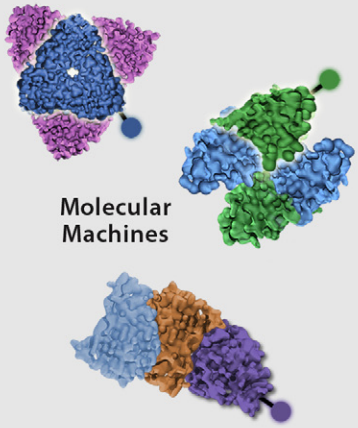
# 5.2. Facility for Characterization and Imaging of Molecular Machines



To accelerate GTL research in the key mission areas of energy, environment, and climate, the Department of Er




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.

## 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.**

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.

## 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

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.

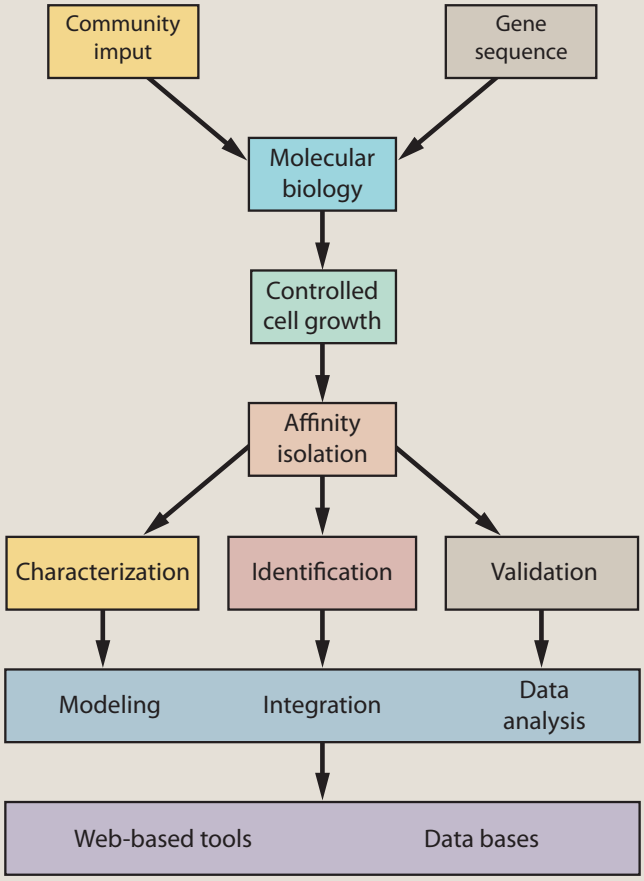



**Fig. 1. Facility for Molecular Machines Analysis Core Capabilities.**

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.

### 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 |
|---|---|---|---|---|
| **Develop technologies for protein machine production: Cell growth and processing**<br>Cultivation systems:<br>• Modified for endogenous protein isolations<br>• Wild type for exogenous complex isolations | Define conditions to express and process active molecular machines<br>Develop methods:<br>• Reproducible growth, real-time monitoring, sampling<br>• Novel culture approaches<br>• High-throughput controlled fermentations<br>• Sample archive documentation<br>• Functional assays for unknown isolated molecular machines<br>Evaluate commercial systems | Controlled cell growth, processing:<br>• Modified cultures for endogenous complex isolation<br>• Wild-type microbes for exogenous complex isolation<br>• Large numbers of microbial clones with encoded tags<br>Database development<br>Controlled bioreactors for cellular imaging<br>Automation and standardization<br>Standards, protocols, costs, QA/QC refinements<br>Evaluation, incorporation of new technologies<br>Development of methods for microbes and machines requiring specialized conditions | Establish high-throughput pipeline based on defined requirements, standards, protocols, costs<br>Scale up parallel processes for multiple organisms<br>Evaluate and incorporate new technologies<br>Use parallel processes for scaleup | Production of well-defined microbial samples for extraction and characterization of active molecular machines<br>Database from controlled cell growth with analysis of protein complexes and associated biocompounds<br>Well-managed biosample archive<br>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 |
|---|---|---|---|---|
| **Develop high-throughput technologies for molecular machine isolation for full population of biomolecular complexes**<br>Soluble stable complexes<br>Membrane-associated complexes<br>Transient complexes | Define needs:<br>• Evaluation of commercial laboratory and LIMS resources, if available<br>• Protocol refinement, automation<br>• QA/QC<br>• Multistage isolation schemes using affinity reagents to minimize background interferences<br>• Microfluidic-based affinity reagent isolations to minimize sample size requirements<br>• Stabilization and cross-linking of less-stable and transient complexes<br>• In vivo validation approaches<br>Develop:<br>• Continuous, automated processing<br>• Multiplexed pulldowns<br>• Novel affinity reagents and isolation schemes<br>Develop:<br>• Solubilization of membrane-associated complexes<br>• Stabilization of complexes<br>Develop stabilization and cross-linking | Pilot-scale isolation method:<br>• Scaleup from 100 assays per week to thousands per week<br>• Automated, continuous processing<br>• Assessment of bottlenecks, costs<br>• QA/QC<br>• Evaluation and incorporation of new technologies<br>• Methods for rapid elucidation of protein complex network linkage maps | Establishment of multiple parallel pipelines<br>Evaluation and incorporation of new technologies | Complexes isolated to permit identification, imaging, and biophysical characterization<br>Protocols<br>Methods<br>Databases and query tools |

