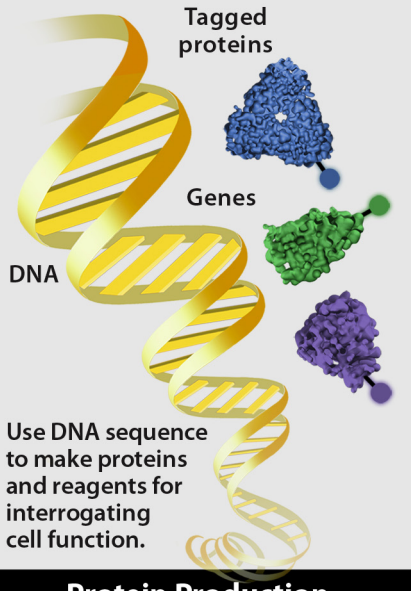
5.1. Facility for Production and Characterization of Proteins and Molecular Tags

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

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



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.

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**

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.

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

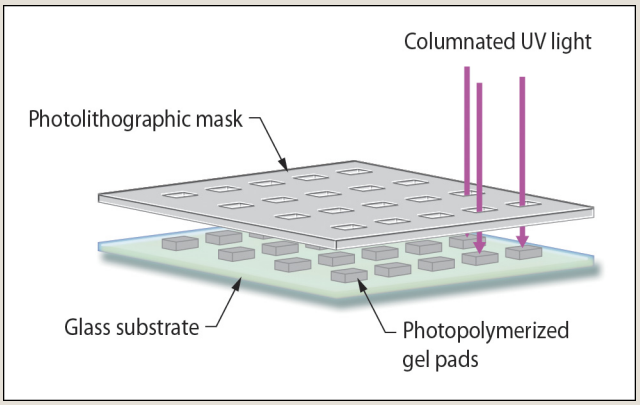
Automated high-throughput and high-quality biophysical and biochemical characterizations of proteins will provide a more rigorous assignment of gene function, resulting in first insights to a mechanistic understanding of microbial capabilities. For the first time, comprehensive reliable data on thousands 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).

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

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, FlAsH 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 10^5proteins/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

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	
	E. coli	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.

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
Protein Production 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
Protein Production 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
Protein Production 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
Protein Production 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.

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-µL reaction volumes, 30-µg 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 that 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

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)
Product Validation, QA/QC	
Protein production, identification 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
Biophysical and Biochemical Characterization	
Prepurification (See items below under postpurification) **Postpurification** 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 (FlE/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)
Ultimate Characterization	
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
Ultimate Manipulation	
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

- 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

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.

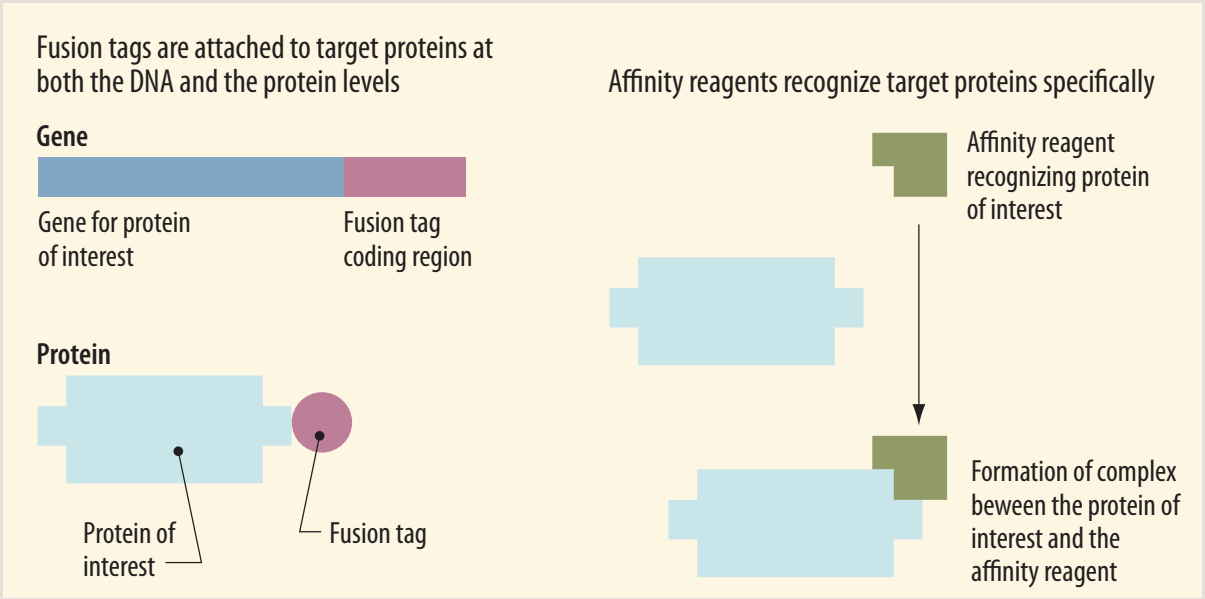
5.1.5. Development of Approaches for Affinity-Reagent Production

