

05-SC-004, Project Engineering Design (PED), Facility for the Production and Characterization of Proteins and Molecular Tags

1. Construction Schedule History

	Fiscal Quarter				Total Estimated Cost (\$000)
	A-E Work Initiated	Completed A-E Work	Physical Construction Start	Physical Construction Complete	
FY 2005 Budget Request (Current Estimate).....	1Q 2005	2Q 2006	N/A	N/A	5,000 ^a

2. Financial Schedule

(dollars in thousands)

Fiscal Year	Appropriations	Obligations	Costs
2005	5,000	5,000	3,000
2006	0	0	2,000

3. Project Description, Justification and Scope

This PED request provides for Title I and Title II Architect-Engineering (A-E) services for the first Genomics: GTL facility - the Facility for Production and Characterization of Proteins and Molecular Tags. The design effort will be sufficient to assure project feasibility, define the scope, provide detailed estimates of construction costs based on the approved design, working drawings and specifications, and provide construction schedules including procurements. The design effort will ensure that construction can physically start or long-lead procurement items can be procured in the fiscal year in which Title III construction activities are funded.

Genomics: GTL User Facilities

Genomic information is providing the starting point for understanding the instructions for the manufacture of all of life's molecular machines and the systems needed to control and operate them. Understanding, not "simply" decoding, the operation, function, and coordination of genome information will be the next transforming phase in biology. From experience gained in sequencing genomes and conducting large scale biology projects, we have learned that the combined capabilities and imagination of biological, physical, and computational scientists will be needed to organize creative new venues for discovery.

^a The full Total Estimated Cost (design and construction) ranges between \$170,000,000 and \$200,000,000. This estimate was based on preliminary data and should not be construed as a project baseline.

The central goal of the Genomics: GTL program is to understand the microbes and communities of microbes, and their molecular machines at the molecular level to address DOE and national needs. The DOE Office of Science has the ability and institutional traditions to bring the biological, physical, and computing sciences together at the scale and complexity required for Genomics: GTL success.

The Facility for the Production and Characterization of Proteins and Molecular Tags will implement high-throughput production of and characterization of microbial proteins.

This resource will help build a bridge between large and small laboratories by making the most sophisticated and comprehensive technologies, materials, and information equally available to all scientists. Using combinations of new equipment and technologies, automation, data management, and data analysis tools, this user facility will provide the Genomics: GTL program and the scientific community with an unprecedented resource for systems biology.

SC has determined that the site selection strategy for this facility will be based upon a competition within the DOE laboratory system. This facility will be a high throughput production facility that will produce or isolate hundreds of proteins or molecular tags. It will likely require a close working relationship with other facilities and resources such as synchrotron light sources and neutron sources that will provide some of the many resources that will be needed to characterize the products of this facility. This facility will also be highly dependent on the development and use of robotics for many aspects of the protein and molecular tag production lines. Finally, high performance computational resources will be required to plan, monitor, and characterize both the production, characterization, and inventory aspects of this facility. All of these features are necessary to ensure that this facility provide high quality, reproducible products to the scientific community and users of this facility. The National Laboratory setting will be essential for this facility, for the National Laboratories provide both the necessary experience in developing and operating large multi-disciplinary high-throughput facilities, and the close proximity to the associated specialized technological resources cited above.

Key performance criteria under consideration for selecting a contract and for inclusion in the resultant contract include maximizing the involvement of the full scientific community in the design, construction, and use of this facility and ensuring broad public notice to universities and other potential users of the need for input in the design, construction, and use of this facility. A tentative schedule for the Facility for the Production and Characterization of Proteins and Molecular Tags includes a solicitation for site selection in February 2004, followed by an information workshop approximately one month after the solicitation, and review and site selection in the summer of 2004. The Project Engineering and Design datasheet will then be updated to identify the project site.

