http://DOEGenomesToLife.org

Chicago, Illinois, May 29–30, 2003

GTL Facility for Production and Characterization of Proteins and Molecular Tags

*Organizers: Lee Makowski and Brian Kay, Argonne National Laboratory; and Jim Brainard, Los Alamos National Laboratory

*Facilitators: George Michaels, Pacific Northwest National Laboratory; Bill Studier, Berkeley National Laboratory; Debbie Hanson, ANL; Eric Ackerman, PNNL; Jim Brainard, LANL; Andrew Bradbury, Los Alamos National Laboratory; and Grant Heffelfinger, Sandia National Labortories

With the sequencing of over a hundred different microbial genomes now complete, the next step in understanding and exploiting microbial diversity is to focus on genome protein products. This is quite a challenge because the typical bacterial genome encodes about 5000 proteins, which are expressed as the cell needs them. Furthermore, many proteins are of unknown function; nor do we know the relative abundance and cellular location of each protein and how those details change with external stimuli. To tackle these important questions, DOE has plans to establish a state-of-the-art facility for the annual production and characterization of tens of thousands of proteins and their affinity reagents. Establishment of this facility would be followed by three others to enable systems biology focused on whole-proteome analysis, characterization and imaging of molecular machines, and analysis and modeling of cellular systems.

Contents

There are many persuasive reasons for establishing the Facility for Production and Characterization of Proteins and Molecular Tags. First, it would permit the proteomes of microbial systems relevant to DOE bioremediation, energy production, and carbon sequestration missions to be analyzed in a highly organized and systematic manner (see sidebar at right). This would save time and money and would provide the basic science to enable timely applications of microbial systems. Second, a dedicated staff with robot workstations would be able to develop a standard operating protocol, minimize variation among samples, and lead to data that are highly accurate. Third, a centralized facility would take advantage of economy of scale in expressing and characterizing thousands of proteins and affinity reagents at a time. Fourth, it would serve as a focal point for user training, reagent exchange, and proteomics information.

*The organizers and facilitators shown above planned and implemented the meeting and prepared this report. Their purpose was to provide a forum for the broad biological research community to discuss scientific and technical issues associated with planned user facilities for the Genomes to Life program. The report does not identify potential sites, leadership teams, final technical details, or funding for the facilities.

Published: September 30, 2003, http://doegenomestolife.org/pubs/prod_protein_mol_tags_workshop_052903.pdf

Users would include scientists participating in the Genomes to Life (GTL) program, the DOE laboratory system, academia, and industry. This facility would provide valuable materials (e.g., clones, proteins, affinity reagents) to the other three (see sidebar, GTL Protein Production Facility Deliverables). Protocols, proteins, and databases also would be distributed throughout DOE labs working on relevant biological systems, as well as to interested scientists in academia and industry. Finally, the availability of these reagents probably would entice other investigators and labs into working on these biological systems.

GTL Protein Production Facility Deliverables

- Production of the entire protein repertoire encoded in each microbe's genome
- Clones, protocols, data for protein production outcomes, and milligram quantities of each protein
- Protein variants designed for specific applications
- Affinity reagents for each protein produced, along with protocols and data
- Protein characterization data (biophysical and biochemical)

Meeting Organization and Topics

The workshop, held May 29–30, 2003, was convened at the Biosciences Division of Argonne National Laboratory. It was attended by representatives from the Department of Energy and National Institutes of Health as well as scientists from the national DOE laboratory system, academia, and industry. Over a 2-day period, six sessions focused on the

- role of bioinformatics in identifying genes, tracking samples, and linking data;
- technical challenges involving the cloning, expression, purification, and characterization of thousands of proteins at a time, including membrane proteins, periplasmic proteins, and very large proteins; and
- scientific challenges of developing affinity reagents specific to each protein produced.

Summaries of these sessions and related topics follow. Major research and development needs are outlined in the sidebar on p. 2.

Microbial Research Targets

Expressing and characterizing the proteomes of microbes with relevance to bioremediation, energy production, and carbon sequestration represent an important goal. Initial targets should be microbial systems that are the focus of existing GTL projects and for which complete or draft genome sequence is available. These candidate microbes include *Shewanella oneidensis*, *Geobacter metallireducens*, *Geobacter sulfurreducens*, *Deinococcus radiodurans*, *Pseuduomonas putida*, *Rhodopseudomonas palustris*, *Prochlorococcus marinus*, and *Synechococcus*. Target selection could be based on prioritization of additional microbial systems that are undergoing complete genome sequencing under DOE funding at the Joint Genome Institute and The Institute for Genomic Research, as well as microbes that become the future focus of the GTL program.

Computational Infrastructure

Computing and informatics activities associated with the protein production facility comprise three elements and associated teams.

The first is a comprehensive protein pipeline control system (PPCS) and laboratory information management system (LIMS) that will automate the collection of data and actively manage the pipeline.