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.



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
Strengths	Good diversity Fusion proteins	Liquid and fluorescence-based screening Affinity maturation Fusion proteins	Good diversity Fusion proteins	Good diversity	Many secondary antibodies available
Weaknesses	Slower screening Plate based	Fluorescent tags required that may complicate recognition Reduced cys 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
Development Targets and Needs	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.

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

Objectives Subtopics	Research	Pilots	Production	Products
Affinity-Reagent Library Development	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
Affinity-Screen Automation	Develop protocols	Scale up to 2k/year Protocol standards QA, standards	Scale up to 25k/year	Affinity reagents Reagent chips Protocols QA, standards
Affinity-Reagent Target Design	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
Obtained by Animal Immunization	
IgG and IgM	Detection, purification
Obtained by in Vitro Methods (Affinity reagents based on antibody-like proteins)	
Fab	Detection, purification, therapeutics
FV	Detection, purfication, crystallization
scFV	Detection, purfication, in vivo perturbation, therapeutics
Domain antibodies (VH, VL)	Detection, purfication, therapeutics
VHH (shark and camel heavy-chain antibody VH domains)	Detection, purfication
Fibronectin type 3 domain	Detection, purfication, in vivo perturbation
Affinity Reagents Based on Other Proteins (Scaffolds)	
Affibody (protein A)	Detection, purfication
Anticalin (lipocalin)	Detection, purfication
Ankyrin repeats	Detection, purfication, in vivo perturbation
Thioredoxin	In vivo perturbation
Affinity Reagents Based on Other Molecules	
Conbinatorial peptides	Detection, crystallization, in vivo perturbation
RNA or DNA aptamers	Detect, purification, in vivo perturbation
Small chemical molecules	In vivo perturbation

Table 7. Examples of Fusion Tags and Their Applications

Examples	Applications
Peptide Tags	
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^{2+} dependent)
Elastin-like peptides	Purification (temperature-dependent aggregation)
Protein Tags	
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^6-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 10^9 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

