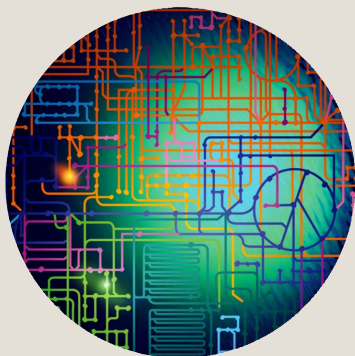


## 5. GTL Facilities

<b>5.0. Facilities Overview</b>	101
5.0.1. Science and Technology Rationale	102
5.0.2. A New Trajectory for Biology	104
5.0.3. Capsule Facility Descriptions	104
5.0.3.1. Facility for Production and Characterization of Proteins and Molecular Tags	104
5.0.3.2. Facility for Characterization and Imaging of Molecular Machines	105
5.0.3.3. Facility for Whole Proteome Analysis	105
5.0.3.4. Facility for Modeling and Analysis of Cellular Systems	106
5.0.4. Relationships and Interdependencies of Facilities	106
5.0.5. Research Scenarios	107
5.0.6. Facility Development	107
<b>5.1. Facility for Production and Characterization of Proteins and Molecular Tags</b>	111
<b>5.2. Facility for Characterization and Imaging of Molecular Machines</b>	139
<b>5.3. Facility for Whole Proteome Analysis</b>	155
<b>5.4. Facility for Analysis and Modeling of Cellular Systems</b>	173

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, the Bioenergy Research Center, will concentrate on overcoming biological barriers to the industrial production of biofuels from biomass and on other potential energy sources such as biophotolytic decomposition of water to hydrogen. 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://doegenomestolife.org/biofuels/>).



To address the analytical and computational capabilities needed to put the GTL research program on track for creating a science foundation for DOE missions, workshops were held between June 2002 and June 2004. Much of the material in this chapter was drawn from the outputs of those workshops, in which nearly 800 different individuals participated. For a list of GTL workshops, meetings, and links to workshop reports, see Appendix D. GTL Meetings, Workshops, and Participating Institutions, p. 239.

## Facilities Overview

The proposed GTL user facilities for 21<sup>st</sup> Century biology and biotechnology will be a major strategic asset in achieving DOE mission goals in industrial biotechnology—a critical arena of national economic competitiveness. The facilities will enable a new era in biology, building on the national investment in genomics.

The research community increasingly is recognizing the need for global analysis of myriad simultaneous cellular activities and is calling for a new research infrastructure. “Progress in microbiology always has been enabled by the technology available, a fact that is still true today. However, many researchers are stymied by lack of access to the expensive instruments that would enable them to make the greatest strides.” (Schaechter, Kolter, and Buckley 2004, p. 13; see also Aebersold and Watts 2002; Buckley 2004a; Stahl and Tiedje 2002).

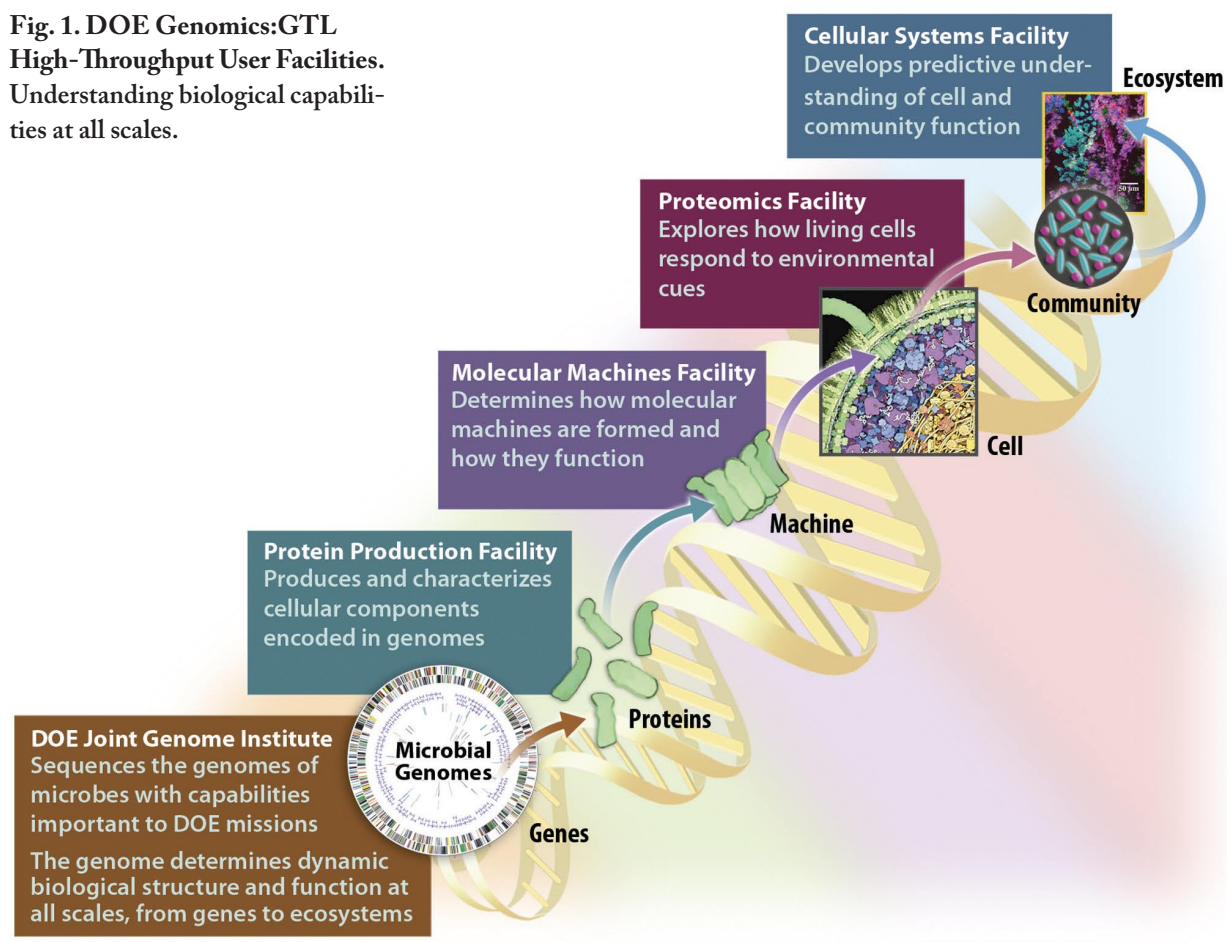
### 5.0.1. Science and Technology Rationale

In the simplest interpretation, an organism’s genome contains encoded information to produce the proteins that are the cell’s workhorses. Central to the strategy of GTL and the GTL facilities suite, however, is the fact that information encoded in the genomes of microbes and the metagenomes of microbial communities goes well beyond encoding proteins. Genomes control microbial structure and function at many spatial and temporal scales—molecular, machine, cellular, community, and ecosystem—through an intricate set of interrelated and communicating regulatory and control processes (see Fig. 1. DOE Genomics:GTL High-Throughput User Facilities, p. 103). Microbes display such strong interactions that capabilities are needed to explore these systems in a comprehensive and integrated way at all levels. Proteins and even microbes are thought to function rarely in isolation.

In the Missions Overview, our example mission descriptions demonstrate that the technical challenges of these analyses and the scale of the systems that must be understood exceed any existing capabilities (see Missions Overview, Tables 1–9, beginning on p. 26; 3.2.2. Science and Technology Milestones, p. 44; and sidebar, High-Throughput Model Guides Future Facilities, p. 6). Facilities must be established to dramatically improve research performance, throughput, quality, and cost.

Examples of performance challenges include producing and characterizing complex proteins (e.g., membrane and multidomain); isolating,

**Fig. 1. DOE Genomics:GTL High-Throughput User Facilities.**  
Understanding biological capabilities at all scales.



characterizing, and modeling large or tenuous molecular machines; measuring the full molecular profile of microbial systems; and imaging molecules as they carry out their critical functions in cells in structured communities. Examples of throughput challenges include providing insight into the functions of hundreds of thousands of unknown genes and their modifications; processing thousands of molecular machines; analyzing molecular profiles of thousands of microbial samples under different conditions; and spanning the full range of conditions and processes governing microbial-community behaviors. Quality control includes developing and implementing strict protocols and providing the most sophisticated diagnostics. High-throughput methods and resource sharing among community members will lower the unit cost for production and analyses.

Figure 1, this page, depicts facilities focused on building an integrated body of knowledge about behavior, from genomic interactions through ecosystem changes. Simultaneously studying multiple microbial systems related to various mission problems is powerfully synergistic because enduring biological themes are shared and general principles governing response, structure, and function apply throughout. The biology underlying the challenges of one mission will inform those of the others. Accumulating the data as it is produced, the GTL Knowledgebase and the computational environment that GTL will create will act as the central nervous system of the facilities and program, allowing this information to be integrated into a predictive understanding.

The Office of Science has a tradition of strategic basic research in a multidisciplinary team environment for national missions. These facilities will bring together the biological, physical, computational, and engineering sciences to create a new infrastructure for biology and the industrial biotechnology needed for the 21<sup>st</sup> Century. DOE's technology programs can work with industry to apply such capabilities and knowledge to a new generation of processes, products, and industries.

## 5.0.2. A New Trajectory for Biology

As we have learned from the genome projects, consolidating capabilities and focusing on aggressive goals will drive dramatic improvements in performance and cost (Fig. 2. Putting Biology on a New Trajectory, this page). As depicted in Fig. 2, GTL facilities will accelerate discovery and reduce the time for useful applications. With this higher level of performance, microbial systems biology is tractable and affordable to support the next generation of industrial biotechnology for the coming decade and beyond.

## 5.0.3. Capsule Facility Descriptions

The GTL facilities provide a complementary set of technologies and products. Two facilities are focused on analysis of properties and functions of cellular components, proteins, and molecular machines:

- Facility for Production and Characterization of Proteins and Molecular Tags
- Facility for Characterization and Imaging of Molecular Machines

Two are focused on analysis of microbial-system responses and functions at the molecular, cellular, and community levels:

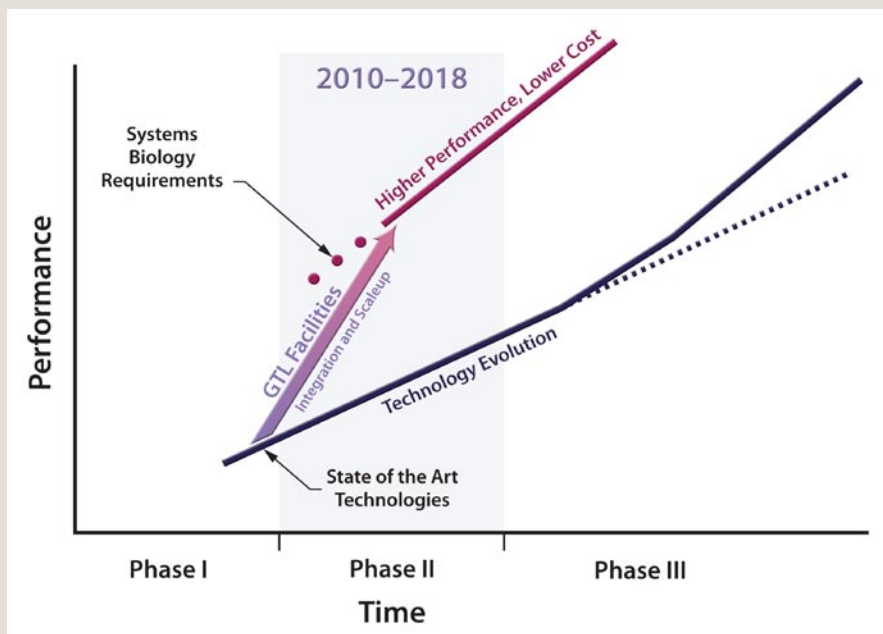
- Facility for Whole Proteome Analysis
- Facility for Modeling and Analysis of Cells and Communities

The ensuing chapters of this roadmap discuss each facility in further detail. Capsule facility descriptions follow.

### 5.0.3.1. Facility for Production and Characterization of Proteins and Molecular Tags

The Protein Production and Characterization Facility will use DNA sequence to make proteins and reagents for interrogating cell function. Specifically, this facility will have the capability to produce all proteins encoded in any genome on demand; create molecular tags that allow each protein to be identified, located, and manipulated in living cells; and, to gain insights into function, perform biophysical and biochemical characterizations of proteins produced. Using high-throughput in vitro and in vivo techniques will lower the

**Fig. 2. Putting Biology on a New Trajectory.**



cost of producing proteins to levels that will allow comprehensive analysis of all proteins within the cell. The facility's products and analysis capabilities will be made available to all scientists.

In parallel with protein production will be the generation of "affinity reagents." These small proteins or nucleic acids will permit the detection and tracking of individual proteins in living systems, including complex molecular assemblages; the intracellular position of all proteins and their spatial dynamics; if exported, the extracellular localization and interaction with other community members; and techniques for manipulating protein activity in the environment.

Core facility 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 system and robotics, and production-strategy determination

### 5.0.3.2. Facility for Characterization and Imaging of Molecular Machines

The Molecular Machines Facility will identify and characterize molecular assemblies and interaction networks. It will have capabilities to isolate and analyze molecular machines from microbial cells; image and localize molecular machines in cells; and generate dynamic models and simulations of the structure, function, assembly, and disassembly of these complexes. The facility will identify molecular machine components, characterize their interactions, validate their occurrence and determine their locations within the cell, and allow researchers to analyze the thousands of molecular machines that perform essential functions inside a cell. It will provide a key step in determining how the network of cellular molecular processes works on a whole-systems basis by completely understanding individual molecular machines, how each machine is assembled in 3D, and how it is positioned in the cell with respect to other components of cellular architecture.

Core facility instrumentation:

- Robotic culturing technologies to induce target molecular machines in microbial systems and supporting robotic techniques for molecular complex isolation
- Numerous sophisticated mass-spectroscopy and other techniques specially configured to analyze samples of purified molecular machines for identification and characterization of complexes
- Various advanced microscopies for intracomplex imaging and structure determination
- Imaging techniques for intracellular and intercellular localization of molecular complexes
- Computing and information systems for modeling and simulation of molecular interactions that lead to complex structure and function

### 5.0.3.3. Facility for Whole Proteome Analysis

The Proteomics Facility will be capable of gaining insight into microbial functions by examining samples to identify (1) all proteins and other molecules that a microbe (or microbial community) creates under controlled conditions and (2) key pathways and other processes. 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 a better understanding of how and why an organism turns portions of its genome "on" and "off." Facility users will achieve a comprehensive understanding of microbial responses to environmental cues by identifying, quantifying, and measuring changes in the global collections of proteins, RNA, metabolites, and other biologically significant molecules. These molecules, including lipids, carbohydrates, and enzyme cofactors, are important in understanding biological processes mediated by proteins. Integrating diverse global



## FACILITIES

data sets, the facility will develop computational models to predict microbial functions and responses, inferring the nature and makeup of metabolic and regulatory processes and structures.

Core facility 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 instrumentation capable of analyzing the molecular makeup of ensemble samples with thousands of diverse molecular species
- High-performance computing and information capabilities for modeling and simulation experiments of microbial-system functionalities under different scenarios to inform the design of experimental campaigns focused on systems-level goals and to infer microbial-system molecular processes from ensuing data

### 5.0.3.4. Facility for Modeling and Analysis of Cellular Systems

The Cellular Systems Facility will be the capstone for the ultimate analytical capabilities and knowledge synthesis to enable a predictive understanding of cell and community function critical for systems biology. The facility will concentrate on the systems-level study of living cells in complex and dynamic structured communities. Imaging methods will monitor proteins, machines, and other molecules spatially and temporally as they perform their critical functions in living cells and communities. Microbial communities contain numerous microniches within their structures that elicit unique phenotypic and physiological responses from individual species of microbes. We need to be able to analyze these niches and the microbial inhabitants within. This grand challenge for biology must be addressed before scientists can predict the behavior of microbes and take advantage of their functional capabilities. Modeling in the facility will describe essential features of these biological interactions with the physicochemical environment and predict how the system will evolve in structure and function.

Core facility instrumentation:

- Highly instrumented cultivation technologies to prepare structured microbial communities to simulate natural conditions under highly controlled conditions
- Instruments integrating numerous analytical imaging techniques that can spatially and temporally determine, in a nondestructive way, the 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

### 5.0.4. Relationships and Interdependencies of Facilities

Each of the facilities is technically distinct in the nature of its instrumentation, methods, and overall goals. All will be centered around either production lines designed to maximize quality and throughput and reduce unit costs, the development and operation of frontier instrumentation or unique suites of instrumentation to reach new levels of performance, or combinations of both. While each can serve a user community for a wide range of independent studies, the suite of facilities has complementary strengths and core technologies that together can help provide complete systems knowledge. Figure 3. GTL Facilities: Core Functions and Key Interactions, p. 108, displays how each facility's core functions are complementary to those of the other facilities. The key interactions shown demonstrate their interdependencies and necessary exchange of all information through the GTL Knowledgebase and the program's communication and computing infrastructure.

### 5.0.5. Research Scenarios

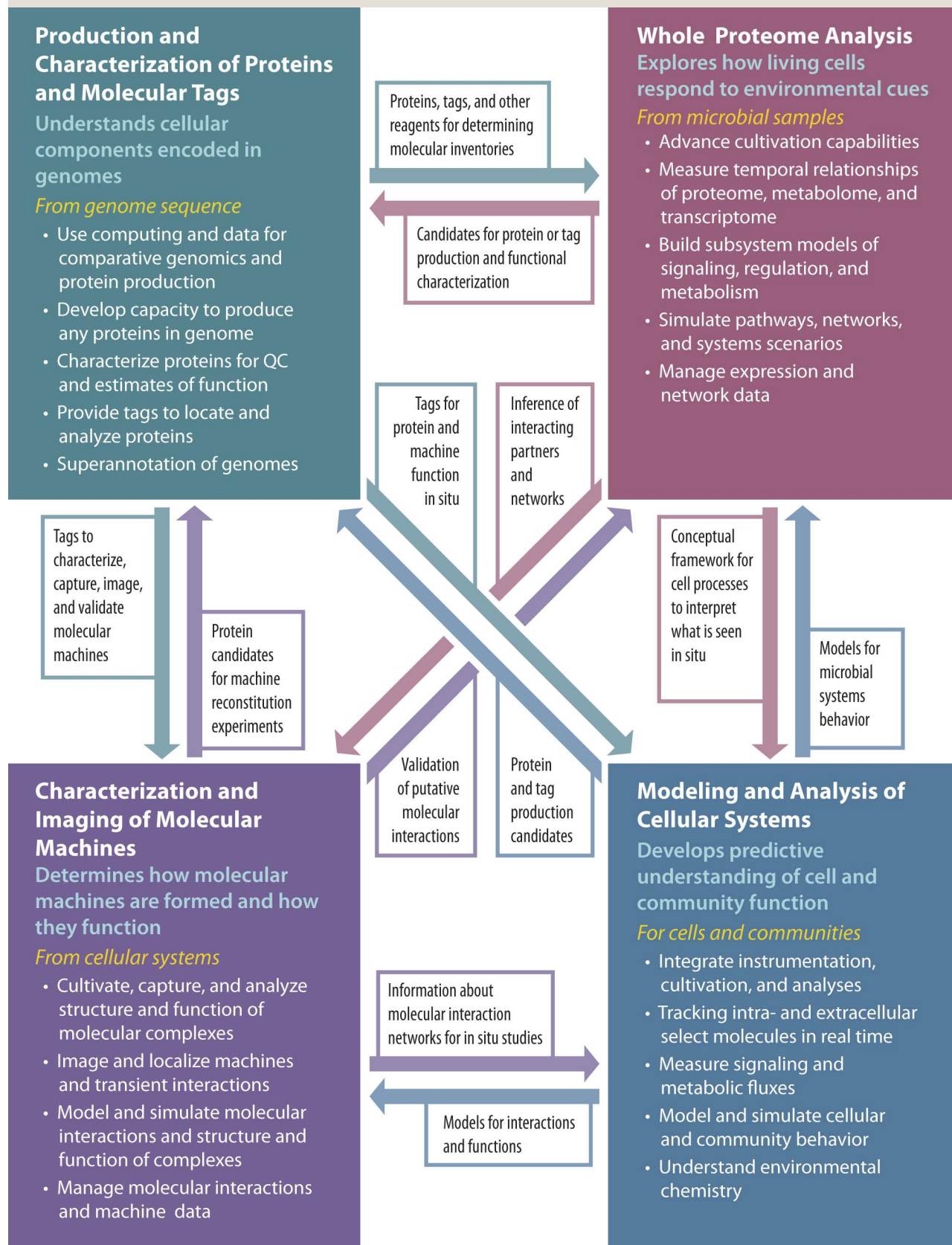
As described in the Missions Overview and related appendices, each mission example has a unique endpoint and research strategy for developing the needed understanding, predictive models, and research capabilities. Table 1. Research Scenarios on Microbial Processes, p. 109, and Table 2. Science Roadmaps for Natural Systems, p. 110, present conceptual research-scenario roadmaps for six cases as illustrations related to Science Milestones and GTL Facilities. Although these systems and problems cover a breadth of microbial phenomenology and system behaviors, they can be studied using the same foundational capabilities. Each of the GTL milestones, as denoted in the left column of Tables 1 and 2, drives the technical core of the facilities, where capabilities resulting from milestone R&D can be scaled up and integrated.

### 5.0.6. Facility Development

The facility acquisition process will employ project-management practices similar to DOE Order 413.3 Facilities Project Management. The facilities budget will include all costs for the conceptualization, design, R&D and testing, and acquisition of the necessary conventional facilities, instrumentation, computers and software, and supporting technologies, training, and installation of fully operational production lines and analytical facilities upon completion of the project. The process will involve participants from national laboratories, academia, and industry in the necessary workshops and working groups to determine the technical scope and scale of the facilities, technical priorities, and technology development. Many of the long-lead and crosscutting development needs are outlined in the GTL Development Summary chapter. This roadmap is meant to be a starting point for the intensive conceptualization and planning that must occur for successful design, acquisition, and operation of these facilities.

# FACILITIES

Fig. 3. GTL Facilities: Core Functions and Key Interactions.





**Table 1. Research Scenarios on Microbial Processes: Relationship to Science Milestones and Facilities**

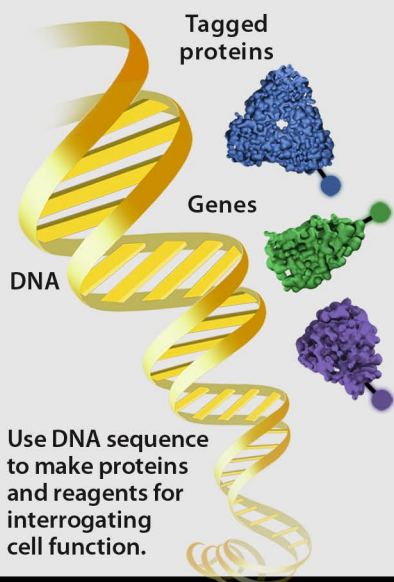
Progression of GTL Science Milestones and Facilities	<i>Conceptual Science Roadmaps for Microbial Energy and Environmental Processes</i>		
	<b>Convert Sunlight to Hydrogen and High-Hydrogen Fuels</b>	<b>Convert Cellulose to Fuels</b>	<b>Reduce Toxic Metals in Subsurface Environments</b>
<b>Milestone 1: Determine the Genome Structure and Potential of Microbes and Microbial Communities</b>  Facility for the Production and Characterization of Proteins and Molecular Tags  Facility for Characterization and Imaging of Molecular Machines	Analysis of hydrogenase families across microbial species: Screen nature for new variants  Range of hydrogenase properties  Suite of heterologous expression hosts  Characterization of partners, energetics, structures, post-translational modifications  Wide range of mutations, variations created and screened  Functional and structural analysis of machines	Wide range of microbes surveyed for cellulases, ligninases, and other glucosyl hydrolases  Partners and structural information established  Structure and imaging of interactions important to efficient function	Survey of subsurface species and genomic potential  Comparative genomics and superannotation  Generation of knockouts, mutations, transmembrane structures to understand function
<b>Milestone 2: Develop a Systems-Level Understanding of Microbial and Community Function and Regulation</b>  Facility for Whole Proteome Analysis  Facility for Analysis and Modeling of Cellular Systems	Oxygen sensitivity of hydrogenases  Electron-transfer reactions and limitations  Reverse-reaction mitigation  Partitioning of electrons between hydrogenases and competing pathways  Light capture  Biophotovoltaic antenna	Proteome analysis of expression and regulation  Fundamental mechanisms of cellulose deconstruction  Transport of sugars  Measurement of electron transport chains' redox state, control of electron fluxes  Carbon partitioning in cells: Carbon, NAD, NADPH, ATP, ADP	Cellular response to environmental stimuli  Proteomics, transcriptomics, and metabolomics to elucidate regulation and responses  Intra- and intercellular communications  Cells in structures such as biofilms  Growth processes, toxicity responses, energy transfer, metabolic responses  Microbe-mineral interactions
<b>Milestone 3: Develop the Knowledgebase, Computational Methods, and Capabilities to Advance Understanding of Complex Biological Systems and Predict Their Behavior</b>  GTL Integrated Computational Environment for Biology	Pathway models—energetics, electropotential, docking, proton fluxes, cofactors  Computational tools for rational design  Suites of hosts, pathway cassettes  Modeling and measurement of pathways, fluxes, regulation	Design of organisms capable of utilizing all sugars  Optimization of sugar transport, regulation  Redesign of cellulose structure  pH- and temperature-tolerant microbes  Principles for enzyme redesign	Modeling capable of visualizing realistic biochemical pathways in cells  Interactions of membrane proteins with contaminants and solid-phase electron acceptors  Design of experiments in cultured and natural systems
<b>Missions Outputs</b>  <b>Systems Engineering</b>	In vivo systems  Processes captured in nanostructures, biomimetic systems  System design: Light harvesting, conversion to hydrogen or fuel, robustness to oxygen, regulation  Transgenic approaches	Improved cellulases and production methods to reduce costs, improve stability  Modularized processing to reduce transportation of feedstock  Sensors for biomarkers and chemical intermediates	Assessment of long-term cellular and system behavior  Remediation strategies  Sensors for coupled biochemical and geochemical measurements in situ

**Table 2. Science Roadmaps for Natural Systems:  
Relationship to Science Milestones and Facilities**

Progression of GTL Science Milestones and Facilities	Conceptual Science Roadmaps for Natural Systems		
	Oceans: Photosynthetically Driven Biological Pumps for Carbon and Energy in Aquatic Systems	Terrestrial: Microbes in Ecological Communities, Carbon and Nutrient Cycles	Deep Subsurface: Microbial Community Processes for Mitigation of Toxic Chemicals and Metals
<b>Milestone 1: Determine the Genome Structure and Potential of Microbes and Microbial Communities</b>  Facility for the Production and Characterization of Proteins and Molecular Tags  Facility for Characterization and Imaging of Molecular Machines	Single-cell and environmental community sequence  Heterotrophs, autotrophs, viruses, and “twilight zone” organisms  Comparative analyses of rhodopsin, hydrogenase genomes  Gene synthesis and manipulation	Single-cell and community sequence in situ and in vitro  Organisms related to processes in soils  Genome annotation	Single-cell and community sequence in situ and in vitro to identify members, functions  Superannotation, genome plasticity effects  Metagenomics, gene transfer  Tags to ID microbes, proteins, metabolites
<b>Milestone 2: Develop a Systems-Level Understanding of Microbial and Community Function and Regulation</b>  Facility for Whole Proteome Analysis  Facility for Analysis and Modeling of Cellular Systems	Photosynthesis, transporters, biomineralization  Proteins, machines, metabolites, and functional assays  Systems responses  Imaging	All GHGs: CO <sub>2</sub> , methane, nitrous oxide, dimethyl sulfide  Molecular inventories vs cues  Systems interactions with soil, rhizosphere, plants: Inputs and outputs (e.g., stable isotope probes)  Proteome and metabolome imaging at cellular and community levels	Community structure and relationship to function  Pathways and networks: Mechanisms of intercellular communication and function  Stoichiometry and kinetics of intercellular fluxes
<b>Milestone 3: Develop the Knowledgebase, Computational Methods, and Capabilities to Advance Understanding of Complex Biological Systems and Predict Their Behavior</b>  GTL integrated computational environment for biology	Modeling of climate-based and mitigational perturbations  Individual and multiple life-scale models (cellular, community, ecosystem): Metabolic budgets  Multiple photosynthetic processes	Modeling of microbial responses to manipulation of plant inputs into carbon cycle  Human inputs directed to soils  Response to environmental change understood	Four-dimensional reactive transport models based on genomic, geochemical, and hydrological data  Scaling of processes through molecular, cellular, community, and environmental levels; and molecular to long time scales
<b>Missions Outputs</b>  Measure environmental responses via sensors	Ecogenomics of sentinel organisms  Cellular, community, and ecosystem biochemical assays  Accompanying environmental assays	Biomarkers: RNAs, proteins, metabolites, signaling  Ecogenomics, functional assays, environmental conditions  Carbon and nutrient inventories	Biology and geochemistry: DNA, RNA, proteins, metabolites, geochemical from single-cell to field scales  Mesoscale simulation of field conditions  Regulatory levels of contaminants
<b>Robust Science Base for Policy and Engineering</b>	Natural behaviors of ocean ecosystems, impact on and of climate change scenarios incorporated into IA models  Assessment of efficacy and impacts of intervention strategies	Biological processes for carbon and nitrogen cycling, impact on and of climate change scenarios incorporated into IA models  Assessment of potential and strategy for terrestrial carbon sequestration	Predictions of transport and fate  Assessment of need for remediation  Remediation strategies, designs, and tests

## 5.1. Facility for Production and Characterization of Proteins and Molecular Tags

5.1.1. Scientific and Technological Rationale .....	112
5.1.1.1. Value of Proteins for Research.....	114
5.1.1.2. Value of Protein Characterization for Research .....	115
5.1.1.3. Value of Molecular Tags for Research.....	115
5.1.2. Facility Description .....	116
5.1.2.1. Facility Outputs .....	117
5.1.2.2. Laboratories, Instrumentation, and Support.....	117
5.1.3. Development of Methods for Protein Production.....	118
5.1.3.1. Production Targets .....	118
5.1.3.2. Specifications for Proteins and Comparisons of Their Production Methods.....	119
5.1.3.2.1. Comparison of Cell-Based Expression Systems.....	120
5.1.3.2.2. Cell-Free Systems.....	122
5.1.3.2.3. Chemical Synthesis .....	122
5.1.3.2.4. Protein Purification.....	122
5.1.4. Development of Methods for Protein Characterization.....	123
5.1.4.1. Requirements, Specifications for Functional Characterization Techniques, Data .....	125
5.1.5. Development of Approaches for Affinity-Reagent Production .....	126
5.1.5.1. Specifications for Affinity Reagents and Their Production.....	127
5.1.5.2. Technologies for Affinity-Reagent Production.....	130
5.1.6. Development of Data Management and Computation Capabilities.....	131
5.1.7. Facility Workflow Process .....	131



### Protein Production and Characterization

- ▶ Produce proteins encoded in the genome.
- ▶ Create affinity reagents that allow each protein to be identified, located, and manipulated in living cells.
- ▶ Perform biophysical and biochemical characterizations of proteins produced to gain insights into function.