- PPCS will enable the facility to be managed by a modest number of dedicated production staff from local and remote control stations. PPCS will control the workflow through the high-throughput pipelines in a completely automated fashion.
- The LIMS system should support real-time status monitoring of sample flow, quality, throughput parameters, and tracking. All sam-

ples, reagents, and work units will be bar coded and tracked. Data input to the LIMS system will be fully automated so humans are not required to input data, thus enabling rapid, accurate, and automatic data collection from key instruments. The LIMS system will construct a data package that parallels the flow of samples through the pipeline and will accompany material shipped to facility end users. It also will support monitoring of production runs, quality assessment, quality control, sample tracking, and status for both production staff and end users. Two LIMS systems are being evaluated for possible use in the protein production facility: the PRISM system, which is in active use at the Pacific Northwest National Laboratory, and the Nautilus system being used in a pilot project at the Oak Ridge National Laboratory. Compatibility with systems at other facilities is considered important. The LIMS system would interface to the bioinformatics environment, facilitating rapid integration of new data with existing databases. Particularly interesting is the rapid incorporation of knowledge of the successful application of a specific expression system or protocol to a gene of interest, enabling more accurate expression system assignment by bioinformaticists.

The second component of the computational infrastructure is the bioinformatics environment (BE). BE will enable the comprehensive characterization of genomes, genes, and protein families necessary for making informed target decisions. The system will support the selection of appropriate expression systems for a targeted set of genes; the integration of diverse information sources to provide a comprehensive view of each protein; and the development of associated characterization databases, quality vectors, and provenance data. The primary external interface to data produced by the facility, BE will be developed in cooperation with partner DOE laboratories and key external user communities.

The third component is an ongoing simulation effort that will construct a mathematical model of the production facility, including key protocols, protein production and characterization workflow, and information flow through production high-throughput pipelines. (See sidebar, Computer Simulations of Facility Processes.) This simulation tool will be used to gain a better understanding of resource management and optimization decision points in the facility and will enable capacity planning, throughput estimation, and control-systems design. Without constant feedback and interaction with the experimental partners in this endeavor, simulation will be inadequate in helping the facility achieve its goals; very focused simulation driven by real applications has the potential to solve problems that otherwise would go unanswered.

Protein Production Research and Development Needs

Computer simulation of the production and characterization process.

- Informatics software to track samples and organize data.
- Annotation, bioinformatics, and genetic analyses of several microbial genomes that will serve as the facility's biological starting point.
- Research into data-handling systems and databases to ensure compatibility with interconnecting systems from other GTL facilities.
- Systems, suitable for automation, for error-free transfer of coding sequences among clones.
- Improved methods for identifying protein complexes and interaction partners.
- Systems for coexpression of multiple proteins in the same cell.
- Cost-effective purification tags and columns for high-throughput, automated protein purification.
- Cost-effective improvements or gel-electrophoresis substitutes, suitable for automation, to analyze protein expression and solubility.
- Application of microfluidics to protein characterization.
- General methods for expression, purification, and characterization of membrane proteins.
- Methods for expressing large proteins or independently folded domains.
- Optimization of refolding protocols and development of high-throughput refolding screens.
- New affinity reagent libraries and automation of the selection process.
- Validation of general conditions for storage of a wide range of proteins.

Computer Simulations of Facility Processes

Setting up the protein production facility will require a great deal of careful planning and optimization to ensure that all parts of the protein-production process run at maximum efficiency. Even incremental improvements in one part could translate into large cost savings, leading to tremendous opportunities for leveraging modeling and simulation to ensure that the facility performs in the most efficient and cost-efficient manner. The DOE complex has great expertise in computer modeling and massively parallel computing, so much of the required simulations technology is readily available. Several potential opportunities in which modeling and simulation could greatly increase Facility I's effectiveness are outlined below.

Design of expression strategies. Perhaps the most important step in mass production of proteins is an effective expression strategy for each protein. Unfortunately, many different expression strategies exist for different proteins, and predicting which will be the most effective for a previously uncharacterized protein is often difficult. Trying different expression strategies can be the rate-limiting step in producing proteins, so informatics techniques that correlate protein sequence data with effectiveness of expression strategies will be highly useful.

Design of detergents for membrane protein processing. Membrane proteins are well known for their difficulty in production and characterization, yet their characterization is vital for a fundamental understanding of cell functioning. Much of the art and science of producing membrane proteins is focused around separating them from a specific membrane with their structures intact. This often is accomplished by detergents specific to each protein or lipid membrane combination. Experimentally testing the entire space of potential detergents for each new membrane protein is practically impossible, but molecular simulation could offer tremendous insight into which detergents would work best for conditions of interest. Techniques that "learn" about the effectiveness (or ineffectiveness) of different detergents can be used to train a molecular design algorithm to suggest the best detergents for the best compounds.

Design of microfluidic processing environments. Any mass-production facility of this size will need built-in miniaturization wherever possible. This not only saves valuable space but also allows many processes to be carried out with smaller quantities of material, which is critical to some hard-to-express proteins. Because of this, microfluidics has great potential for playing a role in building devices to perform many or all steps in producing proteins. Tuning these devices for optimal performance is a tricky process, however, since each new design must be fabricated (often an expensive process for "one-off" designs) and tested. Computer optimization already has been used to design channels to keep suspended material from separating and to perform mixing using special geometrical designs.