**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.**

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 |
|---|---|---|---|---|
| **Develop technologies for complex identification and characterization**<br>Analysis by mass spectrometry (MS):<br>• Identification and quantification of both digested peptides and intact proteins<br>Data processing:<br>• Data interpretation<br>• Data archiving | Develop for mass spectrometry:<br>• High-throughput complex analysis by MS<br>• Improved data analysis<br>• Improved methods for quantitation and determination of complex stoichiometry<br>• Improved MS detection limits and dynamic range<br>• Identification of protein complex modifications via top-down MS<br>• Combined isolation and identification approaches<br>• Improved online separations<br>• Integrated, high-sensitivity analytical tools, eventually for single cells<br>• Improved cleavage and digestion approaches<br>• Improved ionization for broad classes of proteins<br>• Microfluidic-based assays<br>Evaluate commercial hardware, software, and instrumentation | Pilot scale:<br>• Optimization of protocols with regard to throughput, reproducibility, costs<br>• Improved MS data-analysis tools<br>• Database development and query tools<br>Assays:<br>• Integrated "lab on a chip"<br>• Probe-based affinity<br>• Binding affinity<br>• Automated neutron, cryoEM, and X-ray small-angle scattering<br>• New technologies evaluated and incorporated<br>• MS labeling for identification of contact interfaces | Establish high-throughput, automated pipelines:<br>• Scale up via multiple parallel production lines<br>Refine QA/QA protocols<br>Automate data acquisition and data analyses<br>Evaluate and incorporate new technologies | Capability for high-throughput protein complex analysis by MS<br>Highly validated data of identified protein complexes<br>Confirmatory analyses of protein complexes via biophysical techniques<br>New tools for complex-identification analysis<br>Databases for complex identification and characterization<br>Examples:<br>• Binding affinity<br>• Interaction interfaces |
| **Biophysical Characterization**<br>Structural characterization<br>Binding affinities<br>Others | Establish structural and functional assays:<br>• EM, SANS, SAXS, NMR<br>• Approaches to identify contact faces<br>• High-throughput binding affinity assay | | | |

**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

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<br>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<br>(5000 to 10,000 in zoom scan mode) | 0.1 Da | 200 to 4000 |
| Fourier transform ICR-MS | ESI<br>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

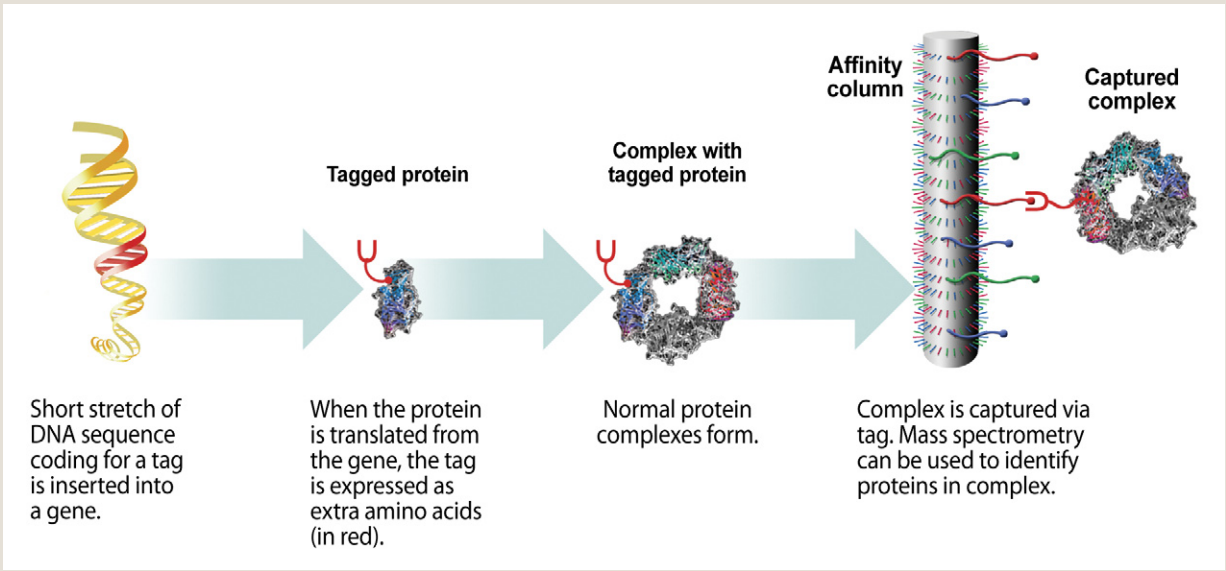



Fig. 2. Capturing Protein Complexes Using Fusion Tags.

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 |
|---|---|---|---|---|
| **Develop technologies for complex validation and characterization**<br><br>Analysis of complexes in vitro and in vivo:<br>• Data processing<br>• Data archiving | Develop:<br>• In vivo imaging for validation and spatial and temporal studies<br>• New labels for optical microscopy<br>• Multimodal imaging approaches<br>• Automated image acquisition<br>• High-throughput image analysis<br>• Improved spatial resolution<br>• Environmental sample-manipulation techniques<br><br>Evaluate commercial hardware, software, and instrumentation | High-throughput EM<br><br>High-throughput optical methods<br><br>Image-analysis software<br><br>Automated sample acquisition<br><br>Multimodal imaging | Automate image acquisition<br><br>Automate data analysis<br><br>Scale up acquisition and analysis<br><br>Establish database<br><br>Evaluate and incorporate new technologies | Data and characterizations:<br>• Existence of complexes<br>• Dynamic spatial relationships of proteins and other macromolecules in complexes<br>• Local chemical and physical environment of complexes in cells |

**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 μM 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

importance is chemical-shift mapping and H/D exchange techniques that can be used to observe dynamics. NMR requires isotopic labeling (e.g., $^{15}N$, $^{13}C$, $^{2}H$) 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.

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