FY 2005 Proposed Design Projects

05-01: Facility for the Production and Characterization of Proteins and Molecular Tags

Fiscal Quarter				Total Estimated Cost (Design Only) (\$000)	Full Total Estimated Cost Projection (\$000)
A-E Work Initiated	A-E Work Completed	Physical Construction Start	Physical Construction Complete		
1Q 2005	2Q 2006	N/A	N/A	5,000 ^a	N/A ^a

(dollars in thousands)

Fiscal Year	Appropriations	Obligations	Costs
2005	5,000	5,000	3,000
2006	0	0	2,000

The Facility for the Production and Characterization of Proteins and Molecular Tags, will surmount a principal roadblock to whole-system analysis by implementing high-throughput production and characterization of microbial proteins. It also will generate protein-tagging reagents for identifying, tracking, quantifying, controlling, capturing, and imaging individual proteins and molecular machines in living systems. Over the next 10 years, our goal is to produce 250,000 proteins in milligram quantities; around 1 million molecular tags for those proteins; and multiple biophysical characterizations of each, beginning with an organism's genomic sequence.

Research is being conducted in the Genomics: GTL program to develop the core technologies that will underpin the high throughput capabilities of this facility including the development of technologies for the high-throughput synthesis of proteins and their biophysical characterization and for the production of molecular tags to identify individual proteins and to characterize multi-protein complexes in microbial cells.

It is recognized that no satisfactory general approach currently exists for the production of proteins in the laboratory from DNA sequences and that not all proteins will likely yield to the same techniques. It is expected that a variety of both cell-free and cell-based systems will be required, as well as multiple characterization methods. Production and characterization technologies should be scalable, economic, and sufficiently robust to work in a production environment. Another early need is the development of improved techniques for predicting from sequence what production and purification approaches are most likely to succeed with each protein. Thus, informatics is an integral component. Algorithms based on data from successful and failed protein expressions are expected to substantially inform and improve future protein production efficiency. Informatics coupled with biophysical characterizations are expected to provide functional insights that may also explain why such a large number of biologically important, full-length proteins either can not be expressed in soluble form, or have structures that cannot

^a The full Total Estimated Cost (design and construction) ranges between \$170,000,000 and \$200,000,000. This estimate was based on preliminary data and should not be construed as a project baseline.

be determined once expressed. These proteins may include substantial disordered regions that adopt structures only after interaction with appropriate protein binding partners. Reliable predictive algorithms based on expression and characterization databases are therefore needed to predict disorder and binding partners.

Research is being and will be supported to:

- Optimize cloning and clone validation techniques to support the protein production process.
- Optimize cell-free and cellular expression methods.
- Optimize protein purification protocols.
- Improve strategies for increasing the fraction of proteins that can be synthesized by automated methods. This may include sequence-based predictions of methods most likely to succeed and insights for optimization of expression protocols.
- Optimize high-throughput, economical approaches for characterizing synthesized protein to assess product quality and to predict protein function such that each protein produced will be characterized biophysically under several conditions.

Research is also being supported to advance the technology needed to mass-produce molecular tags for proteins and protein complexes as tools to be used for determining function. As a top priority, technologies are being developed for mass-producing specific protein recognition tags capable of functioning as capture reagents in affinity extraction and purification protocols and as labeling reagents for intracellular and *'in situ'* localization and mapping studies. As for protein production strategies, these technologies must also be scalable to permit large numbers of useable molecular tags to be produced and characterized per year at affordable costs. None of the many approaches under development to address this problem have yet demonstrated sufficient scalability. It is assumed that purified protein 'targets' will be provided to the researchers in micro-gram to milligram quantities so that tags can be optimized and characterized. Tags that interfere with function as well as those that do not interfere with protein function are both needed to help better define the biological roles of proteins.

Research is being and will be supported to:

- Develop scalable methods for producing 'epitope-directed' affinity reagents of high specificity and affinity for proteins capable of functioning either as affinity extraction and capture reagents or as intra-cellular labeling reagents. High success ratios (fraction of protein epitopes yielding useful reagents) are essential.
- Improve protein-directed affinity tag design to improve tag utility, e.g., to facilitate subsequent purification and or/imaging, to facilitate release of the tagged protein, to image with and without disrupting activity, etc.
- Improve methods for developing tags directed specifically to protein complexes as distinct from their component proteins. Labeling complexes with and without disrupting interactions amongst protein components will provide important functional insights.

method. Consequently a significant component of the protein production and characterization facility will be research into new methods of protein production and into automation of existing methods of expression, purification and characterization. Proteins that cannot at present be readily produced include most membrane proteins, high molecular weight proteins, toxic and unstable proteins, proteins with unknown co-factors and proteins that are integral parts of complexes. Consequently a significant research effort will be needed to (i) address the 50% of all proteins that currently cannot be produced in milligram quantities in soluble, native conformations (ii) automate all portions of the multiple synthetic routes needed (iii) automate the purification and characterization of all proteins (iv) devise informatic methods to predict optimal strategies for each protein to be produced (v) develop novel libraries of affinity tags, advanced methods of library production and methods for screening these libraries.

The primary production facility will include approximately 125,000 to 175,000 sq. ft. consisting of laboratory space for production, as well as research, office and administrative space. A protein characterization network including researchers from multiple national laboratories and universities will utilize the proteins and affinity tags produced in the facility and feed the results of characterization experiments into the central facility database.

4. Details of Cost Estimate ^a

	(dollars in thousands)	
	Current Estimate	Previous Estimate
Design Phase		
Preliminary and Final Design costs (Design Drawings and Specifications).....	3,600	N/A
Design Management costs (13.9% of TEC).....	700	N/A
Project Management costs (13.9% of TEC).....	700	N/A
Total Design Costs (100% of TEC).....	5,000	N/A
Total, Line Item Costs (TEC)	5,000	N/A

5. Method of Performance

Site selection will be made based on a complete scientific, technical, and project management review of offers received. Conceptual design of the facility and technical equipment will be completed by the fourth quarter of FY 2005 to establish an appropriate cost range and conceptual scope. PED will be utilized to perform preliminary and final design of the building; and engineering, design, and development of the technical equipment. Design services will be obtained through competitive and/or negotiated contracts. Site staff may be utilized in areas involving security, production, proliferation, etc.

^a These costs reflect only those associated with design phase activities only.

6. Schedule of Project Funding

(dollars in thousands)

	Prior Year Costs	FY 2003	FY 2004	FY 2005	Outyears	Total
Facility Cost						
PED	0	0	0	3,000	2,000	5,000
Other Project Costs						
Conceptual Design Costs	0	0	850	150	0	1,000
NEPA Documentation.....	0	0	150	50	0	200
Total, Other Project Costs.....	0	0	1,000	200	0	1,200
Total Project Cost (TPC).....	0	0	1,000	3,200	2,000	6,200

- Improve strategies for predicting, from sequence data, what potential protein epitopes are likely to be successful targets for tagging with and without interfering with function, and for predicting what tag development methods are likely to work for a particular protein/epitope.
- Develop imaging and labeling methods for multiplex mapping of proteins within cells. Simultaneously monitoring multiple labeled proteins will provide more comprehensive views of multi-protein complexes and their activities.
- Optimize informatics tools both for managing tag production processes and for managing the data resulting from their use

The Facility for the Production and Characterization of Proteins and Molecular Tags will be a user facility that integrates the necessary basic research, technology and automation to enable (1) the production and characterization of all proteins expressed by a genome and (2) the generation of affinity reagents to each protein. The goal of the facility will be to make possible rapid experimental characterization of the function of gene products on the scale of whole genomes.

Protein production will utilize multiple bacterial expression systems; cell free expression; and chemical synthesis methods. Protocols