Protein Production Process

Overview. A facility capable of producing milligram amounts of tens of thousands of purified proteins per year is eminently feasible with the technology currently available, as demonstrated by the structural genomics centers and by commercial efforts in the pharmaceutical and biotechnology industries. Starting from complete genome sequences of a succession of microbes of interest to DOE missions, some of the facility goals will be to produce as many purified proteins (or protein complexes) as feasible from each microbe in forms and amounts that will be useful to the other GTL facilities and the broader scientific community; implement and improve technologies for obtaining difficult targets such as membrane proteins, insoluble proteins, and multisubunit protein complexes; and continually improve the efficiency and capacity of the overall production process. While less than half of the proteins specified by individual microbial genomes can be readily obtainable in functional form by today's technologies, this fraction is expected to increase substantially as processes and technologies improve.

Deliverables will evolve with user needs and desires. Proteins initially might be fused to a tag suitable for rapid purification, with the option of also producing each protein without any tags or

fused to different affinity tags, fluorescence tags, or other entities, depending on demand. Equally important will be sequence-verified expression clones and protocols for obtaining each protein in the desired form, the ability to generate new types of fusions as the need arises, and, for analysis of function, the potential to transfer coding sequences to vectors suitable for expression of tagged coding sequences in the original host.

The facility must be able to produce and archive at least 10,000 purified proteins a year at the 1- to 10-mg scale and to produce and distribute hundreds at the 100-mg or larger scale. All expression clones will be archived to allow the facility to produce additional amounts of any proteins as needed or to distribute expression clones to dispersed research groups. Potentially useful modes of distributing purified proteins might be large amounts of single proteins in solution, small amounts or pools of individual protein solutions arrayed in microtiter plates or on glass surfaces (i.e., protein chips), all suitable for analyzing enzyme activities, ligand binding, protein-protein interactions, and such. Useful sets might include all available proteins from individual microbes, sets of proteins involved in specific metabolic or signaling pathways, or sets of orthologs or homologs of the same types of proteins from different microbes. To carry out its distribution responsibilities, the facility must have an efficient inventory-and-retrieval system and substantial arraying and chip-making capabilities. Comprehensive information about the production history and characterization of every product must be readily accessible to users through the Internet.

Specifying in detail how an integrated, highly automated, high-throughput protein-production process tailored to GTL goals would look in 5 years is complicated by many viable alternatives for each production step and technology's continuing evolution. A generic production process could use current technology, but specific choices in engineering a production center would depend on such factors as the state of technology at the time decisions are implemented, suitability for integration into an automated production process, expected cost and efficiency of different alternatives, and the specific deliverables desired. Absolutely crucial to success will be comprehensive systems for process management, data and materials tracking, communication, and integration with the needs of other GTL facilities.

To maximize the facility's efficiency and impact, a research and development network of laboratories should be established for close interaction to optimize protocols and help implement the initial production process. Incorporating existing expertise at distributed sites will be an effective means of transferring new protocols, reagents, and expertise.

Bioinformatics. The first stage of the production process must be a comprehensive informatics analysis of each complete genome's coding sequences to be expressed as proteins. Prediction of such sequences depends on the latest tools, including comparisons with non-redundant known and predicted proteins, searches for protein-family or motif signatures, and placement into known protein families. Each predicted protein must be annotated as fully as possible (with regard to known or predicted functions) and linked to relevant sources of information.

Equally important is computational analysis and prediction of the physical, chemical, and biochemical properties of each protein to be produced. This process would include predictions of secondary structures, structural domains, isoelectric point, potential modification, secretion or localization signals, potential membrane insertion or association, potential ligand-binding motifs (e.g., metals, ATP, and cofactors), potential inclusion in multiprotein complexes, and potential interaction partners. As information accumulates in the facility and elsewhere, it also may be possible to predict sequences that influence the ability to produce proteins from clones, such as sequences that may cause pausing or frame-shifting in translation or signals for degradation. Improvements in the ability to predict whether individual proteins are members of tightly associated multiprotein complexes will be particularly important, as individual proteins from such complexes expressed without their interacting partners are likely to be insoluble or unstable.

A comprehensive informatics analysis should allow the grouping of any genome's proteins according to their likelihood of being well

expressed and soluble. This analysis could serve as the basis for assigning coding sequences to different pathways for expression and protein purification, thereby improving the efficiency of the production process. An important facility product should be the generation, testing, and improvement of these informatics predictions, taking advantage of complete tracking of production outcomes and mining the data for informative correlations between amino acid sequence and expression behavior.

Clone Construction. Once the coding sequences to be expressed are identified and classified, the cloning, expression, and purification pathway for each protein must be laid out. This process will evolve with experience but should be specified completely in software. Cloning will include selecting the appropriate vectors, designing oligonucleotide primers for the polymerase chain reaction (PCR), specifying how each PCR product is to be processed and inserted into the appropriate vector, and identifying the transformation process and host. In designing PCR primers, the potential for generating mRNA secondary structure that would occlude the translational start site should be evaluated and, if necessary, minimized by selecting among synonymous codons for the first few amino acids. Factors that affect vector selection might include the predicted expression outcome (soluble, membrane associated, need for interaction partner) and the presence or absence in the coding sequence of specific restriction endonuclease recognition sites that might affect cloning. Cloning and the software that interfaces with it must be modular, allowing easy modification or replacement in response to inevitable changes in goals, vectors, protocols, and automation.