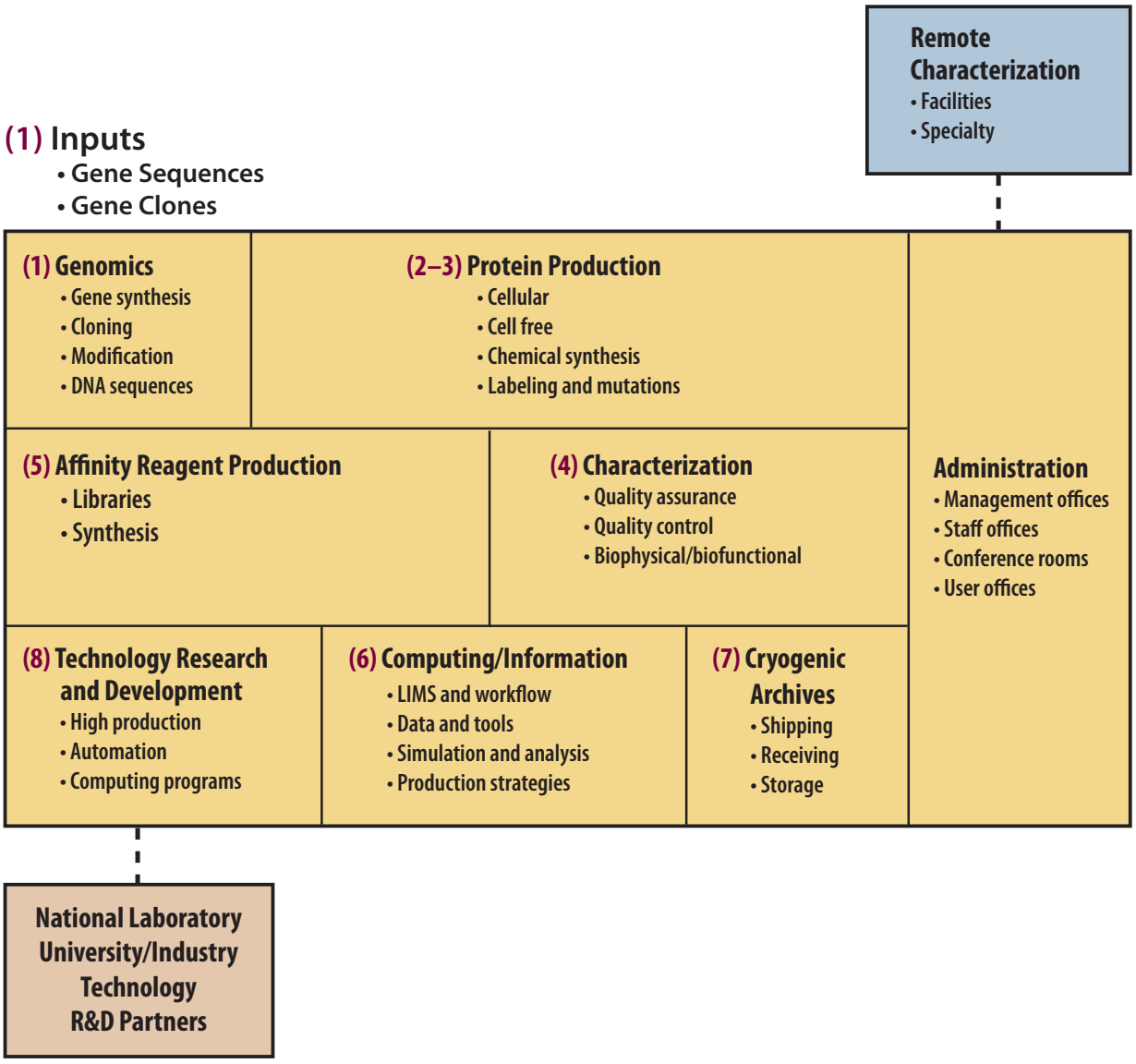
Each step and process in the Protein Production and Characterization Facility will involve very large numbers of biological samples that need to be tracked appropriately through the automated systems. Sophisticated bioinformatics analysis will be greatly needed at all steps so insights can be gained 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
LIMS and Workflow Management 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
Data Capture and Archiving 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
Data Analysis and Reduction 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
Modeling and Simulation 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 scatting data	Production pipeline and end-user interfaces for genome-scale fold prediction Production codes for scattering-data modeling
Community Data Resource 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
Computing Infrastructure 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



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.