## Facility for Production and Characterization of Proteins and Molecular Tags

The Facility for Production and Characterization of Proteins and Molecular Tags will be a user facility providing scientists with an understanding of the components encoded in the genome by using DNA sequence to make and characterize proteins and reagents for interrogating their functions in cells.

### 5.1.1. Scientific and Technological Rationale

Systems biology requires that we understand the proteins that make up a cell and the mechanisms of their function. Individual proteins encoded in the genome are the basic building blocks for biological functions potentially useful in DOE missions. Virtually every cellular chemical reaction and physical function necessary for sustaining life is controlled and mediated by proteins generally organized into macromolecular complexes or “molecular machines,” which might contain proteins, RNAs, or other biomolecules. A typical microbial genome has 2000 to 5000 genes that encode thousands of proteins and regulatory regions that control their expression. The challenge of understanding these workhorse molecules is technically complex and necessitates that very large numbers of them be produced and analyzed. Experimental analysis has determined the functions of only a few thousand of the millions of proteins encoded by the collective genomes on this planet—and even that understanding is incomplete.

#### Example of Mission Problem

### Proteins Provide Insight into Energy Production

Understanding the functions of bacteria, fungi, and algae is important for determining new ways to produce hydrogen or ethanol economically as a fuel. The genome sequences of these organisms provide a first step, but proteins carry out the useful functions encoded by the genes. To study proteins, they must be produced in quantities sufficient for analysis. In addition, studying these molecules functioning in their natural state (i.e., in the cell) requires the generation of affinity reagents or other molecular tags able to recognize specific proteins. Understanding how hydrogen-generating proteins function inside and outside cells will guide optimization of enzymatic hydrogen production for cell and cell-free applications.

# Protein Production and Characterization

We currently have insufficient data and conceptual insights to assign at least one function to about half the proteins found in even the most intensively studied microorganisms. Functional assignments for proteins in unculturable or less-studied organisms often occur by inference from a homologous protein's putative role in an intensively studied organism. A comprehensive understanding of cellular behavior will require experimental data for a significant portion of an organism's proteins (Roberts et al. 2004). We must have the ability to produce and characterize, as needed, essentially all the thousands of proteins encoded in many single genomes and in metagenomes to support functional gene annotation and, ultimately, mechanistic understanding. We also need to be able to produce and screen numerous variants of individual proteins or molecular machines so they can be used for DOE applications.

Having full-length and active forms of proteins in hand for biochemical and biophysical analysis will serve many purposes critical to the next generation of biology. These proteins provide an opportunity for discovery and a starting point for optimizing complex cellular processes from their components and molecular mechanisms. Providing rigorous and comprehensive characterizations for these proteins is invaluable to researchers and frees them to confidently pursue creative experimentation. "Molecular tags" or "affinity reagents" can be produced only by working from the proteins or via protein modification. These tags are critical for detection and potential quantitation of individual proteins and molecular machines in living systems.

The study of microbes, and especially those of DOE relevance, presents a special challenge. Microbial-community systems that we must understand possess millions of genes as opposed to the tens of thousands of even the most-complex higher organisms. The readily available genome sequences and even metagenome sequences of microbial communities have provided our first look into microbes' many functions. Most of the recently sequenced microbial genomes and metagenomes, however, show that roughly 40% of the genes are of unknown function, and, further, the microbes themselves either are not available or are "unculturable." Roughly 200 microbes have been sequenced to date, resulting in a catalogue of unknown genes that now contains 200,000 to 400,000 candidates for investigation. The ability to create and gain insight into proteins from genomic information alone is a crucial first step to understanding these microbial systems. Eventual culture-dependent experimentation on an important subset of microbes will be facilitated greatly by the availability of basic information on proteins and their respective affinity reagents.

Protein production currently is limited by economic and technological constraints and is a widely dispersed and inefficient "cottage industry." While substantial technology exists for generating the easy-to-produce (i.e., small, soluble) proteins, the ability to readily produce large multidomain proteins, membrane proteins, proteins with cofactors, and many other critical proteins is only emerging. For comprehensively understanding microbial systems, access to all proteins in metabolic, signaling, and regulatory pathways and networks is important. The most difficult proteins often are the very ones most vital to cellular function (e.g., those associated with essential transmembrane molecular machines, such as the photosystems in a photosynthetic microbe). In its mature state, the Protein Production and Characterization Facility will spend the greatest part of its effort on hard-to-produce, but critically important, proteins and will enlist the research community to help develop needed methods.

## Facility Objectives

- Perform comparative genomics against GTL Knowledgebase to determine gene function and to inform needed protein production and characterization
- Produce any protein on demand
- Characterize all proteins for quality assurance and quality control, for function, and for determining structure-function relationships as needed
- Produce affinity reagents and other molecular tags to enable location, tracking, and manipulation of proteins and machines in living systems
- Provide clones, proteins, affinity reagents, protocols, and data to scientists



## FACILITIES

A unique benefit of this facility is that, for the first time, a substantial suite of high-throughput, automated, and increasingly sophisticated characterization assays will be performed on proteins. Thus, protein production and characterization both will benefit as the transition is made from widely dispersed efforts focused on easy proteins to the economy of scale made possible by developing technologies capable of producing any desired protein with an accompanying database of reliable characterizations. The situation is somewhat analogous to genomic sequencing as it transitioned from dispersed, somewhat unreliable sequence data to higher-quality, lower-cost data at high-throughput, automated sequencing centers.

Automated high-throughput protein and affinity-reagent production will have several important impacts, including the following, that will enable the expeditious systemic study of chemical and physical interactions of proteins that underlie biology:

- A production environment will establish the necessary standards, diagnostics, control, and quality to develop and execute the demanding protocols for readily and repeatedly producing difficult proteins.
- A production facility will support a comprehensive and sophisticated array of characterization methods, most unavailable to the individual researcher, that can be applied to both production diagnostics and to protein characterization.
- Large-scale robotics, miniaturization, and automation will greatly enhance throughput and reduce costs.
- Making material and data products available to all scientists will leverage the investment to reach a larger community, whose work will facilitate further production, characterization, and understanding.
- Unlike the current situation, in which only selected portions of labor-intensive data are accessible, the facility's strong computational infrastructure will facilitate data mining of both successful and unsuccessful metadata associated with each protein.

### 5.1.1.1. Value of Proteins for Research

Ready and economic availability of proteins and affinity reagents will provide the foundation for the next generation of biological research, building on the national investment in genome sequencing. Having widespread access to cutting-edge technology in protein production will level the playing field, increasing the availability of proteins and protocols and creating a broader biotech industry (see sidebar, Protein Microarrays have Multiple Uses, p. 115). Proteins form the starting point for biochemical and biophysical functional studies, for eventual protein engineering, and for creating chimeric or new (optimized) biochemical pathways or even reactions or pathways that work in reverse directions (e.g., carbon dioxide to formate to methane). They offer the ability to study low-abundance proteins such as important regulatory proteins. Many variants (mutations) can be produced and studied for functional analysis. For nonculturable organisms, proteins can be produced from sequence alone to provide a shortcut to functional genome annotation and allow determination of quantitative biochemical binding or reaction constants. Comparative analyses of the structure and functions of protein families can be used to determine design principles. Proteins are reagents for studying metabolomics, post-translational modifications (substrate identifications), biosynthesis of metabolites and intermediates, binding-partner identification, and affinity-reagent generation. Functional proteins are the starting material for reconstituting molecular complexes, making quantitative and qualitative three-dimensional spectral and structural analyses, and mapping molecular interactions (with DNA, metabolites, and other proteins). They also can serve as mass and spectral standards for enhancement of mass spectrometry (MS) data analysis. Proteins, affinity reagents and other molecular tags, and data produced in the Protein Production and Characterization Facility are needed by users of other facilities to capture molecular machines for MS and other analyses and to identify the machines' components. They also are needed for cellular-imaging studies and verification of models (Roberts 2004; Roberts et al. 2004; see Table 1. Analysis of Technology Options for Protein Production, p. 120, and Table 2. Roadmap for Development of Technologies to Produce Proteins, p. 121).

## 5.1.1.2. Value of Protein Characterization for Research

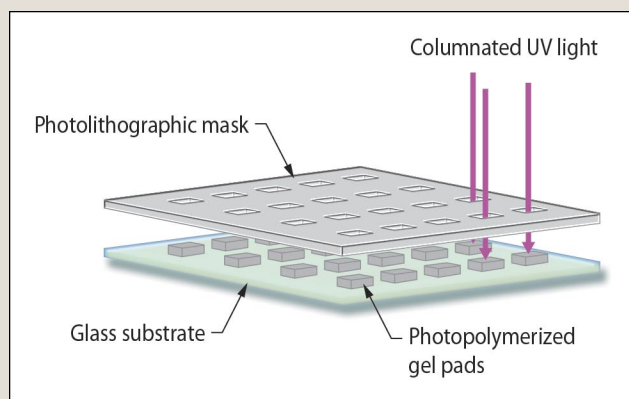
Automated high-throughput and high-quality biophysical and biochemical characterizations of proteins will provide a more rigorous assignment of gene function, resulting in first insights to a mechanistic understanding of microbial capabilities. For the first time, comprehensive reliable data on thousands of proteins will be available to analysts. The production facility can use high-throughput screening to characterize many proteins simultaneously under widely varied controlled conditions. Early analyses will focus on characterizations to determine basic biochemical function and biophysical information (e.g., solubility and insolubility in multiple solutions, multimeric state, presence of metals, and ordered and disordered domains). As the facility matures, the nature and sophistication of these characterizations will expand to determine more complex functions of individual proteins and molecular complexes (see 5.1.4. Development of Methods for Protein Characterization, p. 123, and Table 3. Summary of Characterization Needs and Methods, p. 124).

## 5.1.1.3. Value of Molecular Tags for Research

Two types of molecular tags are discussed here: Affinity reagents and fusion tags. Affinity reagents comprise proteins, peptides, nucleic acids, and small chemical molecules that bind targets of interest with high specificity and affinity. They commonly are used to detect where particular proteins are localized in cells, recover the protein and its associated molecules from cell lysates, and quantitate protein amounts in complex mixtures. Antibodies, popularly used as affinity reagents, can be generated by immunizing rodents or rabbits with the protein target and harvesting immunoglobulins (e.g., IgM and IgG) from the serum several months later. With the advent of various in vitro methods termed “display technologies,” antibody fragments (i.e., scFv, Fab, VH, VL,

## Protein Microarrays Have Multiple Uses

Proteins mass produced in the Protein Production and Characterization Facility or by the commercial sector from facility protocols may be delivered as microarrays to investigators in their labs. These devices provide a platform for directly studying global protein interactions and networks. Protein arrays also can serve as global “pulldown” and “affinity” purification platforms for spatially isolating molecular machines and complexes in the Molecular Machines Facility. Protein chips might serve as prepurification steps or as assays in the Proteomics Facility.



An example of a protein microarray is the “biochip” pictured above. This array supports 3D gel pads covalently attached to proteins, nucleic acids, antibodies, aptamers, and functional enzymes. The gel pads allow a solution-phase test environment that avoids subjecting biomolecules to the potentially harsh effects of a surface (e.g., glass slides). These arrays have been used for DNA and RNA analyses (including actual environmental samples), on-chip polymerase chain reaction and various ligation or amplification reactions, antibody arrays, functional protein assays, and protein-protein interaction studies. Combining these arrays (containing, for example, proteins, enzymes, aptamers, or antibodies) with time-of-flight mass spectrometry and automated spectral analysis allows characterization of the mass and identification of agents interacting with the elements embedded on the biochip. [Source: Argonne National Laboratory]

## References

1. A. Pemov et al., “DNA Analysis with Multiplex Microarray-Enhanced PCR,” *Nucl. Acids Res. Online* 33(2): e11 (2005). Retrieved from <http://nar.oxfordjournals.org/cgi/content/full/33/2/e11>.
2. I. M. Gavin et al., “Analysis of Protein Interaction and Function with a 3-Dimensional MALDI-MS Protein Array,” *BioTechniques*, 39(1), 99–107 (2005).

## FACILITIES

and VHH) can be isolated from naïve libraries in several weeks' time without the use of animals. In addition to modifying antibody-based molecules, scientists are altering other proteins (e.g., lipocalin, ankyrin, fibronectin domain, and thioredoxin) to bind to specific targets of interest. This is accomplished by modifying the open reading fragments through mutagenesis and selecting among the resulting library of randomized proteins for those that bind targets specifically. Finally, affinity reagents can be selected from libraries of combinatorial peptides, nucleic acids (i.e., aptamers), or small organic molecules (see 5.1.5. Development of Approaches for Affinity-Reagent Production, p. 126; Table 4. Analysis of Technology Options for Affinity Reagent Production, p. 127; Table 5. Roadmap for Development of Technologies to Produce Affinity Reagents, p. 128; and Table 6. Examples of Affinity Reagents and Their Applications, p. 128).

The following items focus on affinity reagents:

- Production of affinity reagents must be designed around their many applications. When proteins are in structured environments, some surfaces are exposed while others are hidden because they are in contact with other proteins or molecules. To deal with this contingency, multiple affinity reagents for each protein will ensure that any exposed surface or epitope can be accessed.
- Affinity reagents are needed that either disrupt or preserve protein activity. They can be used to manipulate proteins, including fabrication of biosensors; map post-translational modifications; determine spatial distributions; array targets in a unique spatial configuration; disrupt protein-protein interactions; promote crystallization of proteins; and stabilize membrane proteins.
- Affinity reagents can be used to assess biodiversity and in diagnostic tools for energy-production processes. They are critical for affinity purification of proteins and complexes, for identifying binding surfaces and mapping interactions in protein complexes, and for characterizing functional states (by targeting epitopes unique to active or inactive forms of the proteins). Finally, they are valuable in flow cytometry to sort cells from mixtures and for use in nanotechnology to anchor proteins during fabrication of novel biohybrid materials.

Another type of molecular tags—fusion tags—are short peptides, protein domains, or entire proteins that can be fused at the genetic level to proteins of interest. The target protein then is imparted with the fusion tag's biochemical properties. In general, the type of fusion tag used is dictated by its application. Short peptide tags (e.g., six-histidine, epitopes, StrepTag, calmodulin-binding peptide) regularly serve to permit facile purification of the recombinant protein, allow detection of the fusion protein, or direct the recombinant protein's interaction with other proteins or inert surfaces. Larger fusion partners such as protein domains (e.g., chitin-binding domain) or proteins (e.g., cutinase, GFP, GST, MBP, and intein) usually are employed to promote folding, solubility, purification, labeling, chemical ligation, or immobilization of the recombinant protein. If desired, the fusion tag can be detached from the protein of interest by cleaving a linker region with a site-specific protease that does not affect the protein (see Table 7. Examples of Fusion Tags and Their Applications, p. 129).

### 5.1.2. Facility Description

The facility will bring together comprehensive technologies for high-quality mass production and characterization of microbial proteins produced directly from sequence data or other genetic sources such as gene variants or clones. It also will be capable of generating specific capture and labeling affinity reagents for each protein. To derive insights into gene function and assess the best and most cost-effective protein-production strategies, a key capability will be computational comparison of genomic sequences of unknown organisms against the comprehensive GTL Knowledgebase. This user facility will integrate the basic research and technology development necessary to enable its continued scientific focus and usefulness in working with investigators and technologists in academia, national laboratories, and industry (see 5.1.7. Facility Workflow Process, p. 131, and accompanying sidebar with conceptual diagrams and narrative, p. 133).

## 5.1.2.1. Facility Outputs

Facility products will be distributed to research teams and accessible to the broader community of biologists. In general, proteins will have limited distribution because the facility will establish successful protocols and expression constructs that will allow researchers or commercial concerns to then produce proteins as needed for wider applications. Data and computational analyses will be available freely through the GTL computational environment. Products provided as needed to the user community include:

- Expression vectors (clones) for targeted genes
- Milligram quantities of purified, full-length, functional proteins
- Multiple affinity reagents for each protein, as well as chips with arrayed affinity reagents
- Proteins with a variety of fusion tags
- Initial biophysical and biochemical characterizations of each protein
- Production protocols so researchers and commercial concerns can readily produce proteins for research and biotechnology applications
- Comprehensive production and characterization databases and computational analyses referenced to the subject genome or classes of proteins

## 5.1.2.2. Laboratories, Instrumentation, and Support

The high-throughput facility's 125,000- to 175,000-sq.-ft. building will house core resources for protein production and characterization and the support necessary to ensure its mission. It will have extensive robotics for efficient sample production and processing and suites of highly integrated instruments for sample analysis and characterization of proteins and affinity reagents.

In the facility will be laboratories and instrumentation for production of large numbers of different DNA molecules, including cloning and insertion into expression vectors and, eventually, gene synthesis capabilities; production of proteins from any biological source; purification; quality assessment; and production of protein variants [e.g., isotopically labeled proteins, post-translationally modified proteins, proteins with novel cofactors, proteins incorporating nonstandard amino acids, and site-specific mutant arrays (high-throughput mutagenesis)]. The facility also will involve production of multiple affinity reagents for each protein; production of membrane proteins and multiprotein complexes; multimodal protein biophysical and biochemical characterization; and combinatorial capabilities to screen for complexes under multiple defined conditions. Methods will comprise cellular or cell-free expression and chemical synthesis. Onsite DNA sequencing will be required for several steps in the process. Informatics capabilities will track each gene or clone, protein, affinity reagent, and the associated data. Quality control will be assessed by onsite MS and a range of other biophysical and biochemical analyses.

Automation and computationally based insights are key to achieving high throughput at steadily declining costs, just as they were in DNA sequencing. Over time, as the GTL Knowledgebase matures (see 3.2.2.3.2. GTL Knowledgebase, p. 52), the GTL computational infrastructure will enable use of DNA sequence to predict the following for each protein: Efficient and successful production methods, likely binding partners, appropriate assay conditions, and, ultimately, information about the functions of each gene. Achieving this goal will require experience and the data created from production and characterization of tens of thousands of proteins.

Offices for staff, students, visitors, and administrative support will be included, as well as conference rooms and other common space. The facility will house all equipment necessary to support its mission. The DOE facility-acquisition process will include all R&D, design, and testing activities necessary to ensure a fully functional facility at the start of operations.



## 5.1.3. Development of Methods for Protein Production

Proteins have wide variability in their structure and stability—no single production method and characterization scheme will be applicable to every protein. Thus, several methods will be developed simultaneously, including all appropriate variations on cell-based, cell-free, and chemical synthesis.

Whichever method is selected, nearly all protein production is based on transcription from DNA obtained via cloning or possibly direct chemical synthesis of the gene encoding the desired protein. In cases where only gene sequence is available, chemical synthesis alone will be required. The Protein Production and Characterization Facility, as part of its function as a national resource, will develop a sequence-verified library of publicly available protein-coding microbial genes. This library would be available for translation into protein or for use in transformational studies by the other facilities or the larger scientific community.

Technologies should be scalable, economic, and sufficiently robust to work in a production environment. At least 50% of all proteins are anticipated to pose significant problems for any current method, so development work will be required. Some genes have evolved to generate only very small amounts of protein products. Most proteins are idiosyncratic with respect to conditions; for example, some proteins are not readily soluble or they are relatively unstable and require discovery of special conditions for storage, handling, and use. Others will function only in a properly reconstituted assembly and may need to be produced with their partners under specialized conditions. Consequently, a significant component of the facility will be research into new methods of protein production. In addition, many DOE-relevant systems may require techniques compatible with anaerobic or other extreme conditions. The strategy for success includes high-throughput parallel processing to allow exploration of a very large number of conditions and protocols specific to each protein.

Improved techniques are needed to predict from genome sequence the production and purification approaches most likely to succeed with each protein. We also need methods to identify all DNA sequences in a genome that should encode proteins. Thus, computation and informatics is an integral facility component. Algorithms based on data from successful and failed protein expressions are expected to improve future protein-production and -characterization efficiencies.

**Disorder and the Formation of Molecular Machines.** We need to produce these proteins in their functional state. Disorder is emerging as an increasingly important factor in protein function, particularly in the assembly of protein partners into molecular machines. This key process very often is mediated by disorder-to-order transitions at the binding interfaces as the disordered regions of two proteins become ordered by their interaction. Part of the facility's R&D effort will be to develop characterization methods that will, among other things, allow their general structure (whether ordered or disordered) to be defined and mapped. Whereas disordered protein regions are a hindrance in crystallization for classic protein crystallography techniques, our goal is to allow protein disorder to become a useful tool to predict binding partners and aspects of protein function (Dunker et al. 1998; Romero et al. 1998).

**LIMS.** A laboratory information management system (LIMS) will provide for machine learning from failures and successes of all facility aspects, the larger program, and other facilities. Experience-based decision making will allow selection of optimal expression, purification, storage, and characterization routes based on bioinformatics. Identification of domains that do and do not inhibit activity and strategies for affinity reagent production will be revealed. Inventory tracking and provenance records will be essential. Development will include better integration of instrument data files for generation of provenance records. For more information on LIMS and other computational and information technologies, see 4.0. Creating an Integrated Computational Environment for Biology, p. 81.

### 5.1.3.1. Production Targets

The initial numbers of proteins required are large by any current standard and certainly will increase over time with ongoing guidance and review from the researcher and user communities. In addition, each protein



# Protein Production and Characterization

probably will require exploration of a wide range of conditions to define successful production and characterization protocols. Several independent factors drive the need (see 2.0. Missions Overview, p. 21):

- Producing encoded proteins and characterizing them in a low-cost and high-throughput facility will make tractable and affordable the exploration of large numbers of unknown genes from sequenced microbes.
- Metagenomics is becoming more important as a methodology for studying natural systems critical to DOE mission environments. These studies are revealing millions of genes with the recurring 40% unknown ratio. Although more-sophisticated computational analyses can reduce the numbers that must be produced for analysis and for uncovering culturing techniques for some discovered microbes, potentially millions of proteins could or should be beneficially investigated through production.
- Understanding and eventually optimizing such critical microbial functions as redox processes, cellulose degradation, hydrogen production, and all the ancillary metabolic and regulatory pathways will entail screening potentially thousands of naturally occurring variants of hundreds of protein families. Exploring intentional modifications to understand function and to optimize properties could involve very large multiplicative factors on identified targets—gene shuffling can involve thousands of modifications.
- Exploring microbial function and incorporating nonnatural or isotopically labeled amino acids will be beneficial with or without various fusion tags (e.g., six-His, FAsH tag, and biotin).
- Engineering microbial systems or biobased cell-free systems for energy or environmental applications will require significant exploration of rationally engineered primary and ancillary proteins, machines, and pathways in a concerted and comprehensive way.
- Providing a source of proteins and their characterizations from gene sequence alone would produce a rapid and cost-effective alternative to historical culturing techniques and an important knowledgebase for possible culturing experiments.

Production targets will be determined by research needs and the level of maturity of the particular protein class. Production probably will proceed at multiple scales; the first exploratory pass to determine optimum successful production protocols should be at the smallest and most rapidly executable scale, followed by scaleup of interesting ones accordingly (see sidebar, Workflow Process, p. 133). Three examples follow.

- Screening mode: Microgram quantities, semipure,  $>10^4$  to  $10^5$  proteins/year
- Macroscale: Milligram quantities,  $>90\%$  pure,  $>10^4$ /year
- Large scale: Hundreds of milligram quantities,  $>95\%$  pure,  $>10^2$ /year

Material and data products must be accompanied by protocols that define optimal parameters for production, activity, storage, and use of proteins. The challenge in developing the Protein Production and Characterization Facility is to use various technologies in appropriate ways to cover production needs for all proteins, including small soluble proteins, membrane proteins, multiple domain proteins, and multiprotein complexes. Detailed comparisons of these available options will be a key part of the facility R&D and design process. Table 1, p. 120, provides a summary of technology options for protein production. Table 2, p. 121, is a simplified technology development roadmap covering the necessary research, pilot, and production phases of the R&D process. Each technology application has its own set of challenges. For the easy, soluble proteins, the challenge is scaleup, while the more difficult proteins and complexes require exploration of methods to produce and stabilize them. During facility operations, continued exploration of new techniques for protein production will be needed.

## 5.1.3.2. Specifications for Proteins and Comparisons of Their Production Methods

Methods eventually must be capable of cost-effectively producing on demand all the proteins coded in any microbial genome for which we have sequence, including the ability to coexpress proteins and purify or reconstitute protein complexes, difficult proteins such as membrane and multidomain proteins, metalloproteins, and proteins that cannot be overexpressed in host cells. Proteins must be properly folded and

# FACILITIES

active, incorporate correct cofactors and metals, and have correct post-translational modifications. Eventually, optimized versions of proteins should be available on demand, requiring screening of only dozens rather than hundreds or thousands of candidates. Three key methods for protein production and purification are described in sections 5.1.3.2.1–5.1.3.2.4 and in Tables 1 and 2 below.

## 5.1.3.2.1. Comparison of Cell-Based Expression Systems

Large-scale cell-based expression systems have been used worldwide in structural genomics centers and elsewhere, with *Escherichia coli* as the mainstay system. Yeast and other eukaryotic expression systems have

**Table 1. Analysis of Technology Options for Protein Production**

Comparative Analyses	Technology Options					Purification
	Cell-Based			Cell-Free	Chemical Synthesis	
	<i>E. coli</i>	Alternative Hosts	Homologous Hosts			
Strengths	Established methods, vectors  Renewable  Very cost-effective for industrial-scale quantities	Some higher success rates for certain proteins	Codon bias or missing cofactor issues eliminated	Scalable  Readily automated  Simplified cloning  HT screening under readily manipulated conditions  Cofactors  Labels  Production of toxic proteins	Scalable  Potential for automation  Labels and unusual amino acids incorporated during synthesis	Some tags demonstrated as high throughput, scalable  Numerous chromatography reagents available
Weaknesses	Scalability and high-throughput automation	Less developed methods, vectors  Cost  Not high throughput	Large efforts to develop methods, vectors, strains  Scalability and high-throughput automation	Currently only spontaneous disulfide bond formation	Ligations possible at only a small number of amino acid residues  Refolding required	Tag removal  Tag interference
Development Targets and Needs	More strains, vectors, procedures for difficult proteins	Improved vectors, strains, procedures for difficult proteins	Procedures generalized to engineer uncharacterized microbes	Automation demonstrated  Directed disulfide bond formation  Difficult proteins	Protein folding problem solved  Automated for high throughput	Capability to predict effects of tags  Microfluidics  Integration with characterization  Predictive capability for best purification and storage

June 14–16, 2004, GTL Technology Deep Dive Workshop, Working Group on Genome-Based Reagents

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

# Protein Production and Characterization

been developed for proteins that fail in *E. coli*-based systems. Their use is not as readily automated as with cell-free systems. Various alternatives are contrasted and compared in the three paragraphs below.

***E. coli*.** Use of *E. coli* for protein production is a robust technology (numerous vectors, strains, extant instrumentation infrastructure) that is relatively inexpensive. Bacterial cultures are a renewable resource (from small- to fermenter-sized cultures), and transformants can be stored indefinitely as DNA or frozen cells. Bacterial hosts can be engineered to coexpress certain proteins or chaperones. Shortcomings include scalability (the number of cultures and culture volume required); difficulty in predicting yields and solubility; product subjectability to proteolysis; costly labeling with certain isotopes; possible absence of necessary cofactors or chaperones; and necessarily large freezer storage capacity (and tracking) of transformants. Development needs include miniaturization of cultures for screening and production; improvements in methodologies and strains; and improvements for generating membrane and other difficult-to-produce proteins.

**Alternative Hosts.** Use of alternative hosts (yeast, *Pichia*, *Aspergillus*, insect cell lines) may permit better expression of particular proteins, but they have less-developed vector systems and strains and are more costly than bacterial and cell-free methods. In addition, they have slower growth rates compared to *E. coli*, codon-usage

**Table 2. Roadmap for Development of Technologies to Produce Proteins**

Objectives and Subtopics	Research	Pilots	Production	Products
<b>Protein Production</b> Small soluble proteins	Protocol refinement Optimization for cost-effectiveness	Scale up to 2 k/yr Protocol standards QA standards	Scale up to 25k/yr	Multiple forms of proteins Protein chips Protocols
<b>Protein Production</b> Membrane proteins	Detergents Refolding Novel expression systems Cell-free expression Chemical synthesis Domain identification Domain expression	Evaluate/validate expression systems Protocol standards QA standards	Automate Scale up	Multiple forms of proteins Protocols
<b>Protein Production</b> Multiple domain proteins Proteins with fusion tags	Refolding Novel expression systems Cell-free expression Chemical synthesis Domain identification Domain expression	Evaluate/validate expression systems Protocol standards QA standards	Automate Scale up	Multiple forms of proteins Individual protein domains Protein chips Protocols
<b>Protein Production</b> Multiprotein complexes (when needed for co-expression or stabilization and storage)	Binding-partner identification Refolding Novel expression systems Cell-free expression	Evaluate/validate coexpression systems Protocol standards QA standards	Automate Scale up	Multiple forms of proteins Protocols ID binding partners

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.

## FACILITIES

differences, and possibly missing cofactors or chaperones. These methods require investment in heterologous host systems and improvements for producing membrane and other difficult proteins.

**Homologous Hosts.** Use of homologous hosts has the advantage that cofactors, accessory proteins, modifying enzymes, and chaperones are present, and codons are optimized for open reading frames. These systems are less developed, however, with uncertain scalability, slow growth rates, low yields, nonexistent or difficult genetics and transformation, and the absence of selectable markers. Furthermore, they are not feasible for proteins from currently unculturable microbes. Development needs include defining optimal growth conditions, development of vectors and transformation protocols, and improvements for producing membrane and other difficult-to-produce proteins.

### 5.1.3.2.2. Cell-Free Systems

Cell-free expression systems, such as those based on wheat germ or *E. coli* extracts, hold the greatest potential for full automation and hence lower costs and higher throughput. Successful efforts in Japan using these extracts have yielded hundreds to thousands of proteins per year (Kigawa et al. 1999; Sawasaki et al. 2002; Kawasaki et al. 2003; Endo and Sawasaki 2003). Having the ability to automate these systems and the potential to incorporate labeled or nonstandard amino acids adds to their value. However, these methods have not yet seen widespread use or application. A broader experience base needs to be established.

**Cell-Free Methods.** Amenable to robotics (and microtiter plates), cell-free methods can have either small sample-reaction volumes (30- $\mu$ L reaction volumes, 30- $\mu$ g yields) or large. Cell-free proteins can be produced from PCR-amplified DNA templates, eliminating extensive cloning steps and simplifying rapid testing of many construct variations, thereby making this an attractive method for high-throughput screening. Produced protein molecules exist in simpler mixtures, sometimes permitting functional assessment without purification. Multiple proteins can be coexpressed to assemble complexes. Cofactors and detergents can be added, and certain isotopes can be cost-effectively incorporated. Shortcomings include relatively expensive application, although this is expected to decrease substantially as the method becomes more widely used. Disulfide bonds must form spontaneously when reducing agents are removed. Development needs include advances in directed disulfide bond formation, replacement of cell lysates with recombinant proteins and ribosomes, and improvements in generating membrane and difficult-to-produce proteins.

### 5.1.3.2.3. Chemical Synthesis

Solid-state chemical synthesis is a possible approach for important proteins that fail in all DNA-based expression systems. Currently, this method can produce peptides up to 50 amino acids in length, but longer peptides are made at ever-diminishing efficiencies. Full-length proteins might be synthesized through chemical ligation of multiple peptides. This currently is a costly procedure, and refolding into active protein remains a major problem. This technique has the advantage of producing milligrams of proteins labeled by incorporation of isotopes, chemical modifications, unnatural amino acids, or other chemical groups.

**Chemical Synthesis Methods.** Requiring no DNA, chemical synthesis can have large yields (>50 mg) for small proteins. There is no contamination by cellular proteins, and incorporating unnatural amino acids, labels, and post-translational modifications is easy. Chemical synthesis currently is not high throughput, and it is labor intensive. It is limited to proteins shorter than 200 amino acids, and the product typically requires refolding. Development needs include cheaper production of thousands of peptides, expansion of peptide ligation sites, reliable refolding, and improvements for generating membrane and difficult-to-produce proteins.

### 5.1.3.2.4. Protein Purification

Protein purification after expression presents a number of challenges, particularly in a high-throughput environment. In the Protein Production and Characterization Facility, substantial reliance will be placed on experience-based informatics methods to guide the purification strategy for each protein, with the expectation of achieving significant improvement as the database expands. Automated protocols aimed at

# Protein Production and Characterization

eliminating centrifugation will be developed since this step accounts for the major bottleneck in current protein-production protocols.

**Purification Methods.** Methods based on affinity-purification tags permit generic protocols for purification, but tags can interfere with structure or function and tag removal may be required. Current methods are not high throughput, contaminants may be hard to eliminate, and activity may be lost during purification (i.e., loss of cofactors, denaturation). Development needs include improved instrumentation for high throughput, and the special problems of purifying and storing native membrane proteins should be addressed.

## 5.1.4. Development of Methods for Protein Characterization

Key and largely unique goals of the Protein Production and Characterization Facility are stabilization and extensive characterization of each produced protein under well-defined conditions, with the resulting data made easily accessible to internal and external users. Given the investment in each expressed protein and its scientific value, investigators plan to subject each to a substantial suite of assays. Measurements for thousands of proteins will be generated robotically under standardized conditions, producing voluminous data. Assays must be rapid and inexpensive, requiring miniscule protein quantities to allow data collection from a broad range of conditions. Technologies such as microfluidics and other lab-on-a-chip methods eventually will provide the required versatility and sensitivity, with attendant sample economies and speed (see sidebar, Micro- and Nanoscale Methods, this page). Some of these protocols should reveal additional functional, structural, biological, chemical, and physical insights.

Serving several purposes, characterization first supports production by validating that the right protein has been produced (without sequence or translation errors), that the protein is stable and nominally folded, and that conditions necessary for long-term stabilization and storage have been met. Subsets of these measurements will be made on all protein attempts, including those to generate only screening levels of unpurified proteins. Since no single measurement provides all the answers, suites of techniques will be employed as they are feasible and required (see Table 3, p. 124).

Once we are assured that validated and stable proteins are produced, a more complete set of biophysical and biochemical characterizations will be made as required by the particular research problem and system. According to program and facility governance, user groups and the review process will adjudicate resource allocation with cost and benefit analyses of each characterization. The more complete characterizations likely will be on a down-selected group—10 to 20% of total protein inventory. These measurements will delve more deeply into structure and function. Not all measurements necessarily will be made in this facility but possibly at other facilities or in researchers' laboratories. Various parameters that might be measured are listed below.

### Micro- and Nanoscale Methods Reduce Costs and Improve Performance of High-Speed and High-Throughput Production and Analysis

Recent advances in microanalytical systems support the downscaling of many standard methods, resulting in improved performance and facilitating easier integration of multiple techniques, automation, and parallel material processing. Microfluidic technologies have been used to miniaturize such conventional technologies as chromatographic separations, protein and DNA electrophoresis, cell sorting, and affinity assays (e.g., immunoassays). These methods typically are 10 to 100 times faster (allowing analysis of unstable biological molecules), use 1/100th to 1/1000th the amount of sample and reagents (drastically lowering costs), and offer 2 to 10 times better separation resolution and efficiency than their conventional counterparts. Moreover, the ability to analyze minute amounts of sample reduces sample loss and dilution and allows characterization of low-abundance molecules or screening for exploratory protein-production methods. Microscale miniaturization also enables integration and parallelization of different biochemical processes and components and will be important for all production and analytical processes in the GTL facilities.



**Table 3. Summary of Characterization Needs and Methods**

Properties of Proteins and Affinity Reagents	Analytical Technologies (Computationally Informed)
<b>Product Validation, QA/QC</b>	
<b>Protein production, identification</b> Post-translational modifications Sequence of polymorphisms, isoforms Cloning artifacts Required cofactors, ligands, binding partners (combinatorial approaches) Stability (cofactors, ligands, binding partners) Folding (cofactors, ligands, binding partners) Storage and handling conditions	Mass spectrometry, affinity tag reaction (e.g., arrays, microfluidics, gels), light scattering, spectral matching (IR, UV), 1D/2D gels, liquid chromatography (e.g., affinity, ion exchange) Centrifugation, light scattering, spectroscopy methods (UV, CD) Screening level (UV-CD, dye binding, partial proteolysis/MS, isotope exchange/MS, FT-IR, SAXS/SANS, WAXS, EM) Robotic HT combinatorial methods (e.g., pH, temperature, salts, buffers, solvents), test with stability diagnostics
<b>Biophysical and Biochemical Characterization</b>	
<b>Prepurification</b> (See items below under postpurification) <b>Postpurification</b> Binding partners; identification of reconstitution conditions, intermolecular interactions (dissociation constants) Identification of monomeric or multimeric state Probe of folding landscape, identification of motifs, folding stability, thermodynamics, ordered and disordered regions Discovering substrates (orphan enzymes) Identification of cofactors (e.g., metals, NADH, ATP, ligands) Biological effect of post-translational modifications Identification of DNA and RNA binding, sequence motifs Assignment of function to proteins	HT screening: Dye binding, internal fluorescent labels, metabolite and molecular cocktails/MS (i.e., agonists and antagonists), affinity arrays, MS, biochemical and binding assays, ATP binding, kinase activity, affinity reagent effect on protein activity (neutral or inhibitory) HT, high fidelity: Dye binding, internal fluorescent labels, metabolite and molecular cocktails/MS (i.e., agonists and antagonists), affinity arrays, MS, biochemical and binding assays, ATP binding, kinase activity HT, high fidelity: UV-CD, dye binding, partial proteolysis/MS, isotope exchange/MS, FT-IR, fluorescence emission/lifetime (FIE/L), FRET, SAXS/SANS, WAXS, EM, calorimetry, size-exclusion chromatography coupled with laser light scattering (SEC-LLS) Affinity reagent on protein activity (neutral or inhibitory)
<b>Ultimate Characterization</b>	
Protein primary, secondary, tertiary, and quaternary structures Structural-activity relations Assignment of functions	Computational modeling and simulation Analyses from GTL facilities HT structural measurements: X-ray crystallography, NMR, cryoEM, scanning probe microscopy, FRET, single-molecule spectroscopies
<b>Ultimate Manipulation</b>	
Design of affinity reagents Protein and molecular machine redesign or refinement Pathway redesign Engineering into nanomaterials and devices	Computational modeling and simulation Analyses from GTL facilities Functionalization of nanomaterials, synthetic biology, directed evolution Microbial and cell-free systems design and engineering

# Protein Production and Characterization

- Screen, identify, and measure enzymatic or binding activity, cofactor state and requirements, effect of affinity reagents on proteins (e.g., epitopes, inhibitory or noninhibitory for selected activities)
- Identify agonists and antagonists
- Identify binding partners and determine affinities (dissociation constants) under a suite of conditions, including salts, buffers, pH, temperature, and aerobic or anaerobic
- Identify monomeric or multimeric state
- Identify reconstitution conditions, intermolecular interactions
- Probe the folding landscape, establish structure
- Identify motifs, folding stability, thermodynamics, ordered and disordered regions
- Discover substrates (orphan enzymes)
- Identify cofactors (metals, NADH, ATP, ligands)
- Elucidate biological effect of post-translational modifications
- Identify DNA/RNA binding and sequence motifs

Specific biochemical functions and sensitivities pertinent to DOE applications (e.g., metal reduction, proton or electron transfer, carbon reduction) will be critical. Many of these measurements can be made before the proteins have been purified and thus done in screening mode during the production process. Some measurements could be done with proteins produced to contain sensitive fluorescent probes designed to facilitate inexpensive, high-throughput characterizations with miniscule quantities of protein.

For a set of proteins selected for their unique and mission-relevant properties (e.g., hydrogen and biofuel production, carbon cycling, contaminant immobilization, sensors), the ultimate characterization suite will determine structure at the highest-possible resolution (primary, secondary, tertiary, and quaternary). This approach will use state-of-the-art national synchrotron, neutron, NMR, and electron microscopy facilities and lab-based molecular techniques. These measurements will allow the establishment of structural-activity relations and the understanding of design principles. Computation will be a key part of such analyses.

One of the facility's ultimate roles is to support the refinement and redesign of proteins and affinity reagents for a diverse suite of energy and environmental applications. It will produce and characterize the effects of a wide range of modifications to understand design principles and optimize performance. This includes design of affinity reagents spanning several approaches, not all of which may be proteins or even cellular; protein and molecular-machine redesign or refinement; pathway redesign; and the engineering of biofunctional materials into nanomaterials and devices for energy and environmental applications and research.

As the facility matures, characterizations will shift emphasis from supporting production methods to more advanced characterizations that provide finer detail on structure and function and elucidate design principles.

## 5.1.4.1. Requirements, Specifications for Functional Characterization Techniques, Data

Methods should be sensitive enough to work with screening-mode levels of proteins where possible and should include cost-effective and high-throughput biochemical and biophysical measurements. Individual measurements should be very inexpensive so they can be repeated under a variety of conditions to reflect salt, pH, buffer concentration, cofactors, ligands, and temperature. They also should have a low coefficient of variation to permit statistical analysis. They should be highly parallelized and scalable and provide QA/QC with feedback to the production process. Computational support will include algorithms for cherry-picking samples for retesting and optimizing activity conditions.

Much of the needed instrumentation is laboratory based (i.e., it can be located within the Protein Production and Characterization Facility). Some measurements could benefit from remote instruments like a high-brightness synchrotron or neutron source. For example, at such a synchrotron facility, high-throughput

## FACILITIES

systems (flow or robotic enabled) could be developed and evaluated as a means to provide a cost-effective platform for making certain types of valuable measurements on protein samples [e.g., small-angle X-ray scattering (SAXS) or extended range circular dichroism (or UV-CD)]. Results of such developments could be evaluated for their usefulness in the context of this facility's production goals. To take advantage of such an approach, methods would need to be developed for transporting and automating sample handling, data logging and processing, and comparison of results obtained by these methods. Results would need to be integrated with other laboratory-based measurements.

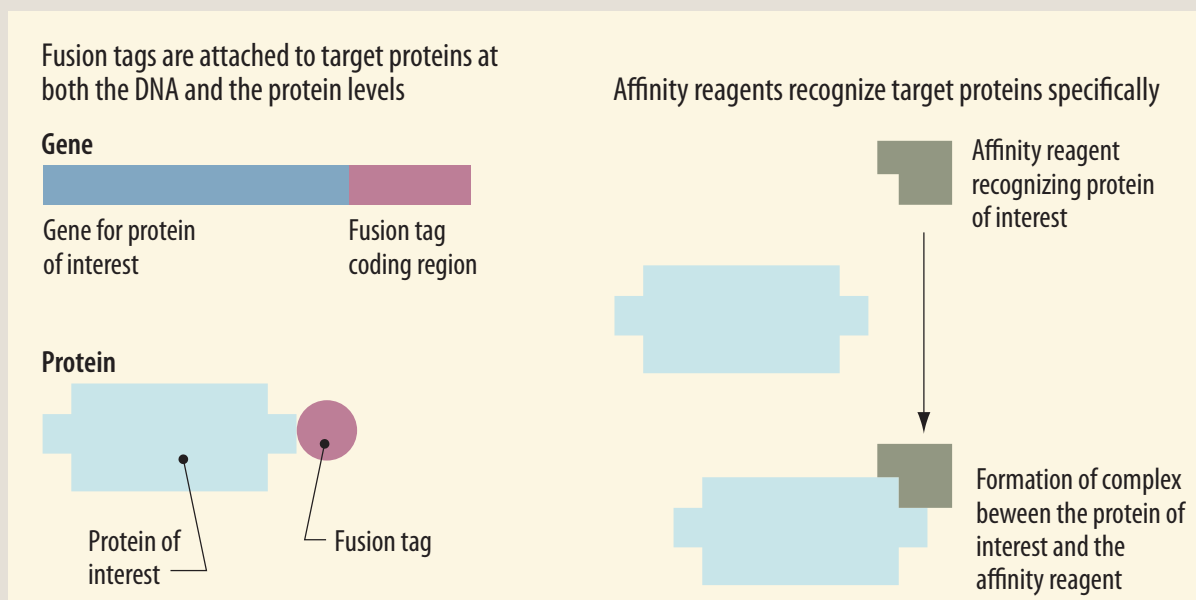
### 5.1.5. Development of Approaches for Affinity-Reagent Production

Production of multiple high-affinity, high-specificity affinity reagents and suitable fusion tags for each protein presents enormous challenges (see sidebar, Molecular Tags: Fusion Tags and Affinity Reagents, this page). Several promising approaches are under development worldwide, although none has yet emerged as an economical and reliable solution to GTL's high-throughput needs. Overcoming this obstacle is therefore a major target for GTL pilot studies and for this facility in particular (see Table 4, p. 127; Table 5, p. 128; Table 6, p. 128; and Table 7, p. 129).

High-throughput systems must be capable of producing numerous affinity reagents that recognize different domains of each protein. This will require multiple new libraries of affinity reagents from which members with desired affinity and specificity to each target protein can be selected. Different and complementary approaches are under development, including phage and yeast display systems and aptamers. When full proteins cannot be produced, these tags might be created for appropriate epitopes that can be determined by computational analyses. In addition, computational insights eventually might recommend the best affinity-reagent approach for particular proteins. These techniques will require substantial development.

#### Molecular Tags: Fusion Tags and Affinity Reagents

Fusion tags (orchid) are short peptides, protein domains, and entire proteins that are fused at the genetic level so the cell's endogenously produced proteins of interest (light blue) will have the imparted fusion tag's biochemical properties. Affinity reagents (green) are proteins, peptides, nucleic acids, and small chemical molecules that bind targets of interest with high specificity and affinity. There are many possible affinity reagents for each protein.



# Protein Production and Characterization

Further developmental areas include improved reagent stability and specificity; improved multiplex screening protocols; and rapid, high-throughput affinity-maturation techniques. Reagents also will be evaluated to determine where they bind to their protein targets and whether they disrupt the target's function, thereby dictating how different affinity reagents can be used. Development of modular affinity reagents also would be extremely useful; selected binding domains could be generated rapidly for such different purposes as protein isolation or live-cell imaging.

In many cases, the most useful affinity reagents may be proteins themselves. They can be produced and characterized using technologies already developed for bacterial proteins. They will be standardized reagents, however, so processes can be developed to allow for their rapid and large-scale production, enabling their distribution to scientists worldwide and greatly enhancing the scientific impact of reagents generated in the facility.

## 5.1.5.1. Specifications for Affinity Reagents and Their Production

Affinity reagent production technologies must be rapid, cost-effective, and amenable to high-throughput automation; they should be capable of being based on antibody fragments, engineered protein scaffolds, combinatorial peptides, and aptamers as the need dictates. They should work with targets that have reduced cysteines or are cell toxic. A computationally based decision process is needed for selecting proteins or epitopes of proteins to serve as targets for affinity-reagent generation. Affinity reagents should bind either individual proteins or complexes, and the collection should recognize three to five different epitopes on a protein and be amenable to epitope subtraction and existing target-detection strategies. The process should identify reagents best suited for particular applications (i.e., Western blot, pulldown, coimmunoprecipitation, staining, complex disruption, inhibited catalytic activity, and inhibited protein-protein interactions).

**Table 4. Analysis of Technology Options for Affinity Reagent Production**

	Phage Display	Yeast Display	Ribosome and Puromycin Display	DNA or RNA Aptamers	Animals
<b>Strengths</b>	Good diversity Fusion proteins	Liquid and fluorescence-based screening Affinity maturation Fusion proteins	Good diversity Fusion proteins	Good diversity	Many secondary antibodies available
<b>Weaknesses</b>	Slower screening Plate based	Fluorescent tags required that may complicate recognition Reduced cysteine on targets problematic	Slower screening	Fewer secondary affinity labels Not protein based, so no fusion proteins	Expensive Not high throughput Nonrenewable unless use mAb Slow
<b>Development Targets and Needs</b>	High throughput demonstrated Improved screening	High throughput Improved screening Secondary antibodies that must be developed	Optimization of scaffolds, screening methods, and automation	Optimization of screening methods	Optimization of screen methodologies DNA immunization and improvements in hybridoma production

June 14–16, 2004, GTL Technology Deep Dive Workshop, Working Group on Genome-Based Reagents

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

# FACILITIES

**Table 5. Roadmap for Development of Technologies to Produce Affinity Reagents**

Objectives Subtopics	Research	Pilots	Production	Products
<b>Affinity-Reagent Library Development</b>	Useful molecular scaffolds developed Useful libraries constructed and evaluated Design validated Expression tested System compatibility tested	Automate library Protocol standards QA, standards	Scale up	Affinity reagents Reagent chips Protocols QA, standards
<b>Affinity-Screen Automation</b>	Develop protocols	Scale up to 2k/year Protocol standards QA, standards	Scale up to 25k/year	Affinity reagents Reagent chips Protocols QA, standards
<b>Affinity-Reagent Target Design</b>	Novel vectors Validate designs	Integrate into protein production system Protocol standards QA, standards	Scale up	Immobilized targets Protein chips Protocols QA, standards

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.

**Table 6. Examples of Affinity Reagents and Their Applications**

Examples	Applications
<b>Obtained by Animal Immunization</b>	
IgG and IgM	Detection, purification
<b>Obtained by in Vitro Methods (Affinity reagents based on antibody-like proteins)</b>	
Fab	Detection, purification, therapeutics
FV	Detection, purification, crystallization
scFV	Detection, purification, in vivo perturbation, therapeutics
Domain antibodies (VH, VL)	Detection, purification, therapeutics
VHH (shark and camel heavy-chain antibody VH domains)	Detection, purification
Fibronectin type 3 domain	Detection, purification, in vivo perturbation
<b>Affinity Reagents Based on Other Proteins (Scaffolds)</b>	
Affibody (protein A)	Detection, purification
Anticalin (lipocalin)	Detection, purification
Ankyrin repeats	Detection, purification, in vivo perturbation
Thioredoxin	In vivo perturbation
<b>Affinity Reagents Based on Other Molecules</b>	
Combinatorial peptides	Detection, crystallization, in vivo perturbation
RNA or DNA aptamers	Detect, purification, in vivo perturbation
Small chemical molecules	In vivo perturbation



# Protein Production and Characterization

**Table 7. Examples of Fusion Tags and Their Applications**

Examples	Applications
<b>Peptide Tags</b>	
Six histidine	Purification by immobilized metal affinity chromatography (IMAC)
Epitope (e.g., myc, V5, FLAG, soft-epitope)	Detection with antibodies, purification, immunoprecipitation
StrepTag	Purification with streptavidin
S tag	Purification, detection
AviTag, Pinpoint	In vitro or in vivo biotinylation
Tandem affinity (TAP)	Purification
Tetracysteine	In vivo labeling, purification
Lanthanide-binding peptide	Labeling
Coiled-coil	Heterodimerization with partner peptide (e.g., E coil with K coil)
Metal, semiconductor, or plastic binding peptides	Immobilization on surfaces, nucleation or growth of nanocrystals, detection of semiconductor materials
Calmodulin-binding peptide	Purification (Ca <sup>2+</sup> dependent)
Elastin-like peptides	Purification (temperature-dependent aggregation)
<b>Protein Tags</b>	
Fusion partners (glutathione-S-transferase, maltose binding protein, cellulose-binding domain, thioredoxin, NusA, mistin)	Promotion of folding, solubility, expression, or purification of fused protein
Chitin-binding domain	Promotion of folding, solubility, expression, purification, immobilization
Green fluorescent protein or alkaline phosphatase	Monitoring of expression, purification, or binding of fusion partner
Cutinase, O <sup>6</sup> -alkylguanine alkyltransferase (AGT), or halo tag	Covalent modification for immobilization, purification, or detection
Intein	Chemical ligation in vitro or in vivo

Affinity reagents should bind their target with modest to high affinity, have lowest-possible failure rate (cross-reactivity, low affinity), be obtainable in reasonable amounts (5 mg, >90% pure) in a cost-effective manner, and be stable and storable. They should be formattable on chips with excellent shelf life and available in fluorescent, biotinylated, or enzyme-linked forms; and formattable for affinity chromatographic methods to purify individual proteins or protein complexes from cells. Ideally, they should be expressible inside cells where they can bind their target and be made conditional or regulatable.

Just as for proteins, no single method will work equally well for producing all affinity reagents, so several methods will be needed. Operationally, methods must be capable of generating reagents from small target amounts (tens of micrograms). They must readily screen diverse libraries with targets and select out the best binders applicable under a variety of conditions; have the capability to screen libraries of more than 10<sup>9</sup> members in a rapid manner for hundreds of targets per day; validate binding to specific target protein; and be amenable to affinity maturation.

Material and data products must be accompanied by protocols that define optimal parameters for production, activity, storage, and use. The challenge is to use various technologies in appropriate ways, including phage display, yeast display, ribosome and puromycin display, DNA or RNA aptamers, and immunization of animals. Table 4, p. 127, provides a summary of technology options for production of affinity reagents.

Table 5, p. 128, is a simplified technology development roadmap covering the necessary research, pilot, and production phases of the R&D process. Each technology application has its own set of challenges. During facility operations, continued exploration of new techniques will be needed.

## 5.1.5.2. Technologies for Affinity-Reagent Production

**Phage Display.** This technology can use libraries of combinatorial peptides, antibody fragments, and engineered protein scaffolds. Phage display is amenable to high-throughput screening with robotics; it is protein based, so functionality is added easily by creating fusion proteins with different functional domains; and it has been used for in vivo and subtractive selections. The resulting output, however, may have to go through a second round of evolution as it tends to isolate weak and strong binders at the same time. In addition, candidates should be sorted according to differences in affinity, specificity, epitope overlap, stability, storage, and application, and the output may be misleading about the strength of binding due to multivalent display. The technology may require different scaffolds, depending on the application. Development needs include the optimization of scaffolds and screening methodologies.

**Yeast Display.** Capable of using libraries of combinatorial peptides, antibody fragments, and engineered protein scaffolds, the yeast display technology can discriminate affinities by flow cytometry, permitting fast assessment and identifying downstream candidates. Good for directed-evolution experiments (enhanced affinity, specificity, expression, or stability) and for epitope identification, yeast display is protein based, so functionality can be added easily by creating fusion proteins with different functional domains. It may need to go through a second round of evolution, however, and its libraries tend to be less diverse than other display formats. Candidates may require sorting by affinity, specificity, epitope overlap, stability, storage, and application. Yeast grow slower than phage, taking more time and effort and needing larger volumes per screening cycle, so making this technology high throughput is more difficult. Yeast display requires different scaffolds, depending on the application. Development needs include optimization of scaffolds and screening methodologies.

**Ribosome and Puromycin Display.** These methods can work with very large libraries (i.e.,  $10^{12}$  members); monovalent display leads to selection of the best binders. The ribosome- and puromycin-display technologies can incorporate mutagenesis during screening and enhance binding during the general selection process. They are protein based, so functionality can be added easily by creating fusion proteins with different functional domains. They are more expensive than phage- and yeast-display technologies, however, and large libraries require more rounds of screening. Candidates need to be sorted by affinity, specificity, epitope overlap, stability, storage, and application; they require different scaffolds, depending on the application. Development needs include optimization of scaffolds and screening methodologies and automation.

**DNA or RNA Aptamers.** Use of DNA or RNA aptamers is amenable to very large libraries (i.e.,  $10^{12}$  members) and high-throughput screening with robotics. Synthesizing large amounts of individual aptamers is relatively expensive, however, and large libraries require more rounds of screening than phage or yeast libraries. Aptamer candidates should be sorted by affinity, specificity, epitope overlap, and application, and they are limited to DNA/RNA. Development needs include optimization of screening methodologies.

**Immunization of Animals.** This traditional, well-established approach requires animals and large amounts of antigen. Repeated injections are necessary, so it is slow. This is a nonrenewable resource unless hybridomas are generated, so the method is expensive; it is limited by the immune response because common epitopes cannot be subtracted. Development needs include DNA immunization and improvements in hybridoma production (see Table 4, p. 127, for strengths and weaknesses and development roadmap).

## 5.1.6. Development of Data Management and Computation Capabilities

Each step and process in the Protein Production and Characterization Facility will involve very large numbers of biological samples that need to be tracked appropriately through the automated systems. Sophisticated bioinformatics analysis will be greatly needed at all steps so insights can be gained from both successes and failures. Processes will generate vast amounts of valuable data on clones and proteins and their characterization. These and other data will be captured properly and disseminated to the scientific user community. Implementation of appropriate LIMS and data-mining capabilities will be absolutely crucial to achieving high-throughput, cost-effective clone and protein production as well as to enable the use of these materials in contributing to the goals of GTL and the Department of Energy. These criteria will require large computing resources and development of the best scientific tools to properly mine the invaluable data being produced. For more details, see Table 8. Computing Roadmap, p. 132.

## 5.1.7. Facility Workflow Process

Conceptual diagrams, shown in the insert starting on page 133, depict prospective major facility equipment layout, process flow, and production targets. The process begins with genomics, which includes comparative genomic analyses against the GTL Knowledgebase to (1) gain insight into an unknown genome and identify its protein production targets and (2) produce clones or synthesized genes. Protein production first is pursued in a high-throughput, low-volume screening mode using appropriate microtechnologies, followed by full-scale production with successful protocols and robotics. Characterization is carried out for QA/QC, for initial biophysical and biochemical analyses, and for in-depth studies as needed. With applicable technologies, affinity reagents to selected proteins are produced using pipelines very similar to those for protein production. Computing and information technologies will support and inform all phases of facility processes and provide protocols, supporting data, and characterizations to the scientific community. The facility will have data and sample archives and distribution capabilities.

**Table 8. Computing Roadmap: Facility for Production and Characterization of Proteins and Molecular Tags**

Topic	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment
<b>LIMS and Workflow Management</b>  Participate in GTL cross-facility LIMS working group	Available LIMS technologies Process description for LIMS system Crosscutting research into global workflow management systems Expert system approaches to guiding production protocols	Prototype production design strategy system Prototype protein production LIMS system Prototype biochemical characterization LIMS system Workflow management system for production and characterization Process simulation for facility workflow	Production design Protein production LIMS system Biochemical characterization LIMS system
<b>Data Capture and Archiving</b>  Participate in GTL cross-facility working group for data representation and standards	Data models for process metadata and biophysical characterization data Technologies for large-scale storage and retrieval Preliminary designs for databases	Prototype storage archives Prototype user-access environments	Archives for key large-scale data types (e.g., biophysical characterization data) Archives linked to this facility's community databases and other GTL data resources Archives for microbial genome annotation with partners
<b>Data Analysis and Reduction</b>  Participate in GTL cross-facility working group for data analysis and reduction	Algorithmic methods for biophysical characterization modalities Grid and high-performance algorithm codes Design for biophysical characterization tools library	Prototype biophysical characterization tools library Prototype analysis grid for biophysical characterization, with partners Analysis tools linked to data archives	Large-scale annotation systems with partners Production-analysis pipeline for biophysical characterization on grid and high-performance platforms Library with production-analysis codes Analysis tools pipeline linked to end-user problem-solving environments
<b>Modeling and Simulation</b>  Participate in GTL cross-facility working group for modeling and simulation	Existing technologies explored for protein-fold prediction Technologies explored for low-resolution modeling from scattering data	Genome-scale protein-fold prediction, with partners Prototype code for protein modeling from scattering data	Production pipeline and end-user interfaces for genome-scale fold prediction Production codes for scattering-data modeling
<b>Community Data Resource</b>  Participate in GTL cross-facility working group for serving community data	Data-modeling representations and design for databases: protein and reagent catalog, protein biophysical characterization, protein-production methods, and protocols	Prototype database End-user query and visualization environments Databases integrated with other GTL resources and databases	Production databases and mature end-user environments Integration with other GTL data resources Integration with other community protein-data resources
<b>Computing Infrastructure</b>  Participate in GTL cross-cutting working group for computing infrastructure	Analysis, storage, and networking requirements for protein production facility Grid and high-performance approaches for large-scale data analysis for biophysical characterizations and establish requirements	Hardware solutions for large-scale archival storage Networking requirements for large-scale grid-based biophysical data analysis	Production-scale computational analysis systems Web server network for data archives and workflow systems Servers for community data archive databases

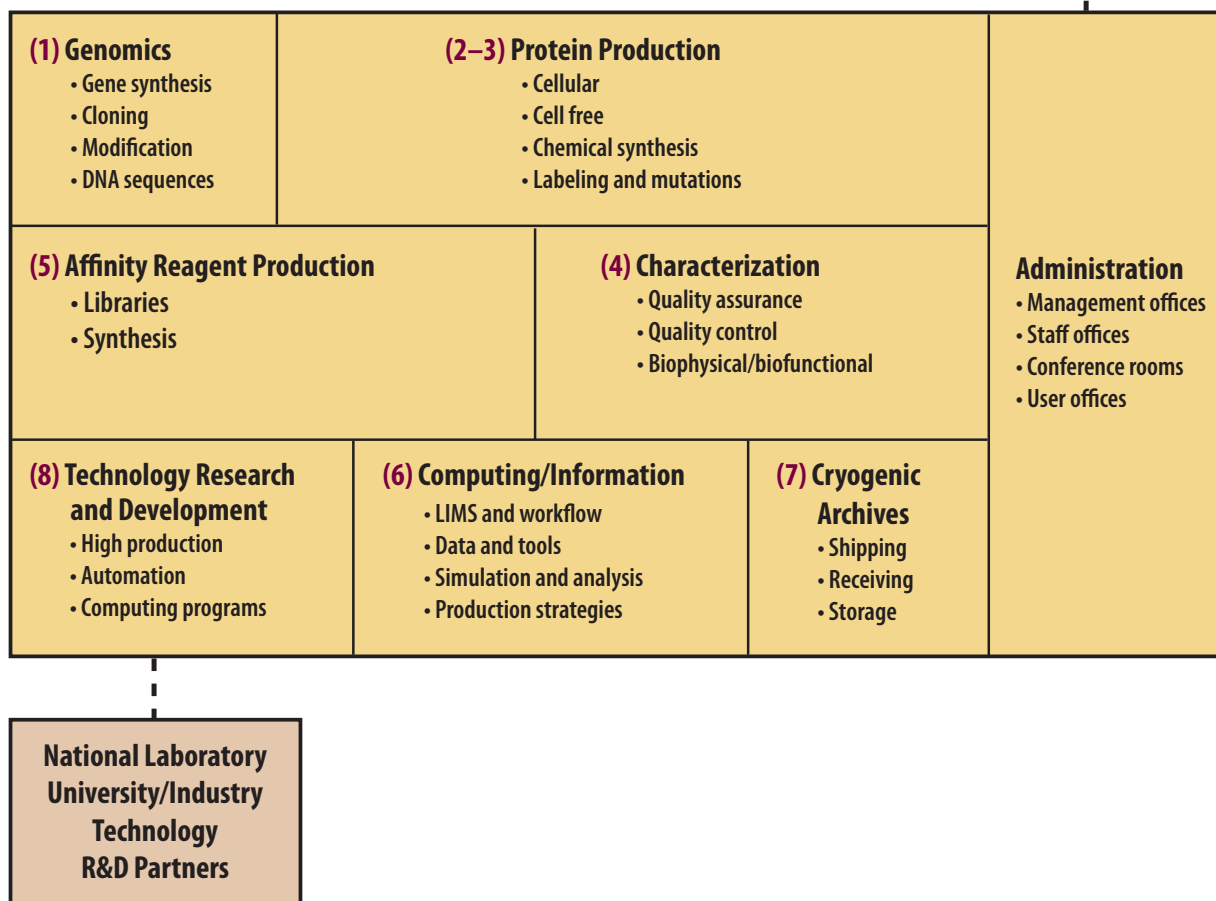
# Protein Production and Characterization

## (1) Inputs

- Gene Sequences
- Gene Clones

## Remote Characterization

- Facilities
- Specialty



## Workflow Process of the Protein Production and Characterization Facility

Note: Numbers and italicized words in parentheses below refer to terms used on charts beginning on next page.

### Inputs (1)

In its DNA sequence, every gene contains information needed by a cell to produce a specific protein. Scientists can use this information to make the same protein in the laboratory. The Protein Production and Characterization Facility will make proteins beginning with one of two inputs: Actual pieces of DNA that serve as molecular templates for producing given proteins (*Gene Clones*) or gene sequence information stored in databases—virtual pieces of DNA (*Gene Sequences*).

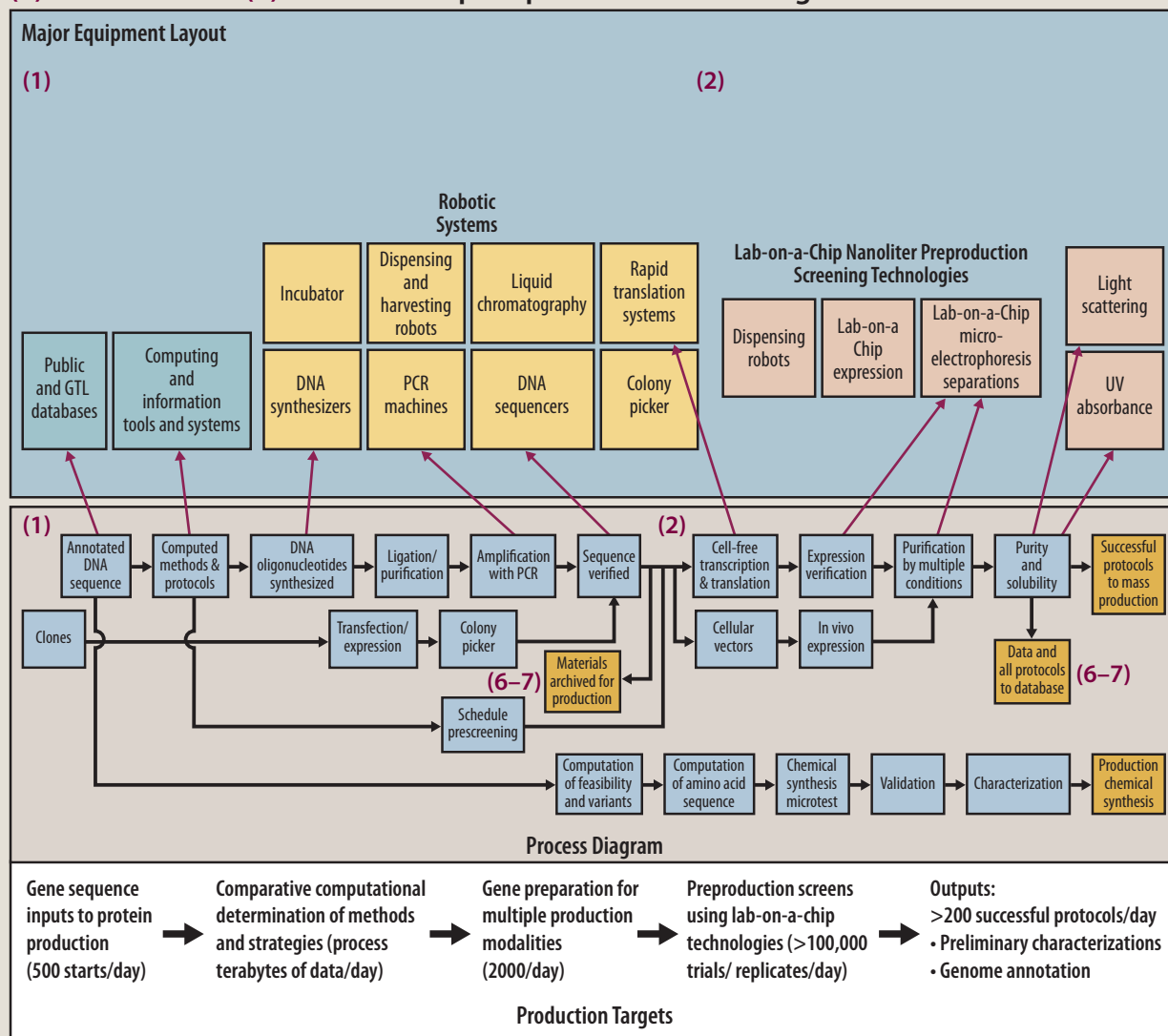
### Genomics (1)

With standard techniques, the gene sequence information can be used to construct a gene clone (*Gene Synthesis*). Cloning is accomplished by inserting the synthesized DNA segment into a cloning vector, usually a specific microbe or bacterial virus designed to over-express the protein of interest (*Cloning*). Choice of vector will vary, since all DNA sequences cannot be cloned in the same vector, nor can all proteins be produced in the same vector. In some cases, specific DNA sequence



# FACILITIES

## (1) Genomics and (2) Lab-on-a-Chip Preproduction Screening Lines



modifications will be needed before insertion [e.g., to increase the resultant protein's solubility or to change the way it interacts with other proteins (*Modification*)].

Cloning and modification can introduce errors into a given DNA sequence. A critical quality-control step, one of several in the protein-production process, is verification that the gene clone's DNA sequence is correct. This process uses the high-throughput DNA sequencing technology developed as part of the Human Genome Project (*DNA Sequencing*).

Virtually all steps in this process can be automated. A technician can obtain gene sequence information from a database and use genomics software to automatically direct a series of robots to produce a gene clone, verify the sequence, insert the clone into the appropriate

vector, and produce DNA samples ready for making proteins. A laboratory can run this process simultaneously on hundreds of different target gene samples.

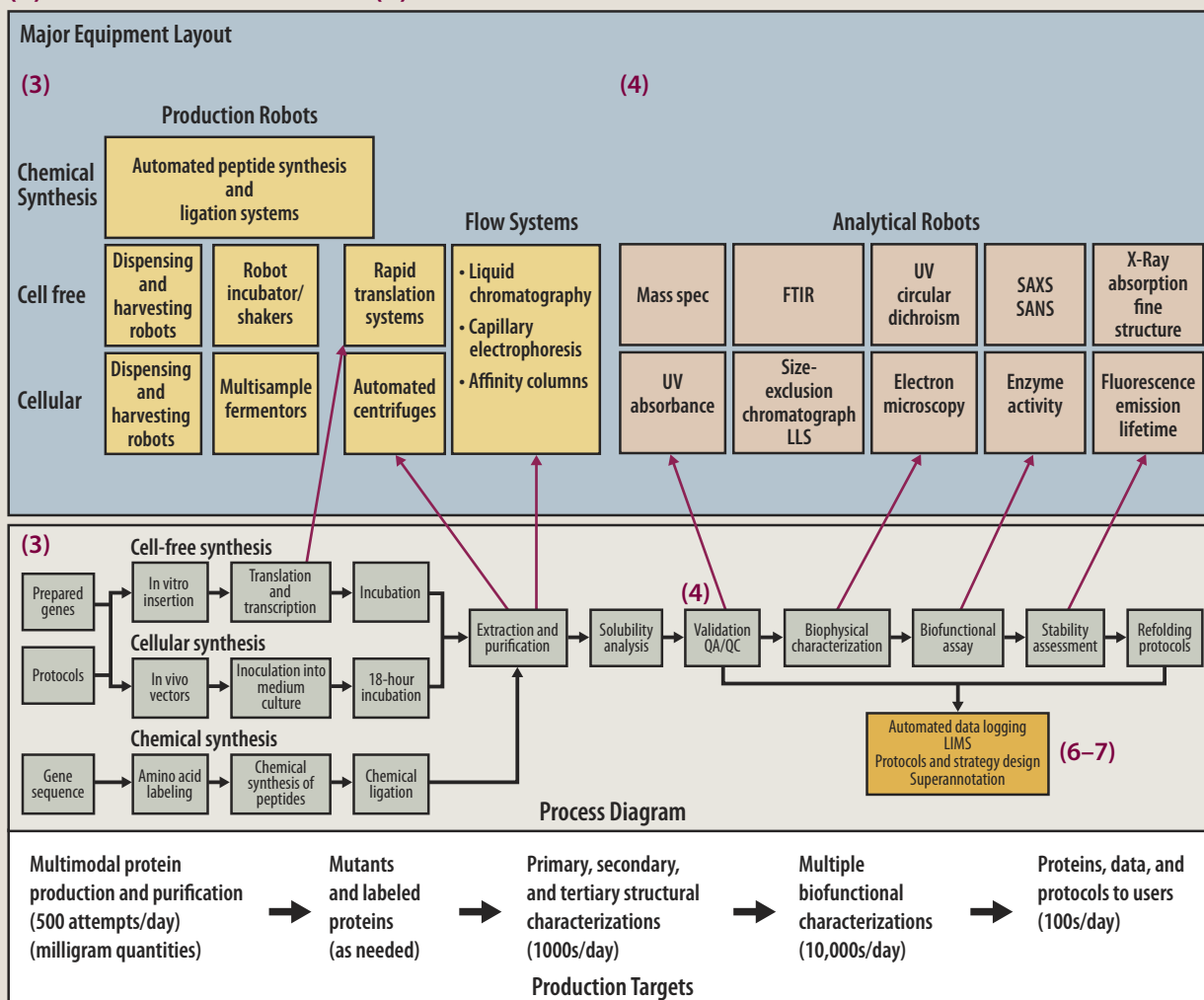
## Protein Production (2–3)

No single method will work equally well for all proteins, so several methods will be needed to produce difference proteins from gene clones or gene sequence information.

Preproduction screening will optimize production and purification methods for each protein of interest. Various production conditions will be tested using nanoliter volumes of reagents and a "lab on a chip" on which large numbers of synthesis and analysis steps can be carried out in parallel. Robotics and microfluidic

# Protein Production and Characterization

## (3) Protein Production and (4) Characterization Lines



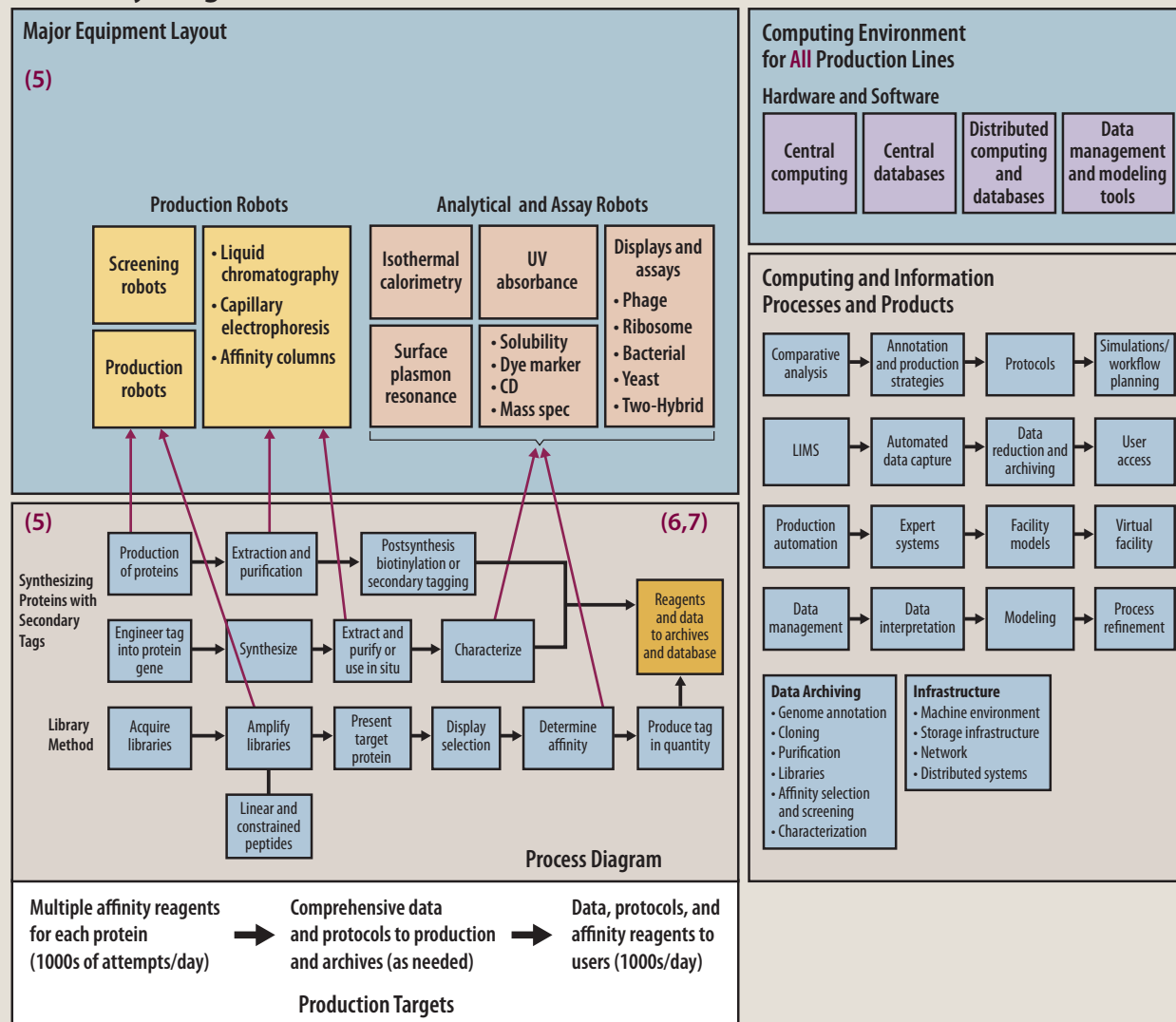
processes will be used to test various combinations of cloning vectors, reagents, and reaction conditions. The presence, level, and purity of protein expression will be checked using microchannel separations of reaction products combined with molecular-weight markers and various detection techniques (e.g., mass spectrometry, ultraviolet absorbance, and light scattering). Data will be entered into a computer and analyzed, and the best conditions and methods for large-scale protein production will be identified automatically.

During cellular protein production, vectors carrying the gene clone of interest are inserted into a bacterial host whose cellular machinery is used to produce the specific protein of interest (*Cellular Production*). The protein is extracted from the host cells and purified. Alternatively, proteins can be produced by mixing a DNA template with a set of purified enzymes and

chemicals normally used by the cell for protein production; only the protein of interest is produced without the need for a living cell (*Cell-Free Production*). Finally, well-established chemical-synthesis methods can be used to make short strings of amino acids that must then be hooked together to make complete proteins (*Chemical Synthesis*). These methods, especially chemical synthesis, can be used to introduce specific changes in a protein sequence such as modification of protein subunits or incorporation of radioactive isotopes needed in downstream analysis. All proteins will undergo purification using a variety of separation technologies (e.g., liquid chromatography, capillary electrophoresis, or affinity columns). Proteins also will need to be collected and maintained under specific conditions that enable them to fold into their natural, functionally active configurations.

# FACILITIES

## (5) Affinity Reagent Production Line



The protein-production process can be automated and run simultaneously on hundreds of samples to generate a vast array of normal or modified proteins ready for characterization.

## Characterization (4)

In addition to verifying the sequences of gene clones, we also need to characterize the proteins produced (and the processes used to produce them) to ensure their purity and biological behavior (*Quality Control* and *Quality Assurance*).

All proteins produced will be run through a battery of tests and screening procedures (*Biophysical Characterization*) to assess their quality and to provide initial

insights into their structures. For each protein, molecular weight, stability, and proper folding must be determined. No single test will be sufficient to characterize every protein adequately and accurately. Instead, a combination of various spectroscopic, separation, and imaging techniques will be used. Some proteins of particular interest to DOE, such as those involved in hydrogen production or cleanup of environmental contaminants, will be characterized further for biological function by assaying for specific enzymatic activity or binding properties.

Automated systems will simultaneously characterize hundreds of proteins for purity and, in some cases, function.

# Protein Production and Characterization

## Affinity Reagent Production (5)

A very useful product of this facility will be affinity reagents that can serve as molecular markers needed to “see” the proteins in cells as parts of multiprotein complexes or as they interact with other proteins or molecules in their normal functions. Multiple affinity reagents, produced by a variety of methods, will be needed for each protein, since each reagent will recognize and bind to a particular feature (e.g., a specific physical conformation or shape as well as specific sites responsible for protein function or activity).

Affinity reagents can be produced from “libraries” of potential binders (*Libraries*). Each contains, for example, millions of different antibody-like molecules. These libraries can be screened rapidly to identify sets of affinity reagents for each protein. Proteins also can be produced or synthesized (see Protein Production above) with molecular markers or tags built into each (*Synthesis*).

Almost all steps in this process can be automated and run in parallel so millions of potential affinity reagents can be made simultaneously and hundreds of proteins can be screened against these large libraries to identify binding markers.

## Computing and Information (6)

Both the production and research components of this facility need robust tools for tracking the many processes and products and associated R&D operations. A laboratory information management system (*LIMS*) is needed to track every sample and product that goes into or out of the facility and every process carried out as part of the facility (*Workflow*). LIMS will enable tracking of process efficiencies, product locations, status and availability of all facility research tools, and status of ongoing user projects. LIMS will allow

facility managers and researchers to monitor production strategies (*Production Strategies*) for both proteins and molecular tags, keep track of all data generated by the facility including successes and failures, and use all that information to predict, for example, which specific strategy would be most likely to work for a given protein (*Data and Tools*). Developing these data-analysis and process-simulation capabilities will increase facility operational efficiency and reduce costs (*Simulation and Analysis*). Moreover, the publicly available protocols of “lessons learned” will be a valuable resource that speeds progress in laboratories of scientists not physically using this facility.

## Cryogenic Archives (7)

Samples (DNA, proteins, affinity reagents) used and produced by this facility will be stored for future use, shipped to current users, and received from new users (*Shipping, Receiving, Storage*). Part of the centralized LIMS, all storage, shipping, and receiving data are key components in operating this high-throughput user facility. Many aspects of sample storage and shipping are automatable.

## Technology Research and Development (8)

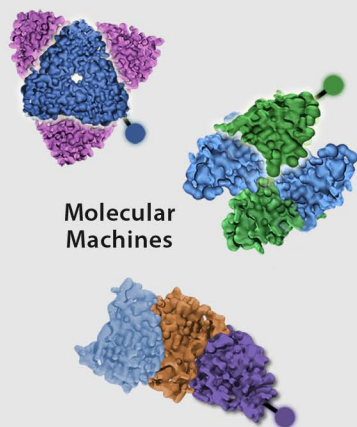
Item 8 is illustrated on first chart only, p. 133. While technologies currently exist to carry out all production and analysis steps described above, additional research and development are needed to make each individual step more efficient, cost-effective, and part of an automated, high-production assembly line (*High Production, Automation*). Development and use of computational tools for all aspects of facility operations will be extremely important (*Computing Tools*).





## 5.2. Facility for Characterization and Imaging of Molecular Machines

5.2.1. Scientific and Technological Rationale .....	140
5.2.2. Facility Description .....	141
5.2.2.1. Laboratories, Instrumentation, Quality Control, Computing, and Support .....	141
5.2.2.2. Production Targets .....	143
5.2.3. Technology Development for Expression, Isolation, and Purification of Molecular Machines .....	143
5.2.4. Technology Development for Identification and Characterization of Molecular Machines .....	144
5.2.4.1. Identification of Macromolecular Complexes.....	145
5.2.4.1.1. Mass Spectrometry .....	147
5.2.4.1.2. Separation-Based Techniques .....	147
5.2.4.1.3. Yeast 2-Hybrid .....	147
5.2.4.1.4. In Vivo Imaging Technologies.....	148
5.2.5. Technology Development for Biophysical Characterization .....	149
5.2.5.1. Structural Techniques .....	151
5.2.5.1.1. Crystallography .....	151
5.2.5.1.2. CryoEM Imaging of Isolated Complexes .....	151
5.2.5.1.3. Nuclear Magnetic Resonance .....	151
5.2.5.1.4. X-Ray Scattering .....	152
5.2.5.1.5. Neutron Scattering .....	152
5.2.5.2. Other Biophysical Techniques.....	152
5.2.5.2.1. Calorimetry.....	152
5.2.5.2.2. Force Measurements.....	152
5.2.5.2.3. Mass Spectrometry for Structural Characterization.....	152
5.2.6. Development of Computational and Bioinformatics Tools .....	153



Identify and characterize molecular complexes and other interactions.

## Molecular Machines

- ▶ Isolate and analyze molecular machines from microbial cells.
- ▶ Image structure and cellular location of molecular machines.
- ▶ Generate dynamic models and simulations of molecular machines.

# Facility for Characterization and Imaging of Molecular Machines

The Facility for the Characterization and Imaging of Molecular Machines will be a user facility providing scientists with the basis for understanding biochemical processes in microbes by determining how molecular complexes are formed and how they function.

## 5.2.1. Scientific and Technological Rationale

Microbes are biological “factories” that perform and integrate thousands of discrete and highly specialized processes through coordinated molecular interactions involving assemblies of proteins and other macromolecules often referred to as “complexes” or “molecular machines.” These biologically important protein-protein interactions (as well as protein-RNA, protein-DNA, and other biomolecular complexes) modify and dictate molecular states, which, in turn, integrate to define cellular physiology in response to genetic and environmental cues.

Understanding molecular machines, key players in various biochemical pathways, is central to systems biology. Many machines are short-lived or unstable and changing in composition, modification state, and subcellular location as they carry out vital functions that dictate how a cell or organism interacts with its environment. Many types of protein complexes exist in cells; complexes are associations that may be precursors to machines, associations that may not form contiguous machines, or associations that include a machine and appended molecules. A large number are assembly intermediates, while others are fully functional molecular machinery.

Key cellular multienzyme complexes can result in increased reaction rates, reduced side reactions, and direct transfer of metabolites, while many truly are machines that have moving parts or move other cellular entities (e.g., folding mechanisms and motors). So-called array machines such as light-harvesting systems, ribosomes, and others carry out intricate conversions in many organisms. Complexes also can be classified in an operational perspective from subcellular fractionation as stable and soluble, transient and soluble, and membrane associated.

As important as these machines are in cellular function, our current knowledge of them is quite limited. This is partly because proteins and other components of the complexes most often have been studied individually and in isolation and partly because they are highly

dynamic and inherently difficult to study. A cell's collection of molecular machines has intricate interrelationships that must be understood to determine how various environmental conditions influence pathways and how they differ from one organism to another. For example, specific pathways that will enhance hydrogen generation might be turned on or off by altering another pathway in an organism. We must determine the location and interactions of the molecular machines as they perform their critical functions in cells. This will require the most sophisticated and modern imaging technologies capable of resolving these details at multiple scales, from hundreds of nanometers to angstroms. Imaging technologies for identifying and locating (and collocating) machines in living cells will be incorporated into the Molecular Machines Facility. More extensive dynamic measurements that might track these machines through the life cycle of a cell will be incorporated into the Cellular Systems Facility, where the internal workings of cells will be monitored within well-defined communities and environments.

The goal of the Molecular Machines Facility is to provide researchers with the ability to isolate, identify, and characterize these functional microbial components and to validate their presence in cells using imaging and other analytical tools. The facility also will generate dynamic models and simulations of the structure, function, assembly, and disassembly of these complexes. Such efforts will provide the first step in determining how the large, dynamic network of cellular molecular processes works on a whole-system basis, how each machine is assembled in three dimensions, and how it is positioned in the cell with respect to other components of cellular architecture. Centralizing these analyses within a specialized and integrated facility will allow them to be conducted with higher performance, efficiency, fidelity, and cost-effectiveness. Many of the technologies discussed in this chapter are part of a long-lead and global development plan described in 6.0. GTL Development Summary, p. 191.

## Facility Objectives

- Discover and define the complete inventory of protein complexes in a microbe.
- Isolate complexes from cells using high-throughput techniques.
- Identify molecular components of complexes.
- Analyze the structure and predict the function of molecular machines. Determine basic biophysical and biochemical properties of these complexes.
- Validate the occurrence of complexes within cells and determine their location.
- Develop principles, theory, and predictive models for the structure, function, assembly, and disassembly of multiprotein complexes. Verify models with experimental data.
- Provide high-fidelity data and tools to the greater biological community.

## 5.2.2. Facility Description

### 5.2.2.1. Laboratories, Instrumentation, Quality Control, Computing, and Support

The Molecular Machines Facility will have several key capabilities to provide detailed insight into the form and function of protein complexes in a cell (see Fig. 1. Core Capabilities for Molecular Machines Facility, p. 142). The high-throughput facility will consist of a 125,000- to 175,000-sq.-ft. building housing core resources for cultivation, isolation, stabilization, identification, and analysis of molecular machines as well as necessary support systems. It will have extensive robotics for efficient sample production and processing and suites of highly integrated analytical instruments for sample analysis and molecular-machine characterization.

Instrumentation in the Molecular Machines Facility will include mass spectrometry (MS) for complex identification; electron, optical, and force microscopes for in vivo and in vitro imaging, localization, and characterization of complexes; and other analytical tools. The facility will make optimum use of state-of-the-art capabilities at such national user resources as synchrotrons, neutron sources, and electron microscopes as needed. Laboratories will be required for microbial cell growth, molecular biology, automated high-throughput sample

# FACILITIES

preparation, gene expression, protein-complex analysis based on MS, imaging of protein complexes, biophysical characterization, and quality assurance. Integrated with these laboratories will be computing resources for sample tracking; data acquisition, storage, and dissemination; algorithm development; and modeling and simulation. For multiprotein machines with structurally characterized components, high-performance computing will play a very significant role in building structural models of the machines and performing molecular dynamics simulations of their intermolecular interactions. The next generation of massively parallel processors in the 40- to 100-teraflop range will allow simulations of sufficient size and fidelity to make important contributions in explaining the mechanisms of machine construction and function.

Stringent quality-control protocols will be applied at each step. To get a complete picture of the complex network of molecular interactions, investigators will culture cells under a number of different conditions. They will work from insights provided by the Proteomics Facility, which will make temporal analyses to determine when and under what conditions specific proteins and machines occur. These protocols will result in potentially thousands of samples to be run through the analysis pipeline for each microorganism. Because of the diverse nature of protein complexes—stable, transient, membrane-associated, and others—multiple isolation approaches must be included. Additional technologies, especially imaging and other structural and biophysical characterization techniques, will be required to validate the machines' presence in living cells and to provide essential data that will enable insight into molecular-level interactions, kinetics, and thermodynamic properties.

The facility's computational requirements will be vast. Handling large amounts of data from diverse sources will be required, and these data must be integrated to provide a more complete view of the cell's interaction networks and to support sophisticated models of intermolecular interactions, structures, and function. In its analysis of protein machines, the facility will use the protocols and vast wealth of data on individual proteins being produced by structural genomics programs in other agencies, including the National Institutes of Health and National Science Foundation.

Offices for staff, students, visitors, and administrative support will be included, as well as conference rooms and other common space. The facility will house all equipment necessary to support its mission. The DOE facility-acquisition process will include R&D, design, testing, and evaluation activities for ensuring a fully functional facility upon completion.

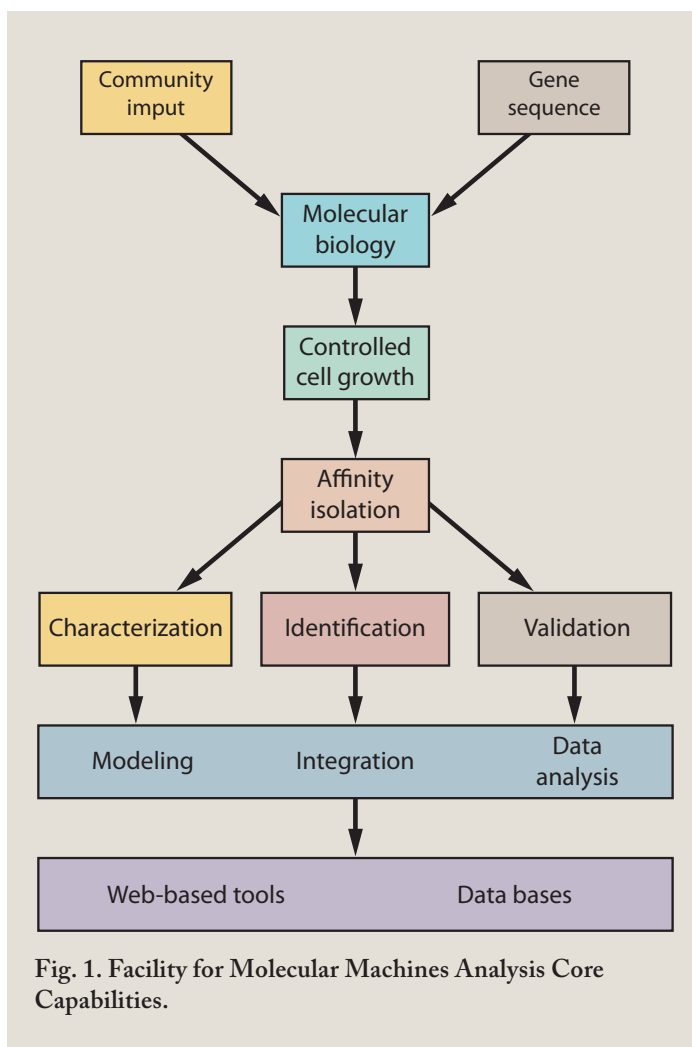


Fig. 1. Facility for Molecular Machines Analysis Core Capabilities.

## 5.2.2.2. Production Targets

To meet GTL program goals, researchers will need to generate protein complexes potentially involving thousands of different proteins (both natural and modified) from each organism studied. This means that many different species of microorganisms (many now unculturable) will need to be grown under a variety of carefully controlled conditions, producing millions of different protein complexes.

For a single microbe, the comprehensive mapping of the entire interactome (the summation of all protein-protein interactions in a cell) in a reasonable timeframe, with thousands of potential targets per microbe, will require throughput of the protein complex purification and identification pipeline of at least 10,000 pull-down attempts per year (to be statistically significant, these procedures must be run in triplicate and have a control, necessitating 40,000 attempts). The GTL program will need to analyze tens of microbes per year, which will require the ability to run about 100,000 pull-down attempts annually. All associated isolation, identification, and characterization procedures must be completed. The exact number of procedures will be determined by the governance processes that adjudicate the allocation of facility resources and set research and production priorities.

## 5.2.3. Technology Development for Expression, Isolation, and Purification of Molecular Machines

Technology must be developed to express intact protein complexes in wild-type and recombinant cultures under well-characterized conditions so molecular machines can be isolated and analyzed in initial studies as well as in those where machine functions are being optimized for specific characteristics. Maintaining high-quality, reproducible growth conditions will be essential for ensuring that high-quality data are generated. Conditions to be controlled must include environment (temperature, pH, media, substrate, light, oxygen); growth state (exponential, steady state, balanced, stationary); operation (batch, continuous); and harvest (age, lag, concentration, handling conditions). Due to the complexity of each process involved in producing the machines and the need for replicates, other quality-assurance and -control (QA-QC) techniques will be paramount to the facility's success (see Table 1. Technology Development Roadmap for Cell Growth and Processing, p. 144).

The isolation of molecular machines from cells is a challenging task. Molecular machines often are held together by weak interactions, making them fragile and difficult to isolate for analysis. Many such complexes are present only briefly or in very low amounts—sometimes just a few per cell (e.g., regulatory complexes, which are singularly important). Current techniques are inadequate for the robust, high-throughput isolation of protein complexes. The development and automation of such improved techniques is therefore an essential early goal of GTL pilot projects for this facility (see 3.3. Highlights of Research in Progress to Accomplish Milestones, p. 55). Data and reagents to be produced by the Protein Production and Characterization Facility will be central to isolating multiprotein complexes; in particular, affinity reagents would be used to isolate or “pull down” complexes (see Table 2. Technology Development Roadmap for Complex Isolation, p. 145). As a long-term goal, novel techniques for analyzing protein complexes in single microbes will be developed. The typical simplified “pipeline” for molecular-machine analysis would involve growing native or wild-type microorganisms under a reference state; using reagents to isolate the complexes from harvested cells; and then analyzing the complexes by MS, imaging, and other analytical tools. This process would be repeated under different growth states established by the Proteomics Facility, p. 155, to enable the comprehensive identification of machines chosen to be studied for the target organism.

Comprehensive identification of multiprotein complexes will require automating current methods for final sample preparation (i.e., desalting, buffer exchange, sample concentration, stabilization, and proteolytic digestion of samples). An important component of this facility is a highly integrated laboratory information management system (LIMS) that will track samples and manage data from cell cultivation through data archiving (see 5.2.6. Development of Computation and Bioinformatics Tools, p. 153).



## 5.2.4. Technology Development for Identification and Characterization of Molecular Machines

This facility is intended to provide detailed information on machine functions and the contributions of each to overall cell function. This analysis is a prerequisite for predicting a microbe's behavior under a range of natural and artificial conditions relevant to DOE missions. Due to the complexity and diversity of functions performed by molecular machines, multiple combinations of techniques and instrumentation must be used to identify and fully characterize all possible machines that a microbial cell is capable of producing.

Integration of multiple analytical and computational technologies will play a key role. Knowledge of a machine's static composition and structures is obtained by a variety of techniques. This information provides a starting point for following the machine's behavior in a living cell, for example, by scientists in their own laboratories and by users of the Cellular Systems Facility, p. 173. Imaging techniques can be used to follow the labeled components of a machine to trace its formation, movement, and dissociation in vivo by nondestructive techniques such as various types of fluorescence microscopy. Similarly, the high spatial resolving power of electron microscopy (EM) and X-ray microscopy can be used to localize machines in cells frozen at key functional time points. Further, X-ray and neutron diffraction and small-angle scattering can be used to help identify structural relationships among complex components.

Many analytical techniques can be used to identify and characterize proteins and protein complexes. Advantages, disadvantages, and potential areas of key method development are discussed in the sections below. Although not an exhaustive summary, they describe technology gaps that must be the subject of this facility's R&D.

**Table 1. Technology Development Roadmap for Cell Growth and Processing**

Technology Objectives	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
<b>Develop technologies for protein machine production: Cell growth and processing</b>  Cultivation systems: <ul style="list-style-type: none"> <li>Modified for endogenous protein isolations</li> <li>Wild type for exogenous complex isolations</li> </ul>	Define conditions to express and process active molecular machines  Develop methods: <ul style="list-style-type: none"> <li>Reproducible growth, real-time monitoring, sampling</li> <li>Novel culture approaches</li> <li>High-throughput controlled fermentations</li> <li>Sample archive documentation</li> <li>Functional assays for unknown isolated molecular machines</li> </ul> Evaluate commercial systems	Controlled cell growth, processing: <ul style="list-style-type: none"> <li>Modified cultures for endogenous complex isolation</li> <li>Wild-type microbes for exogenous complex isolation</li> <li>Large numbers of microbial clones with encoded tags</li> </ul> Database development Controlled bioreactors for cellular imaging Automation and standardization Standards, protocols, costs, QA/QC refinements Evaluation, incorporation of new technologies Development of methods for microbes and machines requiring specialized conditions	Establish high-throughput pipeline based on defined requirements, standards, protocols, costs  Scale up parallel processes for multiple organisms  Evaluate and incorporate new technologies  Use parallel processes for scaleup	Production of well-defined microbial samples for extraction and characterization of active molecular machines  Database from controlled cell growth with analysis of protein complexes and associated biocompounds  Well-managed biosample archive  Protocols

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.

## 5.2.4.1. Identification of Macromolecular Complexes

The four types of macromolecular machines (each containing proteins, nucleic acids, and small biomolecules) are water-soluble stable protein-protein complexes, water-soluble transient protein-protein complexes, membrane-associated complexes, and protein-nucleic acid complexes. Water-soluble complexes typically reside inside the cell, and stable complexes can be tagged readily for isolation and characterization. Technologies for this type of system are the most developed for high-throughput analysis but are by no means sufficiently mature to be applicable to the wide range of macromolecular complexes that conduct life's processes in microbial cells. Transient complexes typically cannot be isolated from cells and therefore must either be identified while in the cell or stabilized before isolation and analysis. Complexes that last for only fractions of a second may best be hypothesized first using computational approaches but can be detected experimentally with emerging techniques. Membrane-associated complexes contain fewer polar (hydrophilic) regions, making them poorly soluble in aqueous solutions. Protein-nucleic acid complexes can fall into any of these categories. Technologies for identifying these various types of macromolecular

**Table 2. Technology Development Roadmap for Complex Isolation**

Technology Objectives	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
<p><b>Develop high-throughput technologies for molecular machine isolation for full population of biomolecular complexes</b></p> <p>Soluble stable complexes</p> <p>Membrane-associated complexes</p> <p>Transient complexes</p>	<p>Define needs:</p> <ul style="list-style-type: none"> <li>Evaluation of commercial laboratory and LIMS resources, if available</li> <li>Protocol refinement, automation</li> <li>QA/QC</li> <li>Multistage isolation schemes using affinity reagents to minimize background interferences</li> <li>Microfluidic-based affinity reagent isolations to minimize sample size requirements</li> <li>Stabilization and cross-linking of less-stable and transient complexes</li> <li>In vivo validation approaches</li> </ul> <p>Develop:</p> <ul style="list-style-type: none"> <li>Continuous, automated processing</li> <li>Multiplexed pulldowns</li> <li>Novel affinity reagents and isolation schemes</li> </ul> <p>Develop:</p> <ul style="list-style-type: none"> <li>Solubilization of membrane-associated complexes</li> <li>Stabilization of complexes</li> </ul> <p>Develop stabilization and cross-linking</p>	<p>Pilot-scale isolation method:</p> <ul style="list-style-type: none"> <li>Scaleup from 100 assays per week to thousands per week</li> <li>Automated, continuous processing</li> <li>Assessment of bottlenecks, costs</li> <li>QA/QC</li> <li>Evaluation and incorporation of new technologies</li> <li>Methods for rapid elucidation of protein complex network linkage maps</li> </ul>	<p>Establishment of multiple parallel pipelines</p> <p>Evaluation and incorporation of new technologies</p>	<p>Complexes isolated to permit identification, imaging, and biophysical characterization</p> <p>Protocols</p> <p>Methods</p> <p>Databases and query tools</p>

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.

# FACILITIES

complexes are summarized in the following pages and in Table 3. Technology Development Roadmap for Complex Identification and Characterization, this page.

Analytical techniques for the identification and characterization of nucleic acid complexes are far less developed, in general, than those for protein-protein interactions. Many of the techniques discussed below also can be applied to this type of complex, but more development will be required, as shown in Table 3, this page.

**Table 3. Technology Development Roadmap for Complex Identification and Characterization**

Technology Objectives	Research, Design and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
<p><b>Develop technologies for complex identification and characterization</b></p> <p>Analysis by mass spectrometry (MS):</p> <ul style="list-style-type: none"> <li>• Identification and quantification of both digested peptides and intact proteins</li> </ul> <p>Data processing:</p> <ul style="list-style-type: none"> <li>• Data interpretation</li> <li>• Data archiving</li> </ul>	<p>Develop for mass spectrometry:</p> <ul style="list-style-type: none"> <li>• High-throughput complex analysis by MS</li> <li>• Improved data analysis</li> <li>• Improved methods for quantitation and determination of complex stoichiometry</li> <li>• Improved MS detection limits and dynamic range</li> <li>• Identification of protein complex modifications via top-down MS</li> <li>• Combined isolation and identification approaches</li> <li>• Improved online separations</li> <li>• Integrated, high-sensitivity analytical tools, eventually for single cells</li> <li>• Improved cleavage and digestion approaches</li> <li>• Improved ionization for broad classes of proteins</li> <li>• Microfluidic-based assays</li> </ul> <p>Evaluate commercial hardware, software, and instrumentation</p>	<p>Pilot scale:</p> <ul style="list-style-type: none"> <li>• Optimization of protocols with regard to throughput, reproducibility, costs</li> <li>• Improved MS data-analysis tools</li> <li>• Database development and query tools</li> </ul> <p>Assays:</p> <ul style="list-style-type: none"> <li>• Integrated “lab on a chip”</li> <li>• Probe-based affinity</li> <li>• Binding affinity</li> <li>• Automated neutron, cryoEM, and X-ray small-angle scattering</li> <li>• New technologies evaluated and incorporated</li> <li>• MS labeling for identification of contact interfaces</li> </ul>	<p>Establish high-throughput, automated pipelines:</p> <ul style="list-style-type: none"> <li>• Scale up via multiple parallel production lines</li> </ul> <p>Refine QA/QA protocols</p> <p>Automate data acquisition and data analyses</p> <p>Evaluate and incorporate new technologies</p>	<p>Capability for high-throughput protein complex analysis by MS</p> <p>Highly validated data of identified protein complexes</p> <p>Confirmatory analyses of protein complexes via biophysical techniques</p> <p>New tools for complex-identification analysis</p> <p>Databases for complex identification and characterization</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>• Binding affinity</li> <li>• Interaction interfaces</li> </ul>
<p><b>Biophysical Characterization</b></p> <p>Structural characterization</p> <p>Binding affinities</p> <p>Others</p>	<p>Establish structural and functional assays:</p> <ul style="list-style-type: none"> <li>• EM, SANS, SAXS, NMR</li> <li>• Approaches to identify contact faces</li> <li>• High-throughput binding affinity assay</li> </ul>			

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.

## 5.2.4.1.1. Mass Spectrometry

This technique, the workhorse analytical tool for all aspects of protein identification, can be adapted readily to the analysis of protein complexes. MS is particularly useful in identifying modified components (e.g., post-translational modifications, mutations, and others) that are important in effecting biological function. In addition, it has high sensitivity and is amenable to high-throughput analyses. Thus, MS currently is recognized as the most broadly applicable tool for large-scale identification of macromolecular complexes. Complexes must be isolated from cells before analysis, which requires the development and use of affinity reagents to isolate the target complex. Further, MS does have limitations for application to membrane-associated complexes because of the requirement to solubilize, separate, and ionize complexes before mass analysis. Although membrane-associated complexes can be solubilized in detergents and other solvents, these modified solutions are not readily adaptable to today's separation and ionization techniques typically employed with MS. Isolated membrane-associated complexes can be digested enzymatically before MS analysis of the resulting peptides, however.

Improved technologies for MS ionization, mass analysis, and detection are needed to handle the full range of complexes in cells with high sensitivity and wide dynamic range. Development in these areas should enhance the ability to analyze membrane-associated complexes. In addition, better sample-handling techniques before mass analysis, including microsample preparation and separation techniques, are necessary to improve detection limits and decrease the amount of sample. Improved methods for isolating complexes from cells are desired, especially affinity reagents and other isolation approaches that are more robust and universal. MS has great potential for quantitative determination of amounts of complexes in a cell and also for establishing complex stoichiometries; however, additional development of quantitative techniques is essential. Application of MS to transient complexes has been reported using cross-linking reagents and other approaches for stabilization, but future work should validate these approaches and make them more robust for routine use. Finally, improved computational tools are needed to provide automated MS data interpretation. Table 4. Performance Factors for Different Mass Analyzers, p. 148, compares available mass-analyzer technologies, their most common ionization modes, resolving powers, mass accuracies, and mass-to-charge ranges. Each of these techniques has some range of applicability in the Molecular Machines Facility.

## 5.2.4.1.2. Separation-Based Techniques

These technologies include a number of methods for characterizing and fractionating a wide range of complexes based on hydrodynamic radius. Separations are achieved, for example, via sedimentation velocity, size-exclusion chromatography, 2D electrophoretic gels, field-flow fractionation, and equilibrium dialysis. Many of these techniques are amenable to microtechnologies. Separation generally is accomplished with a range of such detection techniques as staining, fluorescence, and MS, many of which have wide-capacity capabilities and are fairly low cost. Protein components from complexes, however, can be identified only if standards are available to compare retention characteristics. The exception occurs when MS is used as the detector and the separated peaks can be identified from the resulting mass spectra. Recent developments have shown that microfluidic devices have very high peak resolving powers and very fast analysis times (seconds vs many minutes). Although additional development is required, they have the potential for analyzing components from single cells. In addition, they can be integrated with multiple sample-preparation steps, greatly decreasing the amounts of both sample and reagent needed for analysis. Simple versions of these "labs on a chip" have become commercially available and could be of immediate use for screening samples before full MS analysis is available (see Fig 2. Capturing Protein Complexes Using Fusion Tags, p. 149).

## 5.2.4.1.3. Yeast 2-Hybrid

These assays are applicable to any complex for which the cloned DNA encoding the machine components exists. A readily automatable technique, it provides good coverage of the various types of binary (pair-wise) interactions. It is a very good screening tool but has a number of problems with both false positives and false

## FACILITIES

negatives. The incidence of false-positive results increases as complexes become less stable; thus, the assays have limited use with transient complexes. Moreover, capabilities are needed to enhance applications to domain mapping and obtain low-order structure information. In general, this technique can be very useful as an initial screening tool before analysis by MS and other techniques.

### 5.2.4.1.4. In Vivo Imaging Technologies

Imaging tools can be used to provide high spatial resolution images of complexes in individual living cells. An important application of imaging tools will be to verify the formation of complexes identified by MS and map their locations in the cell as they perform their functions. Affinity reagents modified with fluorescent or other labels (depending upon detection modality) will be produced by the Protein Production and Characterization Facility. These reagents will be used to “tag” specific complex components to identify the locations of complexes within the cell and produce information on the dynamics of their assembly and disassembly. This information will provide additional insights into understanding the function of protein machines and will furnish valuable data for system-wide studies to be conducted in the Cellular Systems Facility.

Many types of imaging technologies can be employed to identify macromolecular complexes, including those based on optical, vibrational, X-ray, electron, and force microscopies. Within these general categories, some specialized techniques have specific applications to the analysis of macromolecular complexes in situ in live, fixed, or frozen cells. The strengths of imaging techniques typically include excellent detection sensitivity (in some cases, single-molecule detection) and the ability to characterize complexes in their natural environments in cells. Imaging techniques are applicable to all classes of complexes, providing that the identity of one or more components of the complex is known and that appropriate labeled molecules can be synthesized. For in situ measurements, the labeled molecules must be introduced successfully into cells in a manner that approximates natural conditions (i.e., does not interfere with protein associations and folding).

Currently, most imaging techniques are labor intensive and slow; robotics and automation, however, have the potential to provide faster sample throughput, and improved computational tools will enhance data

**Table 4. Performance Factors for Different Mass Analyzers**

Mass Analyzer	Most Common Ionization Modes	Resolving Power (FWHM)*	Mass Accuracy	Mass/Charge Range
Quadrupole	ESI	1000 to 2000	0.1 Da	200 to 3000
Time-of-flight (reflection or Q-TOF)	MALDI ESI	2000 to 10,000	0.001 Da	10 to 1,000,000 (200 to 4000 for Q-TOF)
Sector	ESI	5000 to 100,000	0.0001 Da	1000 to 15,000
Quadrupole ion trap	ESI	1000 to 2000	0.1 Da	200 to 4000
Linear trapping quad	ESI	1000 to 2000 (5000 to 10,000 in zoom scan mode)	0.1 Da	200 to 4000
Fourier transform ICR-MS	ESI MALDI	5000 to 5,000,000	0.0001 Da	200 to 20,000

\*Full width at half maximum (FWHM) defines how close two peaks can be and still be resolved (resolving power). The mass divided by the FWHM is the resolving power.

Table 4 compares performance factors for the different mass analyzer technologies envisioned for use in the Molecular Machines Facility. Ionization modes, resolving power, mass accuracy, and mass-to-charge range are important factors qualifying these techniques for various applications.



visualization and manipulation. Issues specific to some of the techniques are summarized here, and some additional information on other imaging tools is given in Table 1, p. 144; Table 2, p. 145; and 5.4. Facility for Analysis and Modeling of Cellular Systems, p. 173.

**Tagged Localization.** This technique can be used with optical, X-ray, or electron microscopies to identify sets of biomolecules labeled with appropriate tags. This in situ method is applicable to live (visible-light), fixed, or frozen cells (X-ray and electron); to tagged transient complexes; and to membrane-associated complexes. Development in optics would improve instrumentation and more versatile excitation sources, and continued probe enhancement is needed. Examples of recently reported tags used with various imaging modalities are lanthanide dyes, quantum dots, nanoparticles, and tetracystein-based ligands.

**Fluorescence Resonance Energy Transfer (FRET).** FRET can identify pairs of biomolecules labeled with tags and provide information on biomolecular interrelationships. This in situ method is applicable to live cells, tagged transient species, and membrane-associated complexes. It is particularly good for structure and binding of extracellular ligands.

**Scanning Probe Microscopy (SPM).** Capable of very high spatial resolution, SPM can identify protein associations by attaching a tagged probe molecule to the scanning tip. Depending on the length of analysis time, the probe can detect single molecules and thus capture information on transient complexes. Labor intensive and slow, this technique is best suited for the study of membrane-associated complexes with whole cells or for the study of isolated complexes. The probe, for example, can be used to identify sites on a cell surface for interactions. Identification is a one-at-a-time process unless multiprobe devices, each with individual probe molecules, can be employed. Now under development, such devices hold promise for allowing this technique to be applied in a highly parallel fashion.

## 5.2.5. Technology Development for Biophysical Characterization

Generating isolated molecular complexes offers a unique but extremely challenging opportunity to characterize a complex with a host of biophysical techniques toward the ultimate goal of fully understanding a specific machine's structure, activity, and underlying interactions and mechanisms. Initially, a suite of well-established techniques will be employed to characterize the basic biophysical properties of an isolated

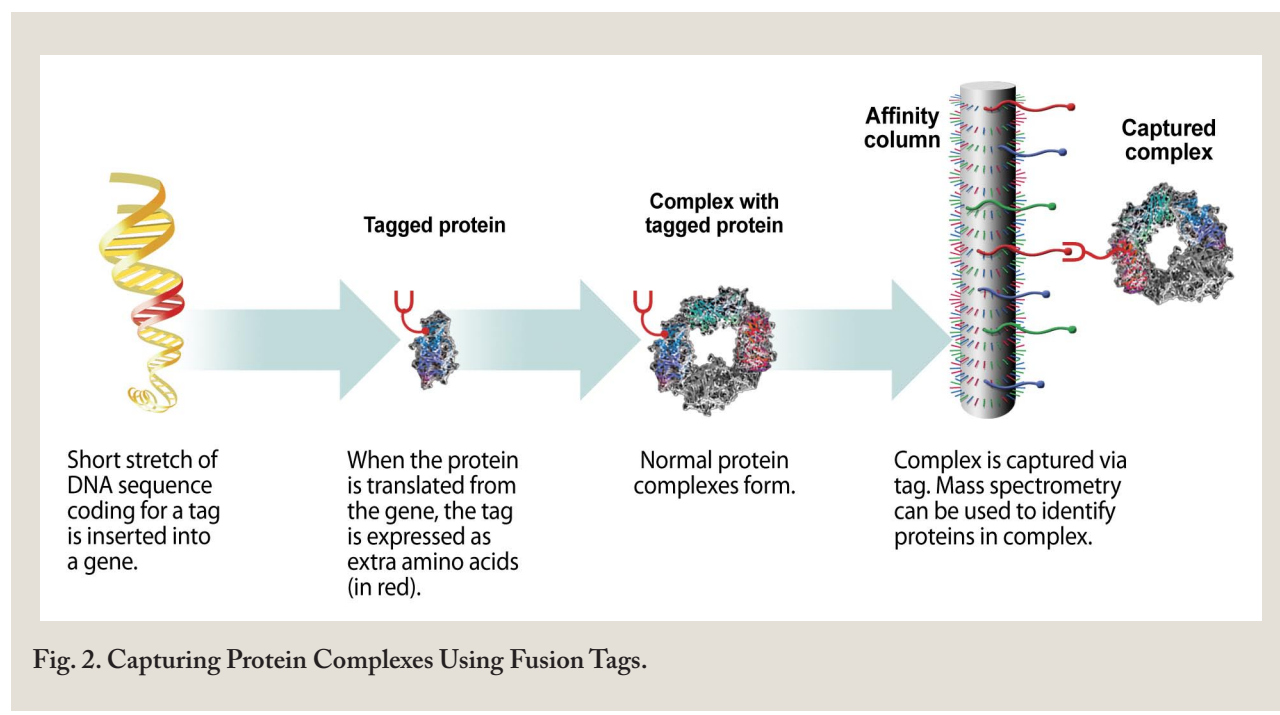


Fig. 2. Capturing Protein Complexes Using Fusion Tags.

# FACILITIES

complex. Obtaining systematic experimental information about dynamic complex behavior (including assembly and disassembly), combined with ongoing improvements in computational tools and modeling methods, will allow accurate simulations of molecular-machine activity at the heart of cellular function.

For stable protein-protein and protein-nucleic acid complexes, mechanistic understanding comes most readily with the highest levels of structural detail (general shape). Thus, atomic resolution generally is the ultimate goal in analysis of any biological structure. Crystallography and some imaging techniques offer this potential but have very specialized sample requirements and limitations, are not high throughput, and provide only a static picture of the complex.

Solution-based techniques such as cryoEM, NMR, and X-ray and neutron diffraction offer information that is lower resolution but can be related more directly to the molecule's structure in a more natural environment. Multiple tools obviously will be needed to obtain a more complete view of the structure of protein complexes, including shape, relationship of interaction faces, and stoichiometry. Three-dimensional images are obtained readily for proteins and protein complexes or machines that can be expressed, isolated, purified, and then crystallized for X-ray diffraction studies or dissolved to a sufficiently high concentration for NMR studies and scattering experiments. Such structural images have been obtained for quite large protein machines, for example, the bacterial ribosome containing some 55 proteins, additional strands of RNA, and other molecules. Some of these structural techniques are described below (see Table 5. Technology Development Roadmap for Complex Validation and Characterization, this page).

**Table 5. Technology Development Roadmap for Complex Validation and Characterization**

Technology Objectives	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
<p><b>Develop technologies for complex validation and characterization</b></p> <p>Analysis of complexes in vitro and in vivo:</p> <ul style="list-style-type: none"> <li>• Data processing</li> <li>• Data archiving</li> </ul>	<p>Develop:</p> <ul style="list-style-type: none"> <li>• In vivo imaging for validation and spatial and temporal studies</li> <li>• New labels for optical microscopy</li> <li>• Multimodal imaging approaches</li> <li>• Automated image acquisition</li> <li>• High-throughput image analysis</li> <li>• Improved spatial resolution</li> <li>• Environmental sample-manipulation techniques</li> </ul> <p>Evaluate commercial hardware, software, and instrumentation</p>	<p>High-throughput EM</p> <p>High-throughput optical methods</p> <p>Image-analysis software</p> <p>Automated sample acquisition</p> <p>Multimodal imaging</p>	<p>Automate image acquisition</p> <p>Automate data analysis</p> <p>Scale up acquisition and analysis</p> <p>Establish database</p> <p>Evaluate and incorporate new technologies</p>	<p>Data and characterizations:</p> <ul style="list-style-type: none"> <li>• Existence of complexes</li> <li>• Dynamic spatial relationships of proteins and other macromolecules in complexes</li> <li>• Local chemical and physical environment of complexes in cells</li> </ul>

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.

## 5.2.5.1. Structural Techniques

### 5.2.5.1.1. Crystallography

X-ray crystallography is employed widely for characterizing proteins and machines. Its strengths include high structural resolution, high reliability, and practically no limit on machine size. Extending these techniques to the scale of machines is a challenge in both data collection and analysis, with problems such as phasing requiring innovations. Although synchrotrons and enhanced detectors have improved greatly the speed of analysis once a crystal is available, difficulties in sample (crystal) preparation ultimately limit throughput and applicability to protein complexes. This technique requires moderate quantities of samples with high purity. Neutron crystallography has inherently lower throughput than the X-ray technique but is the method of choice for certain types of information about protein and nucleic acid complexes. Its attributes also include high spatial resolution and practically no size limit. A particular strength of neutron crystallography is the use of hydrogen-deuterium (H/D) contrast techniques to identify locations of key hydrogen atoms. Difficulties in sample preparation and requirements for large sample quantity and purity, however, greatly limit this technique's applicability.

### 5.2.5.1.2. CryoEM Imaging of Isolated Complexes

Electron cryomicroscopy (cryoEM) is an emerging tool with which the 3D structure of a molecular machine in a single conformation can be determined at subnanometer resolution without requiring a crystal. Studies can be conducted at different chemical or physiological states of the molecular machines so a snapshot of mechanistic processes can be captured. The flexible docking of individual components with the medium-resolution cryoEM map can provide snapshots of the molecular machine as it is being assembled. CryoEM has been applied successfully to several different molecular machines—ribosomes, chaperonins, and ion channels; throughput, however, is slow. New generations of instrumentation allowing higher-throughput data collection will be coupled with more robust and automated image-processing software. The prospect is high that cryoEM can extend molecular-machine imaging to near-atomic resolution in a single conformation. Such advancements would allow a molecular machine's polypeptide backbone to be traced. The challenge of reaching near-atomic resolution lies in software improvement for image reconstruction.

Future excitement in studying molecular machines via structural techniques lies in the interplay among results of multiple methods for refining mechanistic models at the atomic level. For instance, dynamic motion observed via fluorescent microscopy can be used to refine cryoEM structures of a mixture of conformational states. Simulation and modeling will provide feedback to iterative refinement cycles.

Purifying molecular machines in structurally homogeneous states will be difficult because a functional machine may have flexible domains and moving parts. These dynamic characteristics of molecular machines will present a great challenge to obtaining structures of molecular machines that exist only in mixed conformational states. CryoEM can record images of molecular machines with mixed conformations at moderately high resolution. Novel software must be developed for *in silico* separation of molecular-machine images in different conformations. A team effort of experimental and computational scientists will be needed to tackle this problem at both algorithmic and software levels. These types of investigations will require the fastest available computers for data sorting and structural refinement (see 4.2.1.5. Structure, Interactions, and Function, p. 88).

### 5.2.5.1.3. Nuclear Magnetic Resonance

NMR is well suited for detailed studies of select targets in simple mixtures of small molecular assemblies. It can probe the structures of biomolecular complexes at low resolution but requires large quantities of pure complexes at relatively high concentrations (100  $\mu$ M or more) free of nonspecific aggregation. Improvements are needed in data handling and analysis, sensitivity, sample throughput, and mass range (currently <100 kDa). NMR provides information on small biomolecular assemblies at atomic resolution. Of particular

## FACILITIES

importance is chemical-shift mapping and H/D exchange techniques that can be used to observe dynamics. NMR requires isotopic labeling (e.g.,  $^{15}\text{N}$ ,  $^{13}\text{C}$ ,  $^2\text{H}$ ) to identify specific moieties within the complex. Today, NMR has limited usefulness for the analysis of large complexes above about 100 kDa, but this limitation is likely to be circumvented in the future.

### 5.2.5.1.4. X-Ray Scattering

This technique can be applied broadly to a range of macromolecular complexes as long as the complexes can be purified. Small angle X-ray scattering (SAXS) provides moderate-resolution information on complex structure as well as stoichiometry. Quality control and standard database infrastructure are needed. This technique has the potential of being high throughput with the development of specialized robotic sample changers on instruments at synchrotron sources and of improved data-acquisition and -analysis tools.

### 5.2.5.1.5. Neutron Scattering

Small-angle neutron scattering (SANS) has many of the attributes and limitations of X-ray scattering (above). An additional attribute of SANS is that H/D contrast techniques can give more insight into interaction interfaces of macromolecular complex components. Improvements needed are similar to those listed above for X-ray scattering.

## 5.2.5.2. Other Biophysical Techniques

A number of other biophysical techniques, both mature and developing, can be employed to obtain information on kinetics, binding affinities, interaction interfaces, and others. Some of these techniques are outlined below.

### 5.2.5.2.1. Calorimetry

This group of techniques assesses interactions among complex components as well as complex stability. A relatively mature technique that can be used to characterize molecular interactions, calorimetry gives a quantitative measure of thermodynamic parameters associated with the interactions. Data interpretation requires extensive analysis, which would be facilitated by computational-tool development. Calorimetry is limited by its requirement of moderately large quantities (micrograms) of pure materials, although newer techniques may reduce these amounts. Also, the samples must be monodisperse (no aggregation). This technique does have the potential to be high throughput.

### 5.2.5.2.2. Force Measurements

Related to force microscopy (described above in 5.2.4.1.4 under Scanning Probe Microscopy, p. 149), force measurements assess interactions among complex components using chemically modified or tagged probe tips. This technique is capable of single-molecule detection and can assess a large range of forces. It requires a specific probe for each assay, however, and is labor intensive and slow. It is in the early stages of development but eventually could be made highly parallel using multiple probe tips.

### 5.2.5.2.3. Mass Spectrometry for Structural Characterization

MS can provide information on biomolecular interactions at low resolution when gas-phase H/D exchange reactions are used. In that case, surfaces inaccessible to exchange do not incorporate deuterium, providing information to identify solvent-accessible surfaces and protein interfaces. MS is applicable to larger biomolecular assemblies and has high sensitivity. It is most useful when 3D structural data are available. Under development, this structural application of MS is data intensive, requiring improved data handling and interpretation techniques (see Table 5, p. 150).

### 5.2.6. Development of Computational and Bioinformatics Tools

The Molecular Machines Facility has great need of computational tools for sample tracking, data acquisition, data interpretation, quality assurance, modeling and simulation, and many other tasks. A wide variety of these tools are being developed, and some specific application areas are outlined below (see Table 6. Computing Roadmap: Facility for Characterization and Imaging of Molecular Machines, p. 154).

- **Data-Handling and -Integration Techniques.** Not only will huge quantities (gigabytes and more) of MS data be obtained daily, but the data from many other analytical and structural tools must be integrated to understand the (1) complex network of interacting molecules in a microbial cell and (2) temporal and spatial dynamics of these biomolecular complexes. Computational tools for MS, while developed more than for almost any other analytical technique, still need further refinement to allow truly high throughput data acquisition and interpretation. As described above, imaging tools will require improved data acquisition and processing to improve sample throughput. Once all the data are collected, strategies must be designed for archiving and distributing these data to the biological community (see Table 6, p. 154).
- **Probabilistic Sequence or Structure Techniques.** These methods require a priori knowledge of classes of biomolecular interactions, but they can be high throughput and inexpensive. This tool is not CPU limited but needs more algorithm development and continuously updated databases. Also needed is further benchmarking with actual biological applications and improvements in strategies for integrating diverse data and providing reliability estimates.
- **Genome Context Analysis.** Relying on the size and extent of genome databases in its present state, this analysis does not give reliable predictions. The technique, therefore, requires more algorithm development and benchmarking for actual biological use, along with improved strategies for integrating diverse data.
- **Function-Based Inference of Participation in Complexes.** Though inexpensive once the required algorithms and databases are in place, this technique can provide interaction data that may be difficult to access experimentally, especially on short-lived complexes. These methods are not CPU limited but need more algorithm development, continuous database improvements, and benchmarking with actual biological applications.
- **Sequence and Structural-Motif Methods for Predicting Transmembrane Regions.** Limited only by availability of sequence and structural data on these regions, the strengths and development needs of these methods essentially are the same as for function-based techniques discussed above.
- **Computational-Sequence and Structural-Motif Methods for Predicting Regulatory Sites, Nucleic Acid-Binding Domains, and Target Sequence from Protein Structure.** These methods are limited only by availability of sequence and structural data on nucleic acid-binding proteins. Inexpensive to apply once algorithms and databases are in place, the techniques are probabilistic, requiring a priori data and the development of reliability estimates. Although not CPU limited, these methods do need more algorithm development, continuous improvement of databases, and ongoing benchmarking.



# FACILITIES

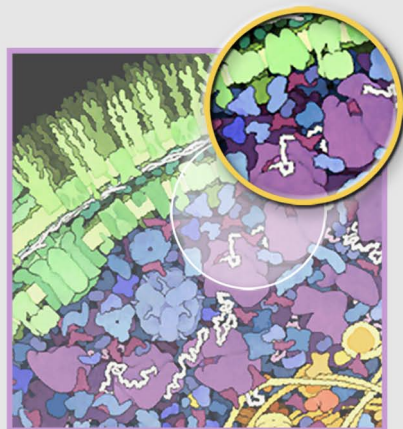
**Table 6. Computing Roadmap: Facility for Characterization and Imaging of Molecular Machines**

Topic	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment
<b>LIMS and Workflow Management</b> Participate in GTL cross-facility LIMS working group	Available LIMS technologies Process description for LIMS system Crosscutting research into global workflow management systems Approaches to guiding experiment-based production protocols to optimize protein production	Prototype molecular machine characterization LIMS system* Characterization design strategy Workflow management for identification and characterization Workflow process simulation	Molecular machines LIMS and workflow system Workflow integrated with other GTL facilities and experimental strategy system
<b>Data Capture and Archiving</b> Participate in GTL cross-facility working group for data representation and standards	Data-type models* Technologies for large-scale storage and retrieval Preliminary designs for databases	Prototype storage archives Prototype user-access environments	Archives for key large-scale data types* Archives linked to community databases and other GTL data resources GTL Knowledgebase feedback
<b>Data Analysis and Reduction</b> Participate in GTL cross-facility working group for data analysis and reduction	Algorithmic methods for various modalities* Grid and high-performance algorithm codes Design for tools library Approaches for automated image interpretation in confocal light microscopy and FRET	Prototype visualization methods and characterization tools library* Prototype grid for data analysis, with partners Prototypes for automated image interpretation in confocal light microscopy/FRET Analysis tools linked to data archives	Production-analysis pipeline for various modalities* on grid and HP platforms Automated image interpretation in confocal light microscopy, FRET Repository production-analysis codes Analysis tools pipeline linked to end-user problem-solving environments
<b>Modeling and Simulation</b> Participate in GTL cross-facility working group for modeling and simulation	Technologies for: Fixed and flexible docking and constrained molecular dynamics Low-resolution cryoEM data modeling and reconstruction Reconstruction of protein interaction and regulatory networks Multiscale stochastic and differential equation network models	Automated production pipeline (experimentally guided molecular docking and machine dynamics; efficient modeling methods for 3D CryoEM data reconstruction) Mature methods for reconstructing protein-interaction and regulatory networks	Production pipeline and end-user interfaces for genome-scale fold prediction Production codes for scattering-data modeling
<b>Community Data Resource</b> Participate in GTL cross-facility working group for serving community data	Data-modeling representations and design for databases: Protein machine catalog, protein machines models and simulations, interaction network models and simulations, protein machine methods and protocols	Prototype database End-user query and visualization environments Integration of databases with other GTL resources	Production databases and mature end-user environments Integration with other GTL resources and community protein-data resources
<b>Computing Infrastructure</b> Participate in GTL crosscutting working group for computing infrastructure	Analysis, storage, and networking requirements for Molecular Machines Facility Grid and high-performance approaches for large-scale data analysis for MS and image data; requirements established	Hardware solutions for large-scale archival storage Networking requirements for large-scale grid-based MS and image data analysis	Production-scale computational analysis systems Web server network for data archives and workflow systems Servers for community data archive databases

\* Data types and modalities include MS, NMR, neutron scattering, X-ray, confocal microscopy, cryoEM, and process metadata. Large-scale experimental data results are linked with genome data, and feedback is provided to GTL Knowledgebase.

## 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



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.

# Facility for Whole Proteome Analysis

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.

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.

## 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.

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.

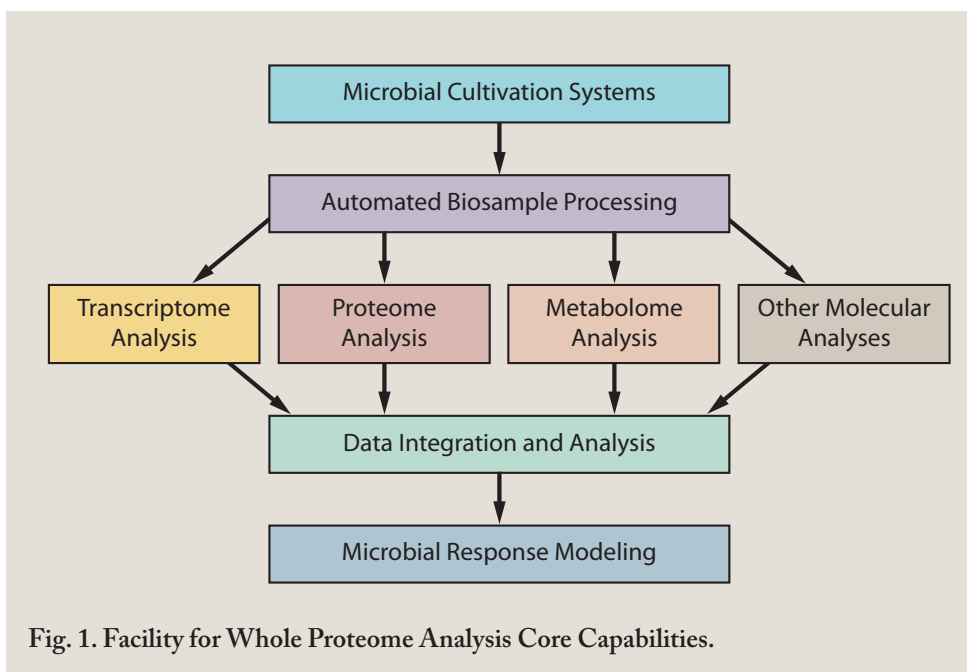
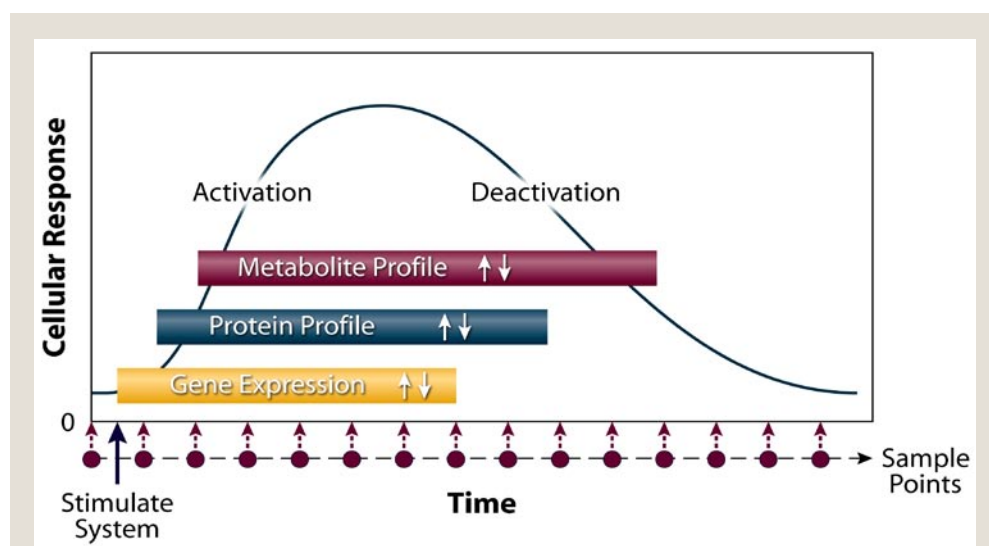


Fig. 1. Facility for Whole Proteome Analysis Core Capabilities.

## 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 at 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 ( $3 \times 10 \times 3$ ) 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
<b>Controlled Cell Growth, Analysis</b>  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: <ul style="list-style-type: none"> <li>• Appropriate parameters, culture variability</li> <li>• Workflow processes</li> <li>• Scale factors</li> <li>• Hardware, software, instrumentation</li> </ul> Develop: <ul style="list-style-type: none"> <li>• Reactor and instrumentation, interfaces, sampling methods</li> <li>• Reactor-based growth models, simulations</li> <li>• Searchable sample archive</li> <li>• Isotope labeling</li> <li>• High-throughput cultivation, isolation of community members</li> </ul>	Pilot: <ul style="list-style-type: none"> <li>• High-throughput controlled cell growth, processing</li> <li>• Methods for large sample collection</li> <li>• Online analytical systems for high-throughput metabolite measurements</li> <li>• Experiment and sample database</li> <li>• Automation, standardization, protocols</li> </ul> Develop methods: <ul style="list-style-type: none"> <li>• Commensal cocultures</li> <li>• Extremophiles</li> <li>• Biofilms and structures</li> <li>• Sample receipt and delivery</li> </ul>	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
<b>High-Throughput Gene-Expression Profiling</b> Sample processing Data processing Quantitation QA/QC	Define: <ul style="list-style-type: none"> <li>Workflow processes</li> <li>Improved detection limits, reproducibility, dynamic range</li> <li>Hardware, software</li> <li>Lab automation, robotics</li> <li>QC instrumentation, processes</li> </ul> Develop: <ul style="list-style-type: none"> <li>Multipurpose, multiorganism array platform</li> <li>In vivo testing platforms</li> <li>Expression database</li> <li>Commercial array applications</li> </ul>	Expression pipeline optimization, scaleup: <ul style="list-style-type: none"> <li>Improved standards, protocols, costs</li> </ul> Pilot: <ul style="list-style-type: none"> <li>Array processing pipeline</li> <li>Expression database</li> <li>In vivo testing pipeline</li> </ul>	Establish high-throughput pipeline based on defined requirements, standards, protocols, costs, and adopted industry standards: <ul style="list-style-type: none"> <li>Array processing pipeline</li> <li>Expression-experiment database</li> <li>In vivo expression-testing pipeline</li> </ul>	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.



## FACILITIES

- **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
<b>High-Throughput Protein Profiling</b>  Sample processing MS for global proteomics Other analysis techniques Data processing and analysis QA/QC	Define: <ul style="list-style-type: none"> <li>Workflow processes</li> <li>Lab automation and robotics</li> <li>QC instrumentation and processes</li> <li>Improved detection limits, reproducibility and dynamic range</li> <li>Hardware, software, instrumentation</li> </ul> Develop methods: <ul style="list-style-type: none"> <li>Peptide identification and quantitation</li> <li>Identification of protein modifications</li> <li>Analysis of intact proteins, including membrane associated proteins</li> </ul>	Whole-proteomics pipeline: <ul style="list-style-type: none"> <li>Optimization and scaleup</li> <li>Improved standards, protocols, costs</li> <li>Pilot of global proteomics database development</li> <li>Determination of global state of modification of cellular proteins</li> </ul> Evaluate and implement: <ul style="list-style-type: none"> <li>Hardware advances</li> <li>Software advances</li> <li>Instrumentation advances</li> </ul>	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

# FACILITIES

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
<b>Strengths</b>	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
<b>Weaknesses</b>	Derivatizing 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
<b>Development Needs</b>	Robustness Improved chromatographic resolving power Improved dynamic range Metabolite databases Computational tools for predicting metabolites		Robustness Dynamic range Cryogenic probes Microprobes and nanoprobe 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  $^{13}\text{C}$ -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
<b>High-Throughput Metabolite Profiling</b>  LC/MS and NMR methods for metabolite discovery  Sample processing  Data processing  Analysis, quantitation, QA/QC	Define requirements: <ul style="list-style-type: none"> <li>• Workflow processes</li> <li>• Detection limits, reproducibility, dynamic range</li> <li>• Lab automation, robotics, QC</li> <li>• Robust LC-NMR techniques</li> <li>• Hardware, software, instrumentation</li> </ul> Develop methods: <ul style="list-style-type: none"> <li>• Identification and quantitation</li> <li>• QA/QC with metrics</li> <li>• Sample processing</li> </ul>	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
<b>LIMS and Workflow Management</b>  Participate in GTL cross-facility LIMS working group Develop technologies and methods to: <ul style="list-style-type: none"> <li>• Manage massive dataflow</li> <li>• Process and integrate data</li> <li>• Manage workflow</li> <li>• Conduct QA/QC</li> <li>• Deploy collaborative tools for shared access to data and processes</li> </ul>	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
<b>Bioinformatics</b>  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: <ul style="list-style-type: none"> <li>• Whole proteome analysis for each GTL organism</li> <li>• Experiment templates and data sets for modeling and simulation</li> <li>• Defined experiment archive integrated with data and analysis from each analytical pipeline</li> <li>• Molecular profiling context-dependent database</li> </ul>
<b>Computing Infrastructure</b>  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

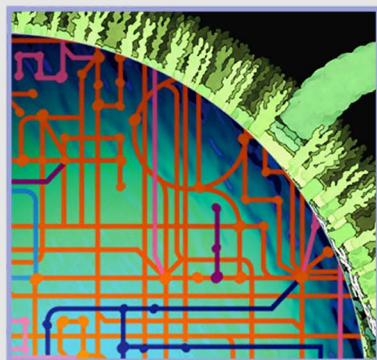


## 5.4. Facility for Analysis and Modeling of Cellular Systems

5.4.1. Scientific and Technological Rationale .....	174
5.4.1.1. Probing Mixed Microbial Populations and Communities .....	175
5.4.1.2. Foundations for Community Analyses .....	177
5.4.2. Facility Description .....	178
5.4.2.1. Laboratories, Instrumentation, Quality Control, Computing, and Support .....	178
5.4.2.2. Performance and Production Targets .....	178
5.4.3. Technology Development for Cultivation of Microbial Communities .....	179
5.4.3.1. Requirement Examples .....	180
5.4.4. Development of Genomic Capabilities .....	181
5.4.5. Technology Development for Imaging and Spectroscopy .....	181
5.4.5.1. Analytical Characterization of Cellular Systems .....	181
5.4.5.1.1. Examples of Analytical Requirements .....	184
5.4.5.1.2. Monitoring and Interacting with Cellular Systems .....	185
5.4.5.1.3. Technology Development Progress and Benefits .....	185
5.4.5.2. Imaging Macromolecular Complexes .....	186
5.4.5.3. Development Options .....	187
5.4.6. Development of Computing Capabilities .....	187



## Facility for Analysis and Modeling of Cellular Systems



Achieve an in silico, predictive understanding of microbes in their natural environments.

### Cellular Systems

- ▶ Integrate knowledge and models to understand the structure and functions of cellular systems, from single cells to complex communities.
- ▶ Integrate imaging and other technologies to analyze molecular species from subcellular to ecosystem levels as they perform their functions.

The Facility for the Analysis and Modeling of Cellular Systems will be a user facility to provide scientists with insight into the responses and functionality of microbes and microbial communities in complex environments. Modeling and real-time functional mapping of processes from the molecular through the ecosystem levels will be used.

### 5.4.1. Scientific and Technological Rationale

The Facility for Analysis and Modeling of Cellular Systems will be the GTL capstone needed to provide the ultimate integration of analytical capabilities and knowledge synthesis critical for systems biology. Users of this facility will investigate how microbial communities and their subsystems of cells function together to sense, respond to, and modify their environment. They will accomplish this by dynamically identifying, localizing, and quantifying molecular machines and all other important biomolecules and their interactions as they carry out their critical roles throughout microbial and community life-cycles. This grand challenge for biology ultimately must be addressed before scientists can develop and test models to predict the behavior of microbes and take advantage of their functional capabilities.

This facility will provide the ultimate testing ground for fully integrated models developed from component models created from ongoing research and from previous facility data, modeling, and experimentation. The experimental capabilities of the facility will drive a new generation of systems models. Essential aspects of the computational challenges and conceptual roadmaps are described in 4.2.1. Theory, Modeling, and Simulation Coupled to Experimentation of Complex Biological Systems, p. 85; and roadmap tables beginning on p. 91 (see Fig. 1. Probing Microbial Communities, p. 176).

The other three GTL facilities will provide new high-throughput production and analysis capabilities to define and understand component parts and processes of microbial systems and analyze physiological and molecular conditions on a global level (i.e., measure the properties of samples comprising large numbers of cells). One of the key insights from recent research, however, is that microbial communities are dynamic and highly structured physically and functionally, suggesting that ensemble measurements that look at properties averaged across many cells can reveal only part of the picture.

To address this challenge, the Cellular Systems Facility will focus on dynamic systems-level studies, ranging from molecular processes within individual living cells to complex, structured microbial communities. Microorganisms in such communities—microbial mats and biofilms, for example—occupy various microniches established as a result of coupled biological, chemical, and physical interactions. Each member of the community carries out unique functions that can vary in space and time but are integral to community stability and overall function. The Cellular Systems Facility will provide the underlying capabilities to allow the spatial and temporal analysis of these complex microbial systems in a concerted and integrated way, from molecular processes to ecosystem functionality. This is a daunting challenge, partly because the complex multicellular drama is playing out at submicron scales. Nonetheless, we will need to dynamically image and functionally analyze the critical substructures and molecular species within microbes and their communities and develop models that describe and predict their behaviors. This capability builds on the Molecular Machines Facility, which will focus on intracomplex imaging to determine molecular makeup and structure and on intracellular imaging to localize machines within the cell. (See box, Cellular Systems Facility Objectives, this page; Fig. 1, p. 176; sidebar, Group Living and Communicating, p. 18; Fig. 1. DOE Genomics:GTL High-Throughput User Facilities, p. 103; and Fig. 3. GTL Facilities: Core Functions and Key Interactions, p. 108, from 5.0. Facilities Overview).

The analytical and conceptual challenges of this ultimate step in systems microbiology will require unprecedented technical and computational resources and infrastructure far beyond the reach of individual investigators.

## Facility Objectives

- Relate community composition, structures, and functions to environmental physicochemical conditions measured at the scale of microbial communities—an overlay of community physical and functional maps.
- Determine community composition, relative positioning of members, and phenotypes.
- Analyze overall community functionality and distributions and fluxes of molecular species.
- Dynamically image critical molecules and substructures as they function intra- and intercellularly.
- Develop models of microbial function at the molecular, cellular, community, and ecosystems levels.
- Provide protocols, data, models, and tools to the community.

### 5.4.1.1. Probing Mixed Microbial Populations and Communities

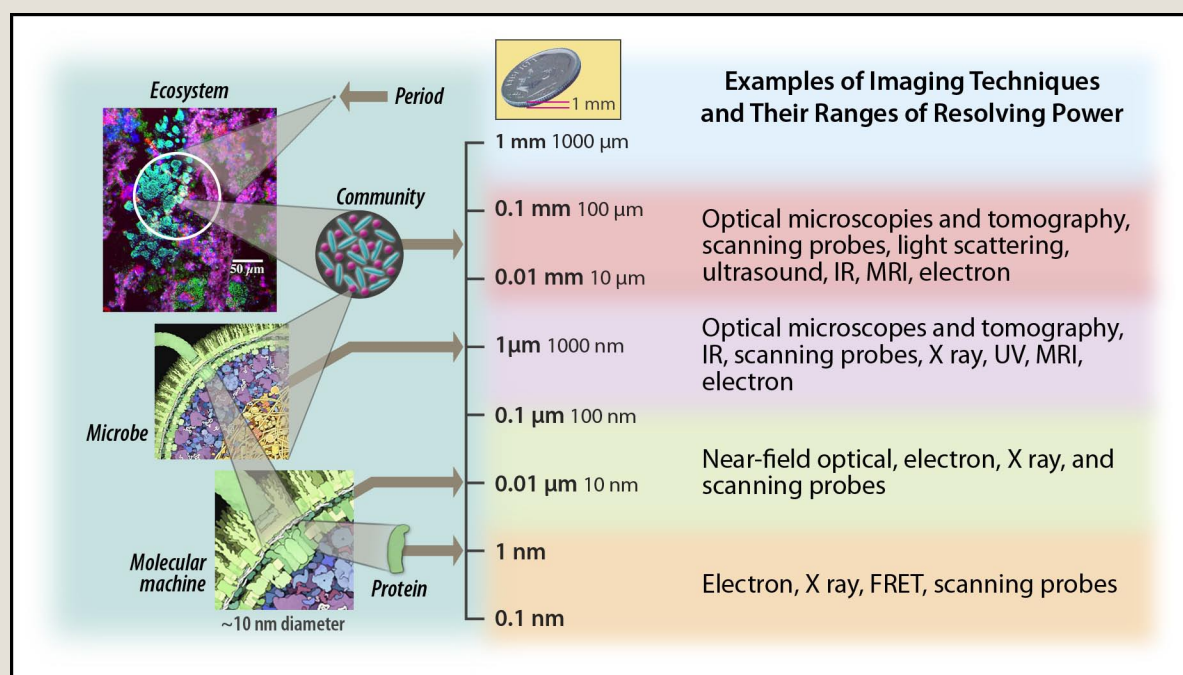
In microbial communities, the complex and dynamic set of interactions that we seek to analyze is taking place in an area far smaller than the period at the end of this sentence. To understand community function, we first must be able to analyze environmental and community structure and composition at high resolution (Fig. 1, p. 176). On an even-smaller scale (roughly 1/500<sup>th</sup> the size of a period), we must be able to peer inside an individual microbe in a nondestructive way to locate and continually track essential biomolecules that reveal the cell's inner workings; these biomolecules include DNA, RNAs, proteins, protein complexes, lipids, carbohydrates, and metabolites (Riesenfeld, Schloss, and Handelsman 2004; Johnston et al. 2004; Schaechter, Kolter, and Buckley 2004). Models then must be developed to describe key features of these biological interactions within the physicochemical environment and predict how the system will evolve structurally and functionally. Robust models are required to conceptualize these intricate systems, formulate meaningful hypotheses, manage the ensuing complexity and sheer volume of information that comes from experiments, and, ultimately, allow incorporation of the resulting science into applications.

Key information we seek about microbial-community function includes:

- Community arrangement and its physical environment
- Members and where they reside in relation to each other and their environment

# FACILITIES

- Phylogenetic, phenotypic, and functional properties of members. Do the genes in one organism regulate gene expression in another organism?
- Microbial interactions among themselves and with their environment
  - Information flow (e.g., genes, signaling molecules) within the community
  - Ecological interactions (e.g., predation, symbioses) among the various members
  - Food webs and communications
  - Excretions, secretions, and consumptions
  - Interactions of secreted components and nutrients with the environment
- Energy and element flow through communities and cells and its regulation
- Intrinsic biological (genetic) and environmental factors that control the structure, stability, and functioning of communities



**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. These and other techniques are discussed in section 5.4.5, beginning on 181.

- Trajectory of community evolution
  - Life strategies of each population in the community
  - Role of lateral gene-transfer processes in microbial evolution and community metagenome
  - Senescence, death, and turnover rate (consequence of death?)
  - Community resilience (biodiversity, stability)

## 5.4.1.2. Foundations for Community Analyses

Before undertaking these analyses, we will have a growing body of knowledge (incorporated in the GTL Knowledgebase) and capabilities from work funded by other agencies and within the GTL program and facilities. These resources will include:

- Cooperative analyses of comprehensively annotated genomes of individual microbes and the community metagenome to estimate the genetic potential of individuals and the community.
- Many critical proteins encoded in the community's genome expressed and characterized to produce a substantial body of functional characterization data incorporated into gene-annotation data sets. These data will provide significant insights into “interesting” processes that need to be pinpointed and analyzed in the context of a complete system. Since studies can be performed on the basis of sequence alone, we will have circumvented the fact that these microbes are largely unculturable.
- Ability to produce multiple affinity reagents for any produced proteins and other such biomolecules as RNA and some metabolites that can be used to locate, track, and manipulate these entities within living systems. Fusion tags can be incorporated in a variety of ways.
- Extensive measurements at the global level on bulk and ensemble samples to ascertain the phylogenetic and physiological state of member microbes or the entire community under relevant conditions. Temporal relationships will be revealed through repeated sampling and process interplay via extensive linked measurements of the transcriptome, proteome, metabolome, and other biological and physicochemical variables.
- Analysis of the structure and function of critical molecular complexes in vitro and insights to determine where and in what context they carry out their cellular functions.
- Extensive databases and exploratory tools to begin deriving underlying principles at the molecular, cellular, and community levels and the ability to begin encompassing complexities in detail as processes play out in real, nonlinear, coupled systems.
- Extensive models at molecular, cellular, and community levels to support creating and simulating hypotheses in a systems context. These models and simulations will be used to design and gain insight into experimental campaigns and protocols and provide advanced knowledge of key experimental variables that must be captured in ensuing research.

Even when all this information is at hand, unraveling how all these entities and processes act together in a continuous, concerted way—from molecular to community levels—will remain a grand challenge in accomplishing DOE mission goals. All technologies that created this body of knowledge must be specialized in innovative ways to provide the same information at a microscopic (actually nanoscopic) scale. This facility will be capable of analytical measurements that are nondestructive and done in real time in living cells within a well-defined global and dynamic system. Understanding how these individual cells interact and function as a unit—a microbial tissue in some respects—to carry out complex processes is key to unlocking their vast potential for important applications and achieving our science goals.



## 5.4.2. Facility Description

The Cellular Systems Facility will combine advanced computational, analytical, and experimental capabilities for integrated analysis of spatial and temporal variations in biological systems—how, when, and why the various system components appear, disappear, function, and remain. The facility will determine the state of cellular systems, from the internal makeup, structure, and dynamics of individual microbial cells to complex communities and their environments. To achieve a systems-level understanding, simulation and modeling must be coupled tightly with experiments to define and analyze the complex regulatory and metabolic processes in microbial cells and communities. This facility will emphasize concurrent and dynamic measurements of proteins, molecular complexes, intracellular metabolites, regulatory molecules, and gene transcripts. The aim is to establish the state of cells within populations and communities as a function of changes in physicochemical and biological conditions, emphasizing measurements at spatial and temporal resolutions appropriate for the entities being measured.

### 5.4.2.1. Laboratories, Instrumentation, Quality Control, Computing, and Support

The facility will be 100,000 to 150,000 gross square feet, with laboratories containing necessary cultivation, isolation, and analysis instrumentation. Its frontier instruments will incorporate capabilities and techniques for extensive environmental control and monitoring, manipulating communities in real time in various ways, and temporal and spatial resolution. The exact configuration of these instruments awaits necessary technical developments as described below. The facility will have requisite offices, common space, and conference facilities for staff, administration, and users.

The Cellular Systems Facility will provide to the user community:

- Frontier instruments incorporating the capabilities to create, sustain, and monitor structured microbial communities and analyze them at the molecular through ecosystem levels.
- Models and the tools and computing and information infrastructure for developing and evaluating such models, along with the user-facility infrastructure needed to undertake such tasks.
- The GTL Knowledgebase as a source of data on all known aspects of microbial systems that have been studied as a foundation for further experimentation, model development, and the ensuing simulations.

The facility therefore will be highly data intensive, providing extensive linked data sets on the dynamic behavior of microbial cells and communities. It also will be compute intensive, providing unprecedented data analysis, modeling, and simulation. In addition, it will involve new computational approaches for data storage, analysis, and use in complex models. These systems-level data sets will be made available to the entire scientific community. This new knowledge will be invaluable for advancing the annotation of microbial and community genome sequence, identifying regulatory and metabolic networks in microbial systems, and understanding microbial contributions to ecosystems function.

### 5.4.2.2. Performance and Production Targets

The Cellular Systems Facility will have the experimental and computational capacities to measure and analyze hundreds of microbial systems per year at unprecedented levels of detail. A key distinguishing feature will be the sophistication and performance of its instrumentation and capabilities in computational modeling and simulation. Some key performance features as described in the following sections include:

- Performance of physical (i.e., structure) as well as functional (i.e., molecular profiling) mapping
- Spatial and temporal resolution to map all levels of the functional processes within a microbial community (i.e., nanometers to millimeters and microseconds to hours)
- Performance of nondestructive measurements
- Culturing capabilities for maintenance of realistic environmental conditions and microbial communities



- Integration of modeling capabilities and supporting computing infrastructure, including data-intensive bioinformatics, compute-intensive molecular modeling, and complexity-dominated systems modeling

This GTL facility, more than the others, must overcome major challenges associated with the lack of available technologies and instruments for measuring the dynamic state of living microbial cells. As stated above, it will benefit from technologies, instrumentation, and data developed by current and future GTL R&D and pilot projects by the time the facility comes online. Along with state-of-the-art capabilities when operations begin, the facility also will include extensive ongoing development of essential instruments and technologies to advance the measurement of activities and characteristics of cellular systems at the single-cell level.

DOE has an extensive and successful history of developing and applying new technologies to complex problems in the physical and chemical sciences. As in the genome projects, the agency can draw on multidisciplinary teams of biological, physical, computational, and other scientists and engineers from national laboratories, academia, and industry. GTL brings a tremendous opportunity for using these same talents in the Cellular Systems Facility to devise technological solutions to some of biology's most complex problems.

### 5.4.3. Technology Development for Cultivation of Microbial Communities

The classic definition of an unculturable microbe is that it cannot be grown in homogeneous suspension. Because of this, there is a dearth of information about the metabolic capacity of microorganisms that resist cultivation under laboratory conditions (Keller and Zengler 2004). This is due primarily to the difficulty or impossibility of simulating the chemistry and interspecies interactions of highly structured communities by suspended cell-culture techniques. Most microbes reside in these structured communities and display unique phenotypic states in response to microenvironments within those communities. Many current technologies require large populations of cells to measure gene expression, proteome, and metabolites, masking the true cell-to-cell heterogeneity. The inability to cultivate most structured communities formed by natural microbial populations limits our discovery of new genes, gene products, and resultant functionalities. New cultivation techniques must support the development of meaningful community structures, and the functions of community members must be measurable in that environment (see sidebar, Laboratory Cultivation Techniques to Simulate Natural Community Structure, p. 180).

The facility's cultivation systems will allow for the precise control and manipulation of environmental conditions and for the monitoring and culturing of microorganisms in meaningful community structures. Many microorganisms of scientific and biotechnological importance, including those sequenced by the Biological and Environmental Research Program, by GTL, and by the Joint Genome Institute's Community Sequencing Program, are relevant to various DOE missions. Some thrive in unusual or extreme environments or those in which gradients or temporal changes occur in physicochemical conditions.

Scientific investigations of these organisms and the communities in which they live thus require flexible, highly controlled, and instrumented systems that can provide a range of environmental conditions and sophisticated measurements. These conditions include monoculture or mixed cultures; nutrient status; extremes of pH, temperature, and salinity; exposure to contaminants and radiation; gas composition and pressure; light intensity; and the presence of solid phases. The ability to control and monitor the environment allows for rigorous investigations and interpretations of gene expression, regulation, and function at the level of individual cells, cell populations, and mixed communities. When required, cells of unusual or difficult-to-culture microorganisms will be cultured in sufficient quantities to provide protein for biophysical, structural, and functional analyses. In other cases, very small numbers of unusual or difficult-to-culture microorganisms might be studied using novel microscale approaches combined with specialized sensitive analyses of gene expression, proteomics, metabolism, and metabolite flux. Ultimately, for meaningful analysis of communities, measurements at the individual cell level will be essential. To select for study any cell in such a structured community, we need to be able to (1) remove it from its environment without inducing significant changes in the properties being measured or (2) conduct analyses of living cells in situ (i.e., without disrupting its environment or harming the cell).

## Laboratory Cultivation Techniques to Simulate Natural Community Structure

Microbes associating within a biofilm surface offer the opportunity for members of a discrete population and individual organisms representing different species to establish fixed spatial relationships over extended periods of time. Surfaces enable microorganisms to establish high cell densities in localized areas. For example, products of cell metabolism in a colony of one type of microorganism diffuse to adjacent surface areas, forming strong concentration gradients within the intercellular volume of a biofilm (Beyenal, Davis, and Lewandowski 2004; Beyenal et al. 2004).

To identify the function of genes preferentially expressed by specific populations in the structured community, new cultivation techniques are being developed that incorporate surfaces for microbial colonization and RNA extraction (Finelli et al. 2003). 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 (FISH) 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).

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. This fermentor allows the end user precise control of culture growth to produce high-quality samples. Future generations of fermentors will be more highly instrumented, possessing sophisticated imaging and other analytical devices developed to analyze interactions among cells in biofilms under an array of conditions.



Pacific Northwest National Laboratory

### 5.4.3.1. Requirement Examples

- Development of techniques to isolate single cells from a natural community and analyze them for the expression of targeted genes, proteins, and other products (e.g., using fluorescent tags produced from metagenome sequences).
- Evaluation of metabolic processes carried out by cells in structured microbial communities, using molecular tags for RNA, proteins, or other reported moieties to map individual cell populations within the community (see imaging discussion below).
- For unsequenced microbes, application of techniques such as intact biofilm polymerase chain reaction (IB-PCR) to construct 16S rRNA clone libraries once a community is formed. The rRNA sequences of each population present could be used to establish phylogenetic links to other known populations (Reardon et al. in press). Techniques such as FISH then could employ 16S rRNA sequence data to construct oligonucleotide probes to locate different populations within the biofilm and identify putative associations between colocalized populations.
- For many systems, capabilities to assay the distribution of properties among a population of extracted cells from structured communities, sediments, or soils. Providing invaluable information, flow cytometry will be a high-throughput method of choice. Novel cultivation approaches also might be combined with single-cell sorting techniques such as emerging microfluidic lab-on-a-chip devices to grow and study currently uncultivable members of microbial communities.

#### 5.4.4. Development of Genomic Capabilities

To test hypotheses about function, genetic manipulation to generate mutants and specific constructs containing tags or reporter molecules is an essential requirement for systems biology research. Highly roboticized capabilities will be essential for high-throughput construction and screening at the genome level. Examples of required basic capabilities include nucleic acid isolation and analysis, sequencing and annotation, expression analysis, gene cloning and expression, fusion tagging of genes (Gaietta et al. 2002), general tools to manipulate members genetically, and cell sorting. Many of these capabilities will be present in other GTL facilities, the JGI, and in researchers' laboratories and may be incorporated into this facility as needed. Molecular microbiology and, in fact, all microbiology support capabilities will require a highly developed system for information management and integration.

#### 5.4.5. Technology Development for Imaging and Spectroscopy

The Cellular Systems Facility will employ a broad range of imaging modalities to monitor the structure of microbial communities and image (spatially and temporally resolving) the many molecular species critical to community function. The imaging capabilities of the Cellular Systems Facility and the Molecular Machines Facility are complementary—Molecular Machines images intracomplex structure and cellular location, and Cellular Systems focuses on spatially and temporally mapping multiple processes through the lifecycle of a community of cells in a complex environment. Imaging modalities available for both facilities are presented in Table 1. Characteristics of Available Imaging Modalities, p. 182; Table 2. Attributes of Available Techniques for Cellular Systems Characterization, p. 183; and Fig. 1, p. 176, all of which list the primary probe methods available, techniques based on them, analytical characteristics, prospects for further development, and computing requirements. To be applicable, these methods must be chemically specific, perform measurements nondestructively, and be capable of functioning in concert with other techniques (Toner et al. 2005).

##### 5.4.5.1. Analytical Characterization of Cellular Systems

Critical to addressing key scientific questions will be the novel application of existing and emerging technologies that characterize systems in a continuous and spatially resolved way. Analyses now conducted on bulk samples must transition to nondestructive processes capable of characterizing systems ranging from a microbial community through multiple processes within a single living cell. Also, the power to view multiple systems with high spatial and temporal resolution must be augmented with the ability to identify, track, and manipulate living microbes in the presence of other strongly interacting species. To achieve this, many classic imaging techniques must be coupled with methods that can detect specific molecules or processes. Physical, chemical, and biological variables must be identified and tracked. Furthermore, the power to observe systems in action will need to be enhanced by the ability to interact with these systems.

Developing the capability to view biological systems in great detail will enable new high-throughput approaches to studying cellular systems. Each cell in a culture, consortium, or community presents a unique reflection of the biological response to the overall system's changing state. Each provides a set of multilevel outputs in response to the effects of changing parameters (e.g., environmental insults, nutrient gradients, and temperature). To the extent that this parallel data stream can be captured in real time, biological experiments can be conducted in a high-throughput manner rather than running as several series of experiments to evaluate each possible response (e.g., cell division, movement, and protein shedding) for each type of environmental change.

To enable these advances, new technologies must address the special requirements for observing biological systems. Ideally, techniques will be nondestructive, noninterfering, and compatible with the analysis of heterogeneous, living systems. They will need to document the state of each cell (or many cells or cell types) as time and environmental conditions change. Furthermore, physical and chemical information must be mapped onto community structure while detailing changes at the molecular scale. These analyses will necessitate the development of new software to provide intelligent processing of data. Ultimately, such tools will

# FACILITIES

**Table 1. Characteristics of Available Imaging Modalities**

	Technique	Unique Characteristics	Future Prospects	Bioinformatics
<b>Visible Light</b> (possible: 50 nm practical: 300 nm)	TIR Absorbance Scattering NLO Adaptive optics FRET Structured light illumination	Noninvasive In situ Wide range of time + length scales Functional analysis Coordinated release of caged molecules Microsurgery, microablation Characterization of individual cells and communities (biofilms)	Better probes, lanthanite dyes, quantum dots, nanoparticles, tetracycline tags, genetically encoded nanoparticle sensors More versatile excitation sources Better detectors	3D visualization (online, offsite) Pattern recognition (spatial, spectral) Multiscale, multimethod data fusion
<b>X Ray</b> (20 nm)	Tomography Spectroscopy Microprobes	Thick, hydrated samples Whole cells Clean spectrum Organic functional group metal redox spectroscopy Molecular localization in ultrastructural context Characterization interactions	More versatile excitation sources Better detectors	
<b>EM</b> (0.3 nm)	Tomography Molecular microscopy: Single particle Cryo	Whole cells or sections High-resolution molecular localizations in ultrastructural context Correlation with fluorescence	More versatile excitation sources Better detectors	
<b>Force Imaging</b>	AFM tapping	Cell wall imaging Imaging of protein, nucleic-acid components	Better tips (higher-aspect ratio: Carbon nanotubes)	
<b>Force</b> (manipulation, perturbation)	Optical tweezers Magnetic tweezers	Mechanical characteristics (cell wall) Thermodynamics and kinetics of transient interactions Characterization of the molecular-machine mechanochemistry Correlated mechanical properties	Combined single-molecule fluorescence, optical tweezers	

help elucidate the large-scale biochemical organization that characterizes community structure. Such new analytical approaches will be essential for assessing the community's physiological and phylogenetic makeup and for testing predictions derived from theoretical models.

A number of these scientific needs will require fundamental new developments in imaging technology—a transformational goal for GTL biology. Revolutionary advances will be essential for determining the dynamics of communities and their functions under various environmental conditions, defining the physical structure of cells and communities, detecting and tracking extracellular and intercellular molecules to define cell states, and, ultimately, understanding how molecular events are communicated in space and time.

**Table 2. Attributes of Available Techniques for Cellular Systems Characterization**

Scale of Analysis	Information Needed	Techniques for Structure and Imaging	Static Characterization Techniques	Dynamic Characterization Techniques
<b>Proteins</b>	Components Abundance Structure	X-ray crystallography (angstrom) Raman spectroscopy (angstrom) Neutron crystallography (angstrom) X-ray spectroscopy (sub-angstrom) Electron microscopy (SEM, TEM, STEM, tomography) Electron crystallography	Infrared spectroscopy Raman spectroscopy NMR (nuclear magnetic resonance) spectroscopy Microsampling Microfluidics Fluorescence Scattering	Infrared spectroscopy Raman spectroscopy NMR spectroscopy Microsampling Microfluidics Fluorescence Scattering
<b>Molecular Machines</b>	Components, active sites Function, role, interchangeability, stressed behavior	X-ray crystallography, Raman spectroscopy, neutron scattering, X-ray scattering, EM, multi- and hyperspectral fluorescence	Infrared spectroscopy Raman spectroscopy NMR spectroscopy Microsampling Sensors	Pump-probe spectroscopy Microsampling Sensors Labels (quantum dots, organic fluorescence) Laue X-ray crystallography
<b>Cellular</b>	Components Active sites Function role Interchangeability Communication Stressed behavior	X-ray microscopy Scanning probes Scanning probe microscopy (SPM) Atomic force microscopy (1.0 nm) Scanning near-field optical microscopy (NSOM or SNOM) Scanning tunneling microscopy Chemical force microscopy Electrostatic force microscopy Magnetic force microscopy Electron microscopy (SEM, TEM, STEM, tomography) Far-field vibrational imaging (>10 microns) Optical microscopy	Mass spectrometry NMR spectroscopy Probes Raman spectroscopy Neutron spectroscopy Infrared spectroscopy SPM PH meter Microsampling Sensors Multi- and hyperspectral fluorescence Optical microscopy (one or multiphoton, scanning optical tomography; 200 nm in conventional mode; 5 nm in FRET/FLIM modes, FISH, CARS, SHM)	Raman spectroscopy X-ray microscopy Scanning probes Mass spectrometry NMR spectroscopy Probe spectroscopy Infrared spectroscopy SPM Microsampling Sensors Fluorescence Labels
<b>Communities</b> In lab In field	Components, active site, function, role, activators, interchangeability, stressed behavior How to communicate?	Far-field vibrational imaging (>10 microns) Optical microscopy (one or multiphoton, scanning optical tomography) NMR imaging Light-scattering spectroscopy Ultrasound	Infrared spectroscopy Raman spectroscopy NMR spectroscopy Microsampling Sensors	Pump-probe spectroscopy Microsampling Stop-flow chromatography Sensors Labels



## 5.4.5.1.1. Examples of Analytical Requirements

**Intracellular Structure.** Intracellular protein, RNA, and metabolite localization and kinetics of localization.

- Proteomics on replicate communities.
- Fine-scale cell ultrastructure.
- New multimodal capabilities for dynamic imaging of targeted intracellular molecules and their interactions (including machines) in individual cells and cell assemblies [e.g., with antibody labeling and electron microscopy (EM)].

**Community Structure.** Analytical instrumentation and techniques for determining overall community structure and identifying and characterizing spatial and temporal variations in metabolites, signaling and regulatory molecules, and the physicochemical environment within communities (see discussions under the Molecular Machines Facility, beginning on p. 143).

- Probes for the in situ measurement of extracellular metabolites in real time.
- Imaging and spectroscopy of population structure, gene expression, and metabolites in cell aggregates and subpopulations within communities.
- Characterization of cells in mixed communities by multispectral imaging of key cellular chromophores, possibly moving to on-the-fly cell-sorting platforms.
- Measurement of elemental distribution, oxidation state of elements, and biomolecules within and among communities.
- Quantitative imaging of metabolite (and signaling molecule) flux between cells in close proximity to or in contact with each other—one of the most critical needs for understanding how microbial communities function.
  - Analysis and assessment of the makeup and role of extracellular polymers in community structure, function, and stability.
  - Detection and frequency of genetic exchange, recombination, and evolution within communities.
  - Determination of macroscopic transport of water, solutes, and macromolecules and their relationship to microbial function.
- Characterization of interface physical and chemical properties.

### Identified Development Needs

- Advanced chemical and biological probes, including engineered microorganisms, tagged biomolecules, and chemical sentinels that will help characterize microbial communities.
- Advanced tools for imaging characterization for use in the laboratory as well as in the field.

A variety of imaging and microspectroscopic techniques are emerging to meet these challenges. In general, imaging relates spatially dependent information. Characterization of additional dimensions, however, will be essential for relating system activity. Some commonly used imaging techniques include:

- **Short-Wavelength Techniques.** Analyses with electrons and X rays typically provide the highest spatial resolution. Although commonly associated with ultrastructural analyses, short-wavelength techniques are being extended for analyses of whole cells at atmospheric pressures. Additionally, X rays are useful for mapping trace metals, while spectroscopic measurements can provide chemical identification.
- **Optical Microscopies.** The current standard for live-cell imaging, these tools are ideal for studying dynamics across a broad range of time scales and are sensitive down to the single-molecule level. A number of physical scales can be assessed, and emerging techniques and new labels are improving the sensitivity and resolution of optical microscopy.

- **Long-Wavelength Techniques.** A variety of these procedures including vibration, magnetic resonance, and terahertz-based imaging can provide essential information on chemical structure, identity, and spatial arrangement. For example, vibrational signatures are molecularly specific and can produce direct chemical information without additional labels.
- **Other Techniques.** A broad range of unconventional imaging approaches are making an impact on biological studies. Most notable, the family of instruments comprising scanning probe microscopy enables molecular-scale resolution; and chemical, electrical, and physical properties can be measured simultaneously. Emerging tools based on optical and magnetic trapping are allowing measurement of mechanical properties while micro- and nanoscale structures permit sensing of chemical, physical, and biological attributes.

Clearly, many current imaging and microspectroscopic techniques possess significant attributes and provide information relevant to the study of biological systems. Significant advances still are needed to adapt many of these tools to the characterization of microbial cellular systems much smaller than eukaryotic cells. Advanced instrumentation, improved biocompatibility, new approaches for targeting and delivery of tags, and improved labels are but a few of the significant challenges that face imaging technologies. More significant, no technique alone can provide the broad range of information needed to understand community structure and system function. A combination of methods will be essential to extend the depth of information required.

## 5.4.5.1.2. Monitoring and Interacting with Cellular Systems

To enable effective systems-level studies, the ability to monitor systems in action must be enhanced with selective construction, manipulation, and interaction with the system. Only then can efficient experimental evaluations and effective iterations be achieved with pursuits in theory, modeling, and simulation. This integration will be a culminating product of the facility and an essential tool for studying microbes, consortia, and microbial communities.

Advanced cultivation systems that allow for precise control, manipulation, and monitoring of environmental conditions must be compatible with advanced imaging technologies. Chemical gradients will need to be controlled and monitored precisely while temporally measuring molecular-scale properties. Genetically defined organisms must be carefully arranged into ordered microbial communities, perhaps through molecular-scale patterning techniques resulting from nanotechnologies. Such highly defined systems will require integration with sensing capabilities and the ability to activate biomolecular networks remotely. The capacity for simultaneously imaging and specifically targeting reagent release or activation, as currently used in biomedical applications, is within reach for GTL systems biology studies. The creation of such compound, multifunctional instruments will enable the collection of information needed to understand and exploit complex biological systems.

## 5.4.5.1.3. Technology Development Progress and Benefits

### 5.4.5.1.3.1. Advanced Optical Methods – Laser or Synchrotron Based

- Optical spectroscopic methods can be used as tools for noninvasive characterization and monitoring of dynamic behavior.
- Measurements of absorption and in vivo fluorescence can be used to monitor the presence and relative concentration of optically active biochemical species.
- Light-scattering spectroscopy can probe the size distribution of community structures.
- Vibrational (infrared and Raman) spectroscopy is a technique for studying the composition of biological materials without perturbing or labeling the sample. Biological components (e.g., lipids, proteins, nucleic acids, and carbohydrates) and biofilm and microbial surfaces (e.g., minerals and polymers) have unique vibrational spectra based on their chemical structures.

## FACILITIES

- Use of these methods will provide new information on the following:
  - Large-scale (1- to 10-micron) biochemical organization.
  - Composition and distribution of extracellular polymer matrices.
  - Concentration and distribution of nutrients, metabolites, signaling molecules, and other macromolecules.
  - Interactions of biofilms and microbial communities with supporting surfaces.

Because vibrational spectromicroscopy is noninvasive, it can be performed on dynamic living systems in combination with other techniques. If synchrotron radiation is used as the photon source, a dynamic system can be studied directly on surfaces of geological materials (see Fig. 4.1, p. 27, Report on Imaging Workshop 2002).

Significant progress already has been made using confocal and two-photon fluorescence microscopy. The specificity of these techniques is provided by the exogenous chromophore targeted through an affinity reagent or fusion tag to a particular protein. The resolution is on the order of a micron and slightly higher for two-photon than for confocal microscopy. Delivering chromophores to remote regions within a community or cell is a particular challenge. Additionally, the identification of probes that maintain activity in diverse environments is required (see Fig. 4.2, p. 28, Report on Imaging Workshop 2002).

All these techniques can be used in an imaging arrangement to monitor changes in community behavior in real time. Improvements are needed in such areas as spatial resolution, the ability to provide quantitative information, and data-acquisition speed. Additionally, advanced light-microscopy techniques can be developed for high-resolution 2D and 3D mapping. Often with specificity to particular components associated with imaging, these techniques include surface-plasmon resonance, surface-enhanced Raman spectroscopy, imaging of second-harmonic generation, optical-coherence tomography, and coherent anti-Stokes Raman scattering.

### 5.4.5.2. Imaging Macromolecular Complexes

Many types of imaging technologies can be employed to identify and spatially and temporally localize macromolecular complexes and their interactions within a dynamic community environment. Some specialized techniques have specific applications to the analysis of macromolecular complexes in situ in live, fixed, or frozen cells or ex situ. The strengths of imaging techniques typically include detection sensitivity and the ability to identify complexes in cells. Imaging techniques are applicable to all classes of complexes. In many cases, however, the identities of one or more components of the complex must be known to prepare tagged probes for imaging analysis. This requirement limits the application of imaging to full identification of protein complexes. Currently, most imaging techniques are relatively slow; automation, however, is providing faster sample throughput, and improved computational tools are enhancing data acquisition and analysis. Imaging techniques relevant to identification and characterization of protein complexes are summarized below, with additional information on other imaging tools in Table 2, p. 183.

**Tagged Localization.** Used with visible, X-ray, or electron microscopies to identify sets of biomolecules labeled with tags. An in situ method applicable to live (visible light), fixed, or frozen cells, it also is applicable to tagged transient complexes and membrane-associated complexes. A limitation is that the complex must be labeled with a tag, requiring tag synthesis and introduction into cells. Spatial resolution in these modalities comes from the instrument response function of the exciting source (i.e., the exciting beam provides the resolution). More developed X-ray optics, more versatile excitation sources, and improved probes are needed. Lanthanide dyes, quantum dots, nanoparticles, tetracysteine-based ligands, and other probes are examples of some recently reported probes used with various imaging modalities.

**Fluorescence Resonance Energy Transfer (FRET).** Used to identify pairs of biomolecules labeled with tags as well as to provide information on biomolecule relationships. This in situ method is applicable to live cells, tagged transient species, and membrane-associated complexes. FRET is particularly good for structure and binding of extracellular ligands. Like other imaging techniques, it requires tag synthesis and introduction into cells.

**Scanning Probe Microscopy.** Identifies protein associations by scanning with a specific molecule attached to the tip, including transient molecules. The technique is capable of very high spatial resolution, depending on the length of probe time, and of single-molecule detection. It is most suited for the study of membrane-associated complexes with whole cells or for the study of isolated complexes. Like other imaging techniques, it requires that the identity of one component of the complex be known so a molecule can be attached to the tip as the probe molecule. The probe, for example, then can be used to identify interaction sites on a cell surface. The technique is labor intensive and slow. Identification is a one-at-a-time process unless multiprobe devices with individual probe molecules are employed. These multiprobe devices are under development to allow technique application in a highly parallel fashion. Computer modeling of protein folds would enhance data interpretation, and improved computation is needed for data visualization and manipulation.

### 5.4.5.3. Development Options

As previously mentioned, many techniques required for the facility have yet to be developed sufficiently to analyze microbes of less than one micron in complex and changing communities. Many potential options must be explored over the next few years to determine probe and detection modalities capable of providing necessary information under these demanding conditions. Options that may be explored regarding available techniques, their range of applicability, and information they might provide are shown in Table 2, p. 183, and Table 3. Cellular Systems Facility Technology Development Roadmap, p. 188. The bulk of intracomplex characterization of molecular machines will be carried out in the Molecular Machines Facility. The sidebar, The Super Imager, this page, details creation of super imagers comprising compound, multifunctional instruments that individually would include many of the capabilities listed. Many of these development issues are summarized in 6.0. Development Summary: Global, Crosscutting, and Long-Lead Issues, p. 191.

### 5.4.6. Development of Computing Capabilities

Computational tools and infrastructure are required for efficiently collecting, analyzing, visualizing, and integrating large data sets to elucidate gene function and to model and simulate regulatory and metabolic networks, cells, communities, and ecosystems). These tools will support the development and validation of

## 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 infra-red 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 of the system's physical parameters with micron-scale resolution
- Atomic force microscopy (AFM) mapping of structures with added information provided by the controlled-interaction light with 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

[Source: *Report on the Imaging Workshop for the Genomes to Life Program April 16–18, 2002* (Office of Science, U.S. Department of Energy, Nov. 2002); [www.doe-genomestolife.org/technology/imaging/workshop2002/](http://www.doe-genomestolife.org/technology/imaging/workshop2002/)]

**Table 3a. Cellular Systems Facility Technology Development Roadmap**

Technology Objectives	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
<b>Technologies for Cultivation of Microbial Communities</b>  Precise control, manipulation, and monitoring of environmental conditions; interrogation  Functional individual microbial cells in the context of characterized physiochemical environment  Support for formation of structured communities	Requirements defined for analyzing individual cells within structured communities:  Mixed microbial cultures <ul style="list-style-type: none"> <li>• Suspended and structured</li> <li>• Biofilms</li> </ul> Methods and approach to identify and track microbes and molecular complexes <ul style="list-style-type: none"> <li>• Tagged probes               <ul style="list-style-type: none"> <li>» Increased variety of signals</li> <li>» Signal interpretation</li> <li>» Incorporation in cell</li> </ul> </li> <li>• Arrays</li> </ul>	Multiple flexible experimental systems to control and manipulate growth and conditions with multiplex measurements of activity including: <ul style="list-style-type: none"> <li>• Chemostats</li> <li>• Microtechnologies</li> <li>• Remote sensing</li> <li>• Imaging</li> <li>• Surfaces to nucleate biofilms and other structures</li> </ul> Multiple probes to identify community members  Temporal monitoring of community structure and function	Integration of culturing capabilities within multiprobe instrumentation for simultaneous control, manipulation, and multimodal analyses of structured communities: <ul style="list-style-type: none"> <li>• Nondestructive</li> <li>• Real time</li> <li>• Linked databases</li> <li>• Environmental, community, cellular, and molecular levels</li> </ul>	Integrated, highly characterized, and real-time manipulatable structured microbial communities that simulate natural communities and niches: <ul style="list-style-type: none"> <li>• Protocols</li> <li>• Extracted samples</li> <li>• Characterizations</li> <li>• Analytical images</li> </ul>
<b>Environmental Communities Sampling</b>  In situ measurements	Lab techniques extended to field use	Planned extension after operations begin		
<b>High-Throughput Cultivation for Single-Cell Analysis</b>  Sampling techniques  Controlled viable growth of single cells	Analysis from within structured communities, in microculture extracts, or in place: <ul style="list-style-type: none"> <li>• Cell sorters</li> <li>• Lab on a chip and microfluidics</li> <li>• Single-cell analysis of “unculturable” environmental samples</li> </ul>	Assessment of compatibility with analytical instrumentation and simulation fidelity of natural environments	High-throughput operational mode combining culturing techniques interfaced with multimodal, analytical, and manipulation modalities	Single cells prepared in conditions that simulate microniche environments in highly structured microbial communities such as biofilms (formerly unculturable)

(continued next page)

theories and models of community growth, function, and environmental response. New theory, algorithms, and implementation on high-performance computer architectures also are needed for modeling and simulating cellular systems. Enabling a broad range of biologists to access the large data sets and computational resources for discovery-based biology will require the development of web- and grid-based technologies (see 4.0. Creating an Integrated Computational Environment for Biology, p. 81, and Table 4. Computing Roadmap, p. 190).



**Table 3b. Cellular Systems Facility Technology Development Roadmap**

Technology Objectives	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment	Facility Outputs
<p><b>Temporal and Spatial Localization of RNAs, Machines, and Metabolites</b></p> <p>Analytical measurement contexts:</p> <ul style="list-style-type: none"> <li>• Environmental</li> <li>• Community</li> <li>• Intercellular</li> <li>• Intracellular</li> </ul>	<p>Requirements defined for multimodal measurements:</p> <ul style="list-style-type: none"> <li>• Environmental physicochemical variables</li> <li>• Intercellular biomolecules</li> <li>• Intracellular biomolecules</li> <li>• Metabolites</li> <li>• Community overall biochemical and biophysical functionality</li> </ul> <p>Examples of needed instrumentation with molecular specificity, sensitivity, and spatial and temporal resolution:</p> <ul style="list-style-type: none"> <li>• NMR for community-scale microscopy (e.g., metabolites, signaling molecules)</li> <li>• Small molecules in living cells</li> <li>• Gene expression in living cells</li> <li>• Proteins and machines in living cells, including dynamics and interactions</li> <li>• Biomolecular mapping microscopies [confocal, CryoEM, SPM (AFM, STM, others)]</li> <li>• Image-interpretation tools</li> <li>• Visualization</li> <li>• Computational systems</li> <li>• Databases</li> </ul>	<p>Modular analytical and imaging instrumentation and methods integrated with culturing, monitoring, control, and manipulation modalities to assess:</p> <ul style="list-style-type: none"> <li>• Viability of integrated approaches</li> <li>• Compatibility with living systems</li> <li>• Intermodal interactions</li> <li>• Ability to meaningfully assess single cells</li> <li>• Data integration</li> <li>• Simulation and modeling integrated into experimental methods</li> <li>• Visualization of multimodal analyses and system monitoring and manipulation</li> </ul>	<p>Integrated culturing capabilities within multiprobe instrumentation for simultaneous control, manipulation, and multimodal analyses of structured communities:</p> <ul style="list-style-type: none"> <li>• Nondestructive</li> <li>• Real time</li> <li>• Linked databases</li> </ul>	<p>Characterizations of microbial communities in realistic environments at the environmental, community, cellular, and molecular levels:</p> <ul style="list-style-type: none"> <li>• Spatial</li> <li>• Temporal</li> <li>• Functional</li> <li>• Process</li> <li>• Molecular</li> </ul> <p>Databases and query tools</p> <p>Protocols</p> <p>QA/QC</p>

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.

**Table 4. Computing Roadmap: Facility for Analysis and Modeling of Cellular Systems**

Topic	Research, Design, and Development	Demonstration: Pilots and Modular Deployment	Integration and Production Deployment
<b>LIMS and Workflow Management</b>  Participate in GTL cross-facility LIMS working group	Available LIMS technologies Process description for LIMS system Crosscutting research into global workflow management systems Approaches to guiding experiment-based production protocols to inform how best to produce a protein as an AI system helps develop strategy for production	Prototype cellular systems LIMS system* Characterization design strategy system Workflow-management system for identification and characterization Process simulation for facility workflow	Cellular systems LIMS and workflow system Workflow integrated with other GTL facilities and experimental strategy systems
<b>Data Capture and Archiving</b>  Participate in GTL cross-facility working group for data representation and standards	Data-type models* Technologies for large-scale storage and retrieval Preliminary designs for databases	Prototype storage archives Prototype user-access environments	Archives for key large-scale data types* Archives linked to community databases and other GTL data resources GTL Knowledgebase feedback
<b>Data Analysis and Reduction</b>  Participate in GTL cross-facility working group for data analysis and reduction	Algorithmic methods for various modalities* Grid and high-performance algorithm codes Design for tools library Approaches for automated image interpretation in confocal light microscopy/FRET	Prototype visualization methods and characterization tools library* Prototype grid for data analysis, with partners Prototypes for automated image interpretation in confocal light microscopy and FRET Analysis tools linked to data archives	Production-analysis pipeline for various modalities* on grid and HP platforms Large-scale experimental data results linked to genome data Automated image interpretation in confocal light microscopy and FRET Repository for production-analysis codes Analysis tools pipeline linked to end-user problem-solving environments
<b>Modeling and Simulation</b>  Participate in GTL cross-facility working group for modeling and simulation	Existing technologies explored for cell-system modeling and simulation Research methods for reconstruction of protein interaction, regulatory networks, metabolic pathways, and community interactions Mathematical methods for multiscale stochastic and differential-equation network models	Experimentally guided metabolic reconstruction Signaling and regulatory-network reconstruction and simulation Efficient modeling methods for community-interaction networks Mature methods for reconstructing protein-interaction and regulatory networks	Production pipeline and end-user interfaces for cellular and community-level combined network reconstruction and simulation Production codes for image time-series analysis
<b>Community Data Resource</b>  Participate in GTL cross-facility working group for serving community data	Data-modeling representations and design for databases: In vivo protein expression and localization, cell models and simulations, community models and simulations, cellular and community methods and protocols	Prototype database End-user query and visualization environments Integration of databases with other GTL resources	Production databases and mature end-user environments Integration with other GTL resources and community protein-data resources
<b>Computing Infrastructure</b>  Participate in GTL crosscutting working group for computing infrastructure	Analysis, storage, and networking requirements for cellular systems data Grid and high-performance approaches for large-scale data analysis for cellular and community networks and simulations and to establish requirements	Hardware solutions for large-scale archival storage Networking requirements for large-scale grid-based MS and image data analysis	Production-scale computational analysis systems Web server network for data archives and workflow systems Servers for community data archive databases

\* Data types and modalities include MS, NMR, neutron scattering, X-ray, confocal microscopy, cryoEM, and process metadata. Large-scale experimental data results are linked with genome data, and feedback is provided to GTL Knowledgebase.