Amplification and cloning technology is not foolproof, so replicate clones must be carried through the process and a correct clone identified by DNA sequencing of the entire coding sequence and critical adjoining vector components. Such quality control is essential because enormous downstream efforts will rely on each clone's correctness and rapid availability. Thus, a DNA sequencing capacity that can handle the most rapid anticipated rate of initial clone production must be an integral part of the facility. Based on current experience, four isolates would be sequenced from each of two types of clones for each protein (a total of eight). One clone would carry the coding sequence in a form suitable for error-free transfer to various vectors for expression as the native protein or coexpressed with interacting partners. The second clone would be capable of expressing the protein fused to an affinity tag suitable for rapid purification. One correct isolate of each type and its purified plasmid DNA would be archived; the remaining isolates would be discarded.

The requirement for error-free expression clones means that any clone generated by PCR must be validated by DNA sequencing. Reliable, efficient, and flexible error-free transfer of coding sequences among different vectors is urgently needed. This need arises because of the probable demand for many different (currently unpredictable) fusion products for each protein, the necessity of combining different coding sequences for coexpression to produce protein complexes, and the anticipated need of other GTL facilities for transfer of coding sequences to vectors for expression of tagged proteins in their original host. Performing such transfers by PCR is likely to impose a significant overhead of DNA sequencing to verify clones. Systems for transfer of DNA fragment from one vector to another through recombination (i.e., lambda insertion or Cre/lox) are attractive but add eight or more additional amino acids within the coding sequence to create the protein fusions, which may or may not modify the protein's structure and function. At least one integrated system with greater flexibility for error-free transfer of coding sequences to fusion or coexpression vectors is under development but has not yet reached the stage of extensive testing. Developing and validating cloning systems suitable for automated, error-free transfer of cloned coding sequences among vectors or for coexpression to produce protein complexes are high-priority near-term activities. Their success could have a substantial impact on the efficiency and success of other GTL facilities.

Generating and archiving a recombinant clone starting from PCR primers and template DNA is a linear process that requires a week or longer. PCR amplification, processing of the PCR product, insertion into a vector, transformation of the host cell, and plating for single colonies typically will take 1 to 2 days. The slowest individual step is to grow a colony from a single cell, which typi-

cally takes 18 hours or more and must be done twice sequentially, first to obtain the initial transformant and then to obtain a pure culture. Plasmid preparation and DNA sequencing to validate the clone typically would take another 2 days. Except for PCR and DNA sequencing, the same processes are involved in transferring coding sequences among vectors or transferring clones among hosts. In small-scale operations plating, picking single colonies, and inoculating cultures usually are done manually. To achieve the throughput envisioned for the protein production facility and to minimize errors, however, these processes must be automated on a scale that can process at least a thousand clones a day reliably and without cross-contamination.

In Vivo Expression and Purification. As expression clones are obtained, they will be tested for expression level and solubility. Some systems, such as the widely used and powerful bacteriophage T7 gene expression system, require that the expression plasmid be transferred to a special host. Recently developed auto-induction procedures greatly simplify the automation of expression and solubility testing in the T7 system, which can be readily accomplished with 100 µL or less of culture in plate format. Cultures may be lysed very simply and soluble and insoluble fractions separated by filtration in 96-well format. A potential bottleneck is the use of gel electrophoresis to analyze total soluble and insoluble fractions. The method is highly informative, but sample loading, gel staining, and capturing images of stained gels electronically are tedious and difficult steps to automate. Alternative methods should be explored for automated generation of required information, perhaps taking advantage of affinity reagents, automated microfluidics, or mass spectrometric methods.

Well-expressed, sufficiently soluble proteins can proceed directly to small-scale production and purification. Cultures grown in parallel in auto-inducing medium in shaking baffled vessels can readily produce hundreds of high-density, well-induced cultures suitable for automated purification procedures that generate purified proteins on the 1- to 100-mg scale. Purification capacity must be at least 96 proteins per day. Automated protein purification on this scale currently requires the initial use of affinity tags. Following the affinity step with automated size-exclusion chromatography would increase purity, provide information about solution size and oligomerization, and place the protein in the desired buffer. Portions of the purified protein would proceed to a pipeline for purity tests and a standard battery of characterizations of physical, chemical, or biochemical properties. The remainder would be aliquoted for storage. Developing conditions for long-term storage, from which useful protein can be recovered reliably, will be essential. After testing and characterization, aliquots of stored protein would be available for distribution to users in ways to be determined.

In Vitro Synthesis. In vitro expression systems derived from Escherichia coli or wheat germ have been implemented recently for high-throughput protein production in Japan. Such systems may be particularly effective for rapid coexpression testing of different protein combinations, screening for protein activities or interactions, and producing proteins under conditions difficult to achieve in vivo. In vitro expression is driven from PCR products (or plasmids) with T7 RNA polymerase, and the ability to use the same vectors for in vivo and in vitro expression could be a big advantage. An in vitro expression module in this facility could test possible uses and the cost-effectiveness of different combinations of in vivo and in vitro expression for different purposes. Developing resident expertise with in vitro synthesis of proteins will be important for the protein production facility's future success.

Chemical Synthesis. Another route to protein production is total chemical synthesis. Over the last few years, several technical advances involving chemical reactions and the use of protein-splicing inteins have permitted the ligation of short (i.e., 20-mer) peptides into longer polypeptides. All of a proteome's short proteins (i.e., <60 amino acids) are envisioned to be synthesized chemically in an expedient manner; difficult-to-express proteins less than 200 amino acids long would be attempted by chemical ligation. One appeal of chemical synthesis is that post-translational modifications, unnatural amino acids, and reporter groups such as biotin, fluorochromes, lanthanides, and isotopes can be designed readily into such proteins.