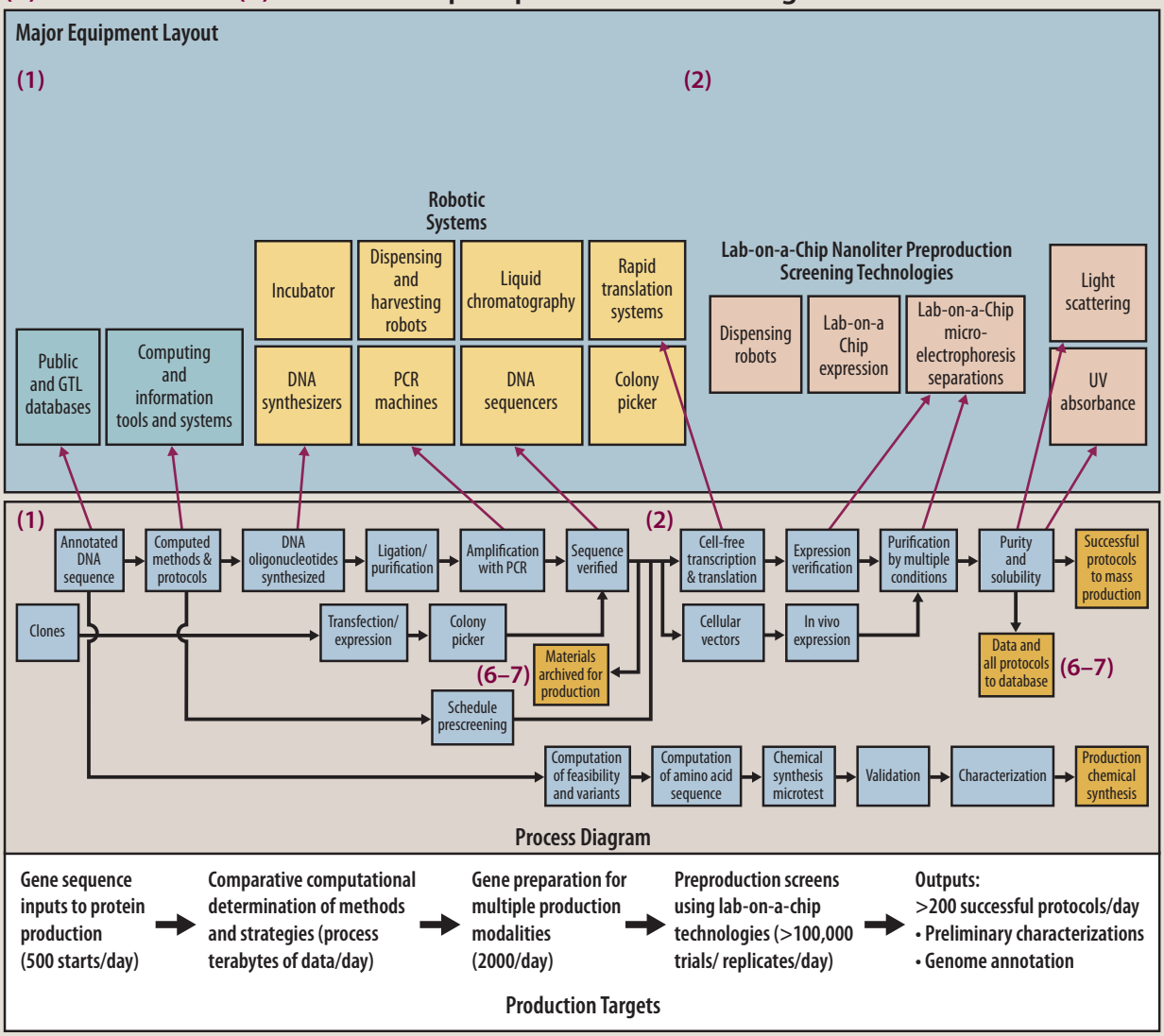
Inputs (1)

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 overexpress 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

(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*).

