: • Whole proteome analysis for each GTL organism • Experiment templates and data sets for modeling and simulation • Defined experiment archive integrated with data and analysis from each analytical pipeline • Molecular profiling context-dependent database
Computing Infrastructure Participate in GTL cross-cutting working group for computing infrastructure Establish scientific computing with massive data reduction, archival storage application development Develop infrastructure: hardware, software, code control, libraries, environments Use ultrahigh-speed internet connection to GTL facilities	Operations process Computational architecture Large-scale data mining Access and security plans and processes Performance and quality metrics of service Capacity planning Backup and recovery strategy Testing plans Workflow Dev-Test-Pro strategy for implementations	Test network Development environment Validation methods Data archive Access methods Storage and retrieval methods Application integration and implementation Production infrastructure Cross-facility data sharing Infrastructure: hardware, software, and network	Production environment and data archive Bulk data archives for key data sets Process to link production activities to local facility data archive Cross-facility data-sharing processes and analysis methods Mature bulk data archives and analysis pipelines

FACILITIES