Challenges. Only about 30 to 50% of proteins specified by a microbial genome are expected to be well expressed and soluble enough to pass directly through a standard production and purification pathway. Membrane proteins and secreted proteins are important classes that usually can be identified by sequence analysis but are difficult to obtain by standard methods. Some membrane proteins can be produced in reasonable amounts in standard expression systems in *E. coli*, but this has not been the general experience. Most membrane proteins cannot be purified with the standard methods for soluble proteins, so development of general methods for extraction, purification, and characterization will be necessary to purify significant fractions of membrane proteins from individual microbes. (See sidebar, Production of Membrane Proteins.) Some secreted proteins may be obtained simply by cloning without the secretion signal, but others may be more readily obtained by expressing in configurations where they are secreted.

Another important class of insoluble or unstable proteins consists of members of multiprotein complexes expressed in the absence of their interacting partners. Informatics analysis sometimes can identify these proteins and their partners, and the capacity to make reliable identifications should expand as the body of information about protein interactions and microbial protein complexes increases. In particular, protein complexes will be a focus of the Facility for Characterization and Imaging of Molecular Machines, and information produced there should rapidly improve the overall ability to predict protein complexes in microbes. If protein complexes can be solidly identified in the initial proteome annotation, components can be coexpressed in cells and purified as assembled complexes. Proteins that are insoluble or appear to be poorly expressed are candidates for unidentified or tentatively identified complexes. If any information can be inferred about potential interaction partners, tests of coexpression would be desirable. In vitro expression may be ideal for rapid tests of likely combinations of proteins, since coexpression can be obtained simply by adding mixtures of expression plasmids (or even PCR products) directly to the translation mixture. If soluble complexes are identified in this way, clones for coexpression in standard expression systems could be generated.

Lack of an interaction partner probably is the reason for a significant number of insoluble or poorly expressed proteins, but other possibilities include improper folding, insufficient amount of a needed chaperone, presence of degradation signals, and problems in the translation process. Screening insoluble proteins under a range of refolding conditions has been remarkably successful in obtaining soluble and active proteins. Careful characterization is needed, however, to distinguish correctly folded proteins from microaggregates or soluble protein micelles. A high-throughput refolding capability will be essential for the facility. Certain fusion partners appear to promote folding or solubility in at least some cases, and a systematic test of their general applicability might be warranted in the facility's context. Large proteins tend to be poorly expressed, and systems with improved expression of large proteins might be identified. Another strategy is to attempt to clone and express independently folded domains, either by random sampling or by informatics predictions. Systematic outcome analysis for every protein that goes through the production process may identify causes for other types of failure and point to remedies that could be implemented in the facility.

Summary. Protein production will involve a highly diverse and complex set of pathways and processes that will evolve continually. To maintain cutting-edge performances, the facility will rely on a distributed network of laboratories engaged in developing technologies critical to the needs of the facility. An ongoing simulation effort of key facility procedures will optimize these processes and enable facility goals to be achieved.

Production of Membrane Proteins: Challenges and Solutions

Membranes and their associated proteins compartmentalize specialized machinery that provides the means by which cells and organelles communicate, generate energy, take up nutrients, excrete wastes, transduce signals, and build gradients of ions and other small molecules used to fuel all normal cellular activities in healthy organisms. In any organism, roughly one-third of genes revealed by genome-sequencing programs encode membrane proteins. Unfortunately, membrane-associated proteins have been notoriously difficult to purify in quantities sufficient for extensive biochemical studies. For significant advances to be made in studying membrane proteins, innovative strategies are needed for identification, expression, purification, characterization, and structure determination of membrane proteins.

Areas for improvement include:

Identification. Improvements are necessary in bioinformatics capabilities to enhance recognition of genes encoding membrane proteins. Especially lacking are methods for reliable prediction of the number and location of membrane-spanning domains, their secondary structures, and their transmembrane topology. Identification transmembrane protein sidedness will imply potential sites for ligand binding and protein-protein interactions.

Expression. A variety of heterologous expression systems must be developed, adapted, and employed because no single system is likely to prove useful for all membrane proteins. The protein production facility must be capable of expressing membrane proteins in *Escherichia coli* and alternative prokaryotic hosts, yeasts, insect cell culture, Semliki Forest virus culture, and mammalian cell culture. Vectors or systems that couple expression of membrane protein biogenesis partners, chaperones, and protein ligands should be designed and employed within every heterologous expression system. Since expression in some systems will be at lower levels than those observed for soluble proteins, highly sensitive methods must be developed to assess membrane protein-expression levels.

Purification. High-throughput assays must be developed that can screen protein-detergent or protein-lipid interactions and monitor the integrity of membrane proteins during extraction and purification. Success in purifying membrane proteins that retain native functionality will be aided by the continued evolution of “designer” detergent and lipid molecules and by thorough characterization of membrane lipid content in each native organism. If the expression systems above do not yield natively folded, functional proteins, capabilities will be required for purifying and defolding membrane proteins from inclusion bodies.