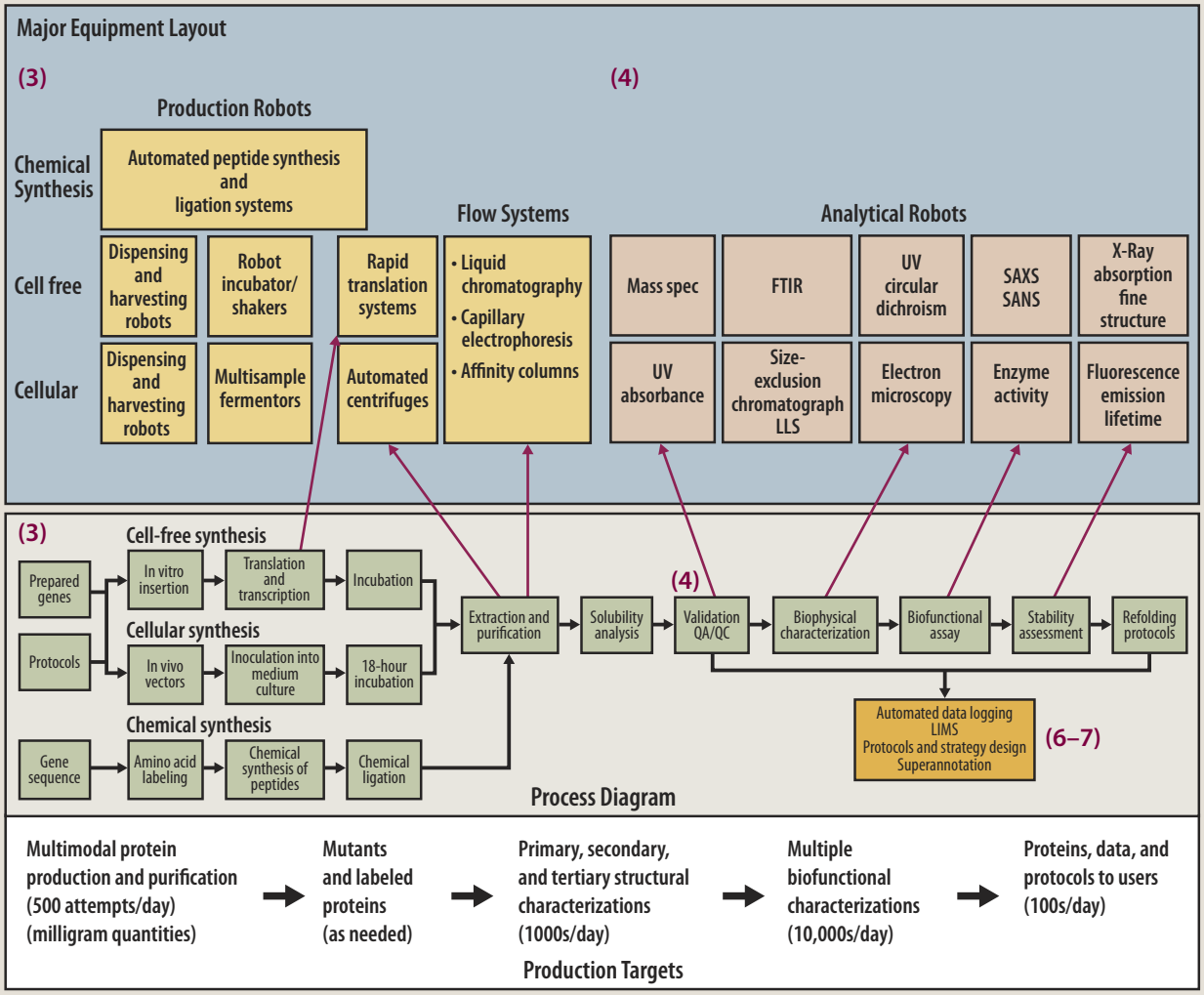
Virtually all steps in this process can be automated. A technician can obtain gene sequence information from a database and use genomics software to automatically direct a series of robots to produce a gene clone, verify the sequence, insert the clone into the appropriate vector, and produce DNA samples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

