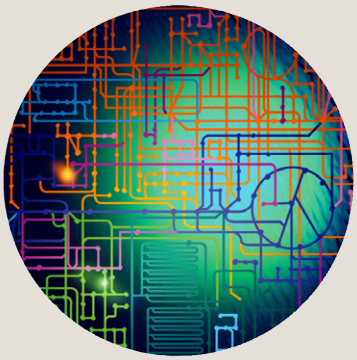
-

6.0. GTL Development Summary: Global, Crosscutting, and Long-Lead Issues



GTL Development Summary: Global, Crosscutting, and Long-Lead Issues

6.1. Coordinated GTL Program and Facility Development

In the preparation of this roadmap, numerous crosscutting and long lead time issues have emerged that require a long-range perspective as well as a globally coordinated approach to technology development (see 1.3.6.2. Integrated Management and Development, p. 11). These include scientific, technological, computing, and governance issues across the GTL program and facilities. Key topics, drivers, and issues include the following:

6.2. Biology Drivers and Issues

Some technical advances needed in the biology underlying GTL production and analytical technologies are discussed below.

6.2.1. Recalcitrant Proteins and Complexes

Drivers: Production and analysis methods are needed for membrane, secreted, transient, and large proteins; proteins having unusual cofactors; those normally found in complexes; and others.

Issues: What should we know about these proteins and complexes? What are the most promising technologies for production and analysis?

- For all proteins, the relative efficacy of in vivo vs in vitro technologies must be determined.

6.2.2. Biosample Growth and Culturing

Drivers: High-fidelity measurements demand well-defined biosamples.

Issues: New experimentation will require intricate design and control, including simulation and modeling and real-time environmental and biological characterization.

- Techniques must encompass complex biology such as biofilms, mixed cultures, environmental samples, and "unculturables."

Note: Long-lead items are those for which the acquisition time (including development and procurement) is longer than the time allotted for a given facility construction project. Crosscutting items are critical to more than one of the GTL facilities and elements.

6.2.3. Affinity Reagent Libraries

Drivers: Comprehensive coverage of affinity reagents is a key product of the Protein Production and Characterization Facility and a critical analysis tool for the whole GTL enterprise.

Issues: Current libraries of affinity reagents, such as single-chain variable domains of antibodies, are not appropriate for high-throughput analyses and downstream applications in GTL. Many possible molecular scaffolds could be used as foundations for reagent libraries, and the usefulness and limitations of these molecular scaffolds should be evaluated. More exploration of novel libraries and development of improved affinity reagent libraries also will be needed.

6.2.4. Characterization of Proteins and Complexes

Drivers: Subsequent utilization and analyses will require a minimal set of characterizations for each protein and complex.

Issues: Community involvement will help to generate specific requirements for the characterization of proteins and molecular machines. These include biophysical characterization such as measurements of size, shape, stoichiometry, and organization, as well as biochemical assays designed to screen for functional activity.

6.3. Technology Drivers and Issues

GTL is highly dependent upon and should stimulate next-generation analytical and imaging approaches for measuring the parameters of microbial molecular and cellular systems. Many of these technologies are undergoing rapid advancement, and some will impact multiple GTL facilities. Requirements for successful facility operations, technical challenges and gaps, and the roadmap for acquiring necessary technologies should be defined.

6.3.1. Technologies for Measurement of Proteins, Metabolites, and Molecular Machines

Drivers: GTL will set requirements for measuring the presence and absolute quantity of many types of molecules in microbial systems. Technologies that can accomplish these measurements should be explored and evaluated regarding their potential for achieving the prerequisite detectivity, sensitivity, accuracy, dynamic range, and throughput to support facility operations. Critical considerations in proteomics, machines, and microbial-system analytical protocols must be resolved based on science goals for the research programs and technology capabilities and limitations.

Issues: Investigators need small sample volumes, measurement reproducibility, sensitivity, dynamic range, high sample quality, low cost, and the achievement of high throughput with technologies that currently are largely manual.

- Challenges include quantitative methods for mass spectrometry, reproducibility, dynamic range, robustness, and global coverage for gene products and regulatory molecules; sample preparation for different molecular classes and cellular fractions; stoichiometry of molecular machine partners; and high-throughput operations.
- Other technologies including arrays also must be investigated.

6.3.2. MEMS, Microfluidics, and Nanotechnology

Drivers: Technology miniaturization has the potential to make huge impacts in protein characterizations and other experimentation by reducing the amount of material and reagents needed. The nature and scope of GTL facilities could be affected dramatically through broad application of such technologies. Micro- and nanotechnologies have the potential to provide new functionalities in detection, manipulation, and analysis of biological systems. These technologies include microfluidics and microelectromechanical systems (MEMS).

Issues: Concerted effort is required to define and develop the use of microtechnologies for all facets of GTL science.

6.3.3. Single-Cell Analysis

Drivers: Many key questions and challenges will require a single-cell capability, including population heterogeneity for proteomics, culture issues, sample minimization, validation testing for presence of machines, in situ analysis, analysis of cellular specialization in communities, and temporal and spatial resolution of cellular systems processes as the ultimate test of systems models.

Issues: Single-cell analysis has challenges in instrumentation, robustness, detection limits, dynamic range, and handling.

6.3.4. Imaging

Drivers: New imaging modalities are required for localization and validation of complexes, dynamics, docking, intercellular communication, extracellular matrix, and metabolite distribution.