Characterization. High-throughput biophysical and biochemical assays that determine the integrity and purity of target membrane proteins will be needed for quality control. These methods include mass spectrometry (for confirming molecular identify of the purified materials and evaluating their chemical purity), sizing high-performance liquid chromatography (for evaluating sample integrity), quasielastic or dynamic light scattering (for determining the molecular masses of protein-detergent complexes and oligomeric states of purified membrane proteins), and circular dichroism and Fourier transform infrared spectroscopies (for estimating secondary structure content of proteins in detergent micelles and lipid bilayers). For determining the activity of hypothetical membrane proteins, a battery of enzymatic and binding assays used for soluble proteins must be adapted to conditions that require the presence of detergent or a lipid bilayer. Characterization of purified membrane proteins potentially may be assisted by affinity reagents that stabilize them throughout the process.

Protein Characterization Process

The ultimate goal of the protein production facility is to provide the resources needed to create a dynamic knowledge base for understanding a living system. This facility will supply reagents (proteins, affinity reagents, and clones for protein production) to users in the science community and to the other three GTL facilities. Characterization and utilization of these reagents is essential to ensuring their quality and maximizing the return on GTL's scientific investment. The facility will fulfill both production and R&D roles integrated through process controls and rigorous information-management architecture and reporting requirements.

Standards and consistency of characterization data will enable an understanding not achievable by comparisons of data obtained by multiple labs under various conditions. Especially in the project's early stages, distributed end users should be allowed to verify a subset of facility characterizations. This will validate the quality-assurance program and represent an important confidence-building measure. Examples of analytical methods that should be considered in the initial suite of characterization capabilities in the facility include:

- Capillary sequencing to validate the clone or DNA template producing the protein or affinity reagent.
- Proteolytic-digest mass spectrometry to compare predicted and analytically measured molecular weight and amino acid sequence.
- Capillary electrophoresis or chromatography to monitor product purity and contamination. Mass spectrometric analysis also can contribute significantly to establishing product purity.

Examples of methodologies to evaluate and monitor protein and affinity reagent solubility, stability, polydispersity, and aggregation in solution include:

- Dynamic light-scattering measurements or scattering of other particles (e.g., X rays and neutrons) to probe at other length scales or resolving power.
- Differential scanning calorimetry to characterize stability.
- UV-VIS and fluorescence spectra using intrinsic chromophores and reporting dyes and tags.

Methods should be established for high-throughput characterization of elements of protein structure and dynamics. Although throughput of the following methods might very widely, all would provide valuable information and can be envisioned to scale. Candidates include:

- UV-VIS and fluorescence spectroscopies, again using intrinsic or extrinsic chromophores.
- Circular dichroism, especially with short-wavelength synchrotron radiation.
- Single- and multidimensional nuclear magnetic resonance spectroscopies for identification of folded and unfolded proteins and protein segments. Multidimensional spectroscopies (e.g., 1H-15N HSQC) have the advantage that they can provide more specific and unequivocal information but require ~0.3- to 1.0-nm concentrations of isotopically labeled proteins.
- Partial proteolysis or D_2O exchange and mass spectrometric analysis for identification of ordered and disordered regions of the protein.
- More specialized analysis tools for isotopic composition of labeled products: carbohydrate, lipid, detergent, metal, and cofactor content; and presence or absence of fusions and affinity and reporter tags.

Since the ultimate goal of GTL is to understand how proteins produced by this facility function within molecular machines, initial assays that would contribute directly to understanding function would be very useful. There are advantages and disadvantages to performing functional screens at the facility rather than at individual user laboratories. Detailed investigations of protein function requiring significant resources devoted to a specific protein's peculiarities are more efficiently performed in a distributed network of user laboratories where in-depth knowledge can be best applied to specific problems. For many technologies, however, economies of scale and uniformity in measurement of quality-assured materials argue for a single facility. Several examples include high-throughput assays for ligand, cofactor, and protein binding, as well as enzyme activity.

A distributed network of characterization is one model to consider. This model would engage a

wider community and utilize intellectual and capital resources for characterization beyond the facility. An additional advantage of a distributed characterization network is that the users become true collaborators vested in the results. This proposed model presents a number of challenges from data standards to investigator expectations for data sharing and intellectual property.

A tiered network of characterization needs should be established for the protein production facility and piloted over the next few years. DOE must invest in pilot projects to adapt existing technologies to high-throughput analysis of proteins and affinity reagents on platforms that would interface well to upstream and downstream processes. Pilot projects also would establish and validate experimental and data standards, implement a single information-technology system to reliably capture and share information with downstream users and the national scientific community, and establish a communication protocol to accomplish these goals.

Affinity Reagent Production

One of the facility's stated aims is to provide affinity reagents against protein targets from any proteome. Although affinity reagents such as antibodies have been invaluable in studying protein function (see sidebar, What Have Antibodies Done for the Study of Proteins?), generating affinity reagents for thousands of proteins in a proteome will be a major challenge. Such an enterprise will require multiple cross-disciplinary components, including protein scaffolds, selection platforms, facile cloning procedures, modular vectors, automation, custom software development, and LIMS. Affinity reagents must be easily derived and usable in real-world experiments, and producing them in quantities and quality sufficient for carrying out such experiments should require minimal efforts.