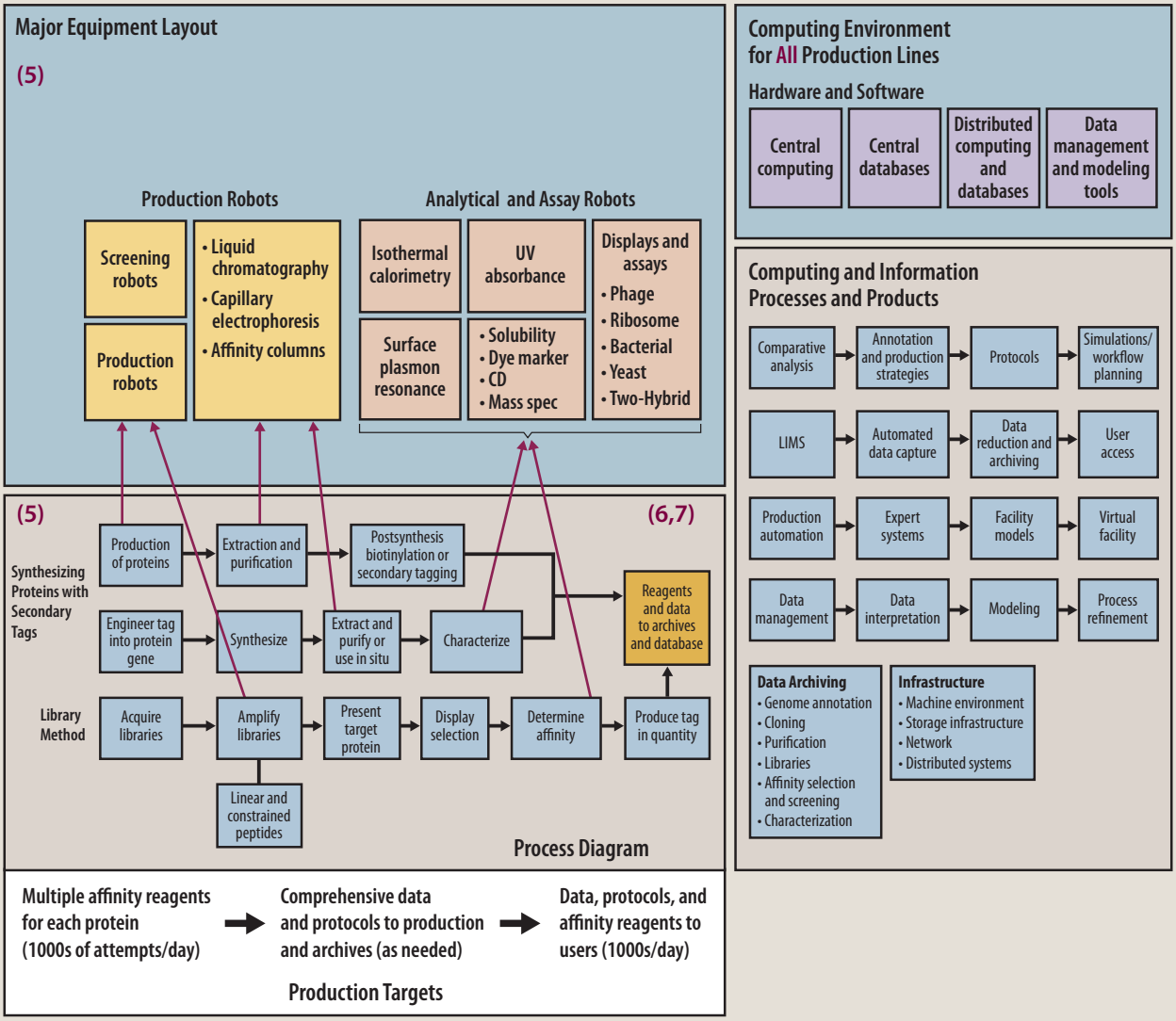
(3) Protein Production and (4) Characterization Lines



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.

(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.

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*).

FACILITIES