As technologies mature to enable the examination of molecular machines in vivo, high-throughput and automated image-acquisition and analysis capabilities will be needed.

Issues: Technical challenges include consistent and benign label incorporation to preserve the functionality of proteins and machines and provide for data acquisition and interpretation, multimodal imaging, high-throughput considerations, chemical analyses, sensitivity, and spatial and time resolution.

6.3.5. Data-Quality Standards

Drivers: Standards are the critical foundation in experimental design, data capture and analysis, and informatics and computational approaches in a data-intensive environment.

Issues: Issues extend across all facilities and throughout the research programs—including data definition, integration, "error" expression, and reference points.

6.4. Computing, Communications, and Information Drivers and Issues

To be successful, GTL must coordinate the analysis of vast amounts of data, share metadata about experimental processes and workflow, manage the combined output of joint experiments, and provide common gateways for the user community to access the data, models, and simulations of microbial systems. The program also must provide for shared hardware, tools, and network infrastructure. This will require long lead times and a coordinated research and development approach.

Key computing, communication, and information needs include the following.

6.4.1. Computational Methods for Experimental Data Analysis

Drivers: Vast amounts of data will be produced by many different methods. These data must be analyzed quickly enough to keep up with data production, using algorithms sufficient to extract information needed by researchers.

Issues: Challenges include advancement in algorithm design to more accurately extract and quantitate observations from raw data (e.g., mass spectrometry, NMR, scattering, expression analysis). Other challenges are the development of methods for large-scale distribution and management of analysis processes such as computing grids; configurable analysis tool pipelines linked to integrated data resources; and environments for researchers to facilitate large-scale analysis processes.

6.4.2. Process Control, LIMS, Workflow Management

Drivers: Operations at the GTL facilities will be coordinated using an integrated workflow process. This workflow environment is significantly more complex than those used for sequencing facilities. Large-scale

experiments will require development of a shared experimental strategy, a uniform pipeline with strongly coupled measurements, sharing of process metadata, and electronic investigator collaboration and coordination.

Issues: The GTL facilities will require a laboratory integrated management system (LIMS), electronic notebooks; collaboratory environments; process optimization; dynamic process scheduling; and sample archiving, tracking, prioritization, and storage. These technologies must be developed in a consistent fashion across the facilities using common technologies and data standards.

6.4.3. Data Architecture, Modeling, and Integration

Drivers: High-throughput processes will generate massive amounts of information that must be shared across facilities and with experimental planners and the community. Collected knowledge of all facility experiments and measurements needs to be captured as endurable data. Reduced data incorporated into models of biological systems and simulations based on these models must be managed as part of the data structure. All aspects of the data infrastructure must be integrated across the program.

Issues: Global database development, the most complex and important element of GTL, needs a representative working group to define and model data types, explore data management and user access technologies, and establish working standards.

6.4.4. Computing Hardware and Networking Infrastructure

Drivers: Analysis, modeling, and simulation from large data sets will require computing and networking capabilities and capacities well beyond existing infrastructures and those proposed for the next generation of computing platforms. This information-intensive undertaking will need terascale communications; distributed (grid) approaches to capacity computing problems; and environments for petascale, numerically intensive, physics-based simulations.

Issues: New hardware architectures will require extensive acquisition lead time. A working group should specify performance specifications that drive new architectures and communication technologies.

6.4.5. Computational Models for Establishing Networks and Simulations

Drivers: Methods and mathematical approaches will be used for modeling, simulating, and visualizing complex types of cellular processes and interactions. Substantial enhancement and maturation in mathematical methods and theory are required before systems with realistic complexity can be considered.

Issues: Existing modeling has not dealt with levels of complexity and uncertainty or the magnitude of data with which we will be working. Stochastic effects of such systems should be better represented. Also, the robustness and sensitivity of system models to data errors or omissions need to be better understood. The preference is to use models to drive experimental designs.

6.4.6. Genome Annotation

Drivers: GTL research is based on the availability of fully finished and annotated (nondraft) genomes. Annotation is the foundation for experimental planning to coordinate activities of the four facilities and provide significant initial information about protein function and significance in each genome.

Issues: Many enhancements in annotation are needed, especially those related to recognition of interacting proteins, regulatory signals and structures, infrastructure for large-scale genome analysis processes, and databases for microbial genomes.

6.4.7. Computing, Communications, and Information

Working groups on the following topics should be established to examine issues of global value to GTL:

- Microbial genome annotation and data management
- LIMS and workflow management
- Data-analysis algorithms and large-scale processing
- Data infrastructure, data modeling, databases, and data standards
- Regulatory, metabolic, and cell modeling
- Hardware, grid, and networking infrastructure

6.5. Other Issues

6.5.1. Ethical, Legal, and Social Issues (ELSI)

GTL is largely a microbiology program, but it encompasses many scientific activities that might be expected to impact society in a number of ways. DOE is committed to stressing the close coordination of ELSI studies with the ongoing science.

6.5.2. Technology Transfer

Transfer is key to infusing new technologies into the facilities and to rapidly making innovative concepts and technologies commercially available for broader biological research and DOE mission applications.

6.5.3. Industrial Involvement

Industrial vendors and developers will be important in the rapid development of new instruments and modalities for facilities and their subsequent introduction to the broader community. Clear policies and procedures involving industry need to be developed and put in place to facilitate this process.