Scaffolds. Monoclonal antibodies (mAbs) are stable proteins of high affinity and specificity that have been used in many research procedures, but generating them is time consuming and labor intensive and requires immunization of mice. It appeared 10 years ago that mAbs eventually would be replaced by single-chain fragments of variable regions (scFvs) or fragments of antigen-binding domains (Fabs), which could be selected from large naive phage display libraries. Such libraries offered the advantages of diversity, high affinity, and specificity in a potentially high throughput format; eliminated the use of animals; and avoided problems of poor immunogenicity. While scFvs and Fabs have been very successful in some cases and essential in the development of antibody therapeutics, they often have suffered from poor stability and low expression yields. These limitations have prompted scientists to search for improved scaffolds that offer properties more appropriate for high-throughput proteomic applications (see sidebar, Attributes of the Ideal Affinity Reagent). In general, most of these scaffolds (such as domains derived from immunoglobulins, fibronectin, ankyrin, lipocalin, or protein A) offer promise in overcoming problems of working with antibody fragments as well as in enhanced stability and in the ability to bind their targets in inside cells.

Another promising scaffold is the green fluorescent protein (GFP), which has been engineered to contain antibody binding loops and combines the advantages of specific, sensitive, high-affinity mAbs with those of GFP (i.e., intrinsic fluorescence, high expression, stability, and solubility). In addition to protein scaffolds, peptides also have been used as affinity reagents. They offer some advantages in the study and identification of protein-protein interaction partners and sur-

What Have Antibodies Done for the Study of Proteins?

- Confirmed gene assignments by identifying proteins
- Determined protein levels under different conditions
- Identified and characterized post-translational modifications (> 100 described)
- Localized proteins (tissue, cell, organelle)
- Identified function (inhibition or mislocalization)
- Identified interacting partners in complexes
- Promoted crystallization for structure determination

Attributes of the Ideal Affinity Reagent

- Well displayed in all selection formats
- Functional production in all cellular compartments
- Stable
- Monomeric
- Expressed at high levels
- Selective, strong binding
- Intrinsic detection function
- Functionality instantly assessable
- Easy measurement of concentrations without purification

faces as well as enzyme inhibition, for which larger protein scaffolds are less well suited. Finally, DNA and RNA oligomers, termed aptamers, are a promising source of affinity reagents.

Selection Platforms. As different genome projects are in different phases of development, a single selection platform is unlikely to be appropriate for all genomes. Where purified proteins are available, physical selection methods (e.g., phage, yeast, or RNA displayed libraries) probably will be most effective. Although phage or mRNA displayed libraries most likely are amenable to high-throughput selection, yeast display will be extremely useful in characterizing selected binders where affinity and epitope identification can be carried out relatively easily. A system that permits rapid transfer of coding regions for affinity reagents from phage to yeast (and vice versa) may well combine the best of both platforms. When only the sequenced genome is available, two-hybrid or protein-complementation genetic methods are likely to be most effective, eliminating the need to produce purified protein. In these methods, interaction between binding ligand and target confers survival on the cell containing the interacting pair.

Downstream Use. Whichever selection method is used, downstream use must be straightforward. As the selected affinity reagents are likely to be put to very varied uses (see sidebar, Antibody Methods Applicable to Affinity Reagents), they must be easily formatted for use in different systems. This requires robust and facile mechanisms to transfer specific binding reagents from selection vectors to use vectors (e.g., prokaryotic expression, yeast display, intracellular expression, enzymatic fusion, and immobilization vectors). Systems based on Cre/lox offer one solution, with libraries of affinity reagents flanked by nonhomologous lox sites created in selection vectors that can be moved easily on either the single clone or library scale into downstream use or screening vectors.

Antibody Methods Applicable to Affinity Reagents

- Western blots
- Quantification assays (ELISAs, RIAs)
- Immunoprecipitation, mass spectrometry
- Immunofluorescence, immunohistochemistry
- Immunoelectron microscopy
- Intracellular perturbation
- Enzyme inhibition studies
- Immunopurification
- Protein/antibody arrays

Customers. One of the many advantages for large-scale production of affinity reagents is completeness (see sidebar, Potential Advantages of Affinity Reagents on a Proteomic Scale). This concept has been well appreciated in the DNA chip arena, where the ability to assess the transcriptional activity of all an organism's genes, rather than a readily isolated or identified subset, has led to completely unexpected discoveries in gene activation. Only when completeness of coverage is available can insights of this kind be made. In addition, the provision of high-quality affinity reagents on such broad scales will lead to greater collaboration and reproducibility among different experimentalists, as well as the elimination of animal use for immunization. Although initial customers probably will be DOE grantees, the facility likely will lead to unexpected scientific, biotechnological, and commercial opportunities and thereby to an increase in the potential user base.

Potential Advantages of Affinity Reagents on a Proteomic Scale

- Completeness: A set of affinity reagents for all gene products and post-translational modifications far more useful than a "sum of subsets" and allowing characterization and identification of unexpected protein roles
- Greater understanding of protein function on a proteomic scale
- Consistency of reagents leading to more reproducible interlaboratory experiments
- Elimination of animal use
- Facilitation of current research projects
- Low cost from economies of scale
- Forum for collaborations

Possible Education and Outreach Functions

- Make a major attempt to extract generalizations about protein expression, purification, and characterization, and make them accessible to the general protein community.
- Initiate a Web site where instructions and cautions can be posted, collected, and organized.
- Teach courses onsite to disseminate accumulated knowledge and technical experience.
- Encourage visiting scientists and establish a minisabbatical program to allow specific queries of the data set or to probe the effects of many variables. These studies are more efficient with high-throughput facilities and a large number of proteins.
- Establish a fellows program to offer relevant formal affiliations with the facility to leading scientists around the country. This group could act as a board of scientific advisors to gain general acceptance of the concept; lend prestige, expertise, and visibility to the facility; and ensure that the facility is linked to the academic and industrial communities.
- Start an intern program for graduate students or postdoctoral fellows to work and learn.
- Form an information-sharing cooperative with various structural genomics centers around the country that will be generating data, protocols, and expertise while focusing on their target microbes.
- Share best-practice protocols used in solving technical problems and identifying robust, effective, reliable methods. Facility staff will accumulate extremely useful information that, if carefully written up and disseminated, would help the entire protein community. Many of these best practices already have been identified in other labs and centers, but they need to be accumulated, tested, refined, and made generally available.
- Conduct protein outreach by preparing effective materials and identifying outstanding spokespeople who can explain to Congress and the general public the value of protein research and why it is important in truly understanding the biological processes underlying all living organisms.

Facility Workshop Attendees

May 29–30, 2003 — Argonne, Illinois

Name	Institute	E-mail
Eric Ackerman	Pacific Northwest National Laboratory	eric.ackerman@pnl.gov
Joanna Albala	Lawrence Livermore National Laboratory	albala1@llnl.gov
Carl W. Anderson	Brookhaven National Laboratory Biology Department	cwa@bnl.gov
Andrew Bradbury	Los Alamos National Laboratory	amb@lanl.gov
Jim Brainard	Los Alamos National Laboratory	jbrainard@lanl.gov
Richard Burgess	University of Wisconsin, Madison	burgess@oncology.wisc.edu
Mark Chiu	Abbott Laboratories	mark.chiu@abbott.com
Frank Collart	Argonne National Laboratory	fcollart@anl.gov
Mark Donnelly	Argonne National Laboratory	mdonnelly@anl.gov
Charles Edmonds	National Institutes of Health National Institute of General Medical Sciences	edmondsc@nigms.nih.gov
David Eiznhamer	Advanced Life Sciences	deizhamer@advancedlifesciences.com
John Flanagan	Brookhaven National Laboratory	jflanagan@bnl.gov
Dax Fu	Brookhaven National Laboratory	dax@bnl.gov
Tony Grabski	Novagen	Tony.Grabski@novagen.com
Deborah K. Hanson	Argonne National Laboratory	dkhanson@anl.gov
Grant Heffelfinger	Sandia National Laboratories	gsheffe@sandia.gov
Tom Holzman	Abbott Laboratories	tom.f.holzman@abbott.com
John C. Houghton	U.S. Department of Energy OBER (SC-74)	j.houghton@science.doe.gov
Andrzej Joachimiak	Argonne National Laboratory	andrzejj@anl.gov
Brian K. Kay	Argonne National Laboratory	bkay@anl.gov
Mike Kennedy	Pacific Northwest National Laboratory	ma_kennedy@pnl.gov
Steve Kennel	Oak Ridge National Laboratory	kennelsj@ornl.gov
Steve Kent	University of Chicago	skent@uchicago.edu
Rosie Kim	Lawrence Berkeley National Laboratory	r_kim@lbl.gov
Youngchang Kim	Argonne National Laboratory	ykim@anl.gov
Shohei Koide	University of Chicago	skodie@uchicago.edu

Frank Larimer	Oak Ridge National Laboratory	larimerfw@ornl.gov
Scott Lesley	Novartis Research Foundation Genomics Institute	slesley@gnf.org
Phil LoCascio	Oak Ridge National Laboratory	i15@ornl.gov
Lee Makowski	Argonne National Laboratory	lmakowski@anl.gov
Betty Mansfield	Oak Ridge National Laboratory	mansfieldbk@ornl.gov
George Martin	Roche	george.martin@roche.com
George Michaels Bioinformatics Director	Pacific Northwest National Laboratory	george.michaels@pnl.gov
Len Napolitano	Sandia National Laboratories	lmnap@sandia.gov
Bob Novy	Novagen	bob.novey@novagen.com
Lee Opresko	Pacific Northwest National Laboratory	lee.opresko@pnl.gov
Danny Rintoul	Sandia National Laboratories	rintoul@sandia.gov
Rick L. Stevens	Argonne National Laboratory	stevens@mcs.anl.gov
Bill Studier	Brookhaven National Laboratory	studier@bnl.gov
Cliff Unkefer	Los Alamos National Laboratory	cju@lanl.gov
Etti Van Etten	Purdue University	eharms@bilbo.bio.purdue.edu
Andy Walker	Sandia National Laboratories	awwalke@sandia.gov

Prepared by Eric Ackerman (Pacific Northwest National Laboratory), Andrew Bradbury (Los Alamos National Laboratory), James Brainard (LANL), Richard Burgess (University of Wisconsin-Madison), Deborah Hanson (Argonne National Laboratory), Brian Kay (ANL), Lee Makowski (ANL), George Michaels (PNNL), Danny Rintoul (Sandia National Laboratories), Rick Stevens (ANL), and William Studier (Brookhaven National Laboratory).

Edited and formatted at Oak Ridge National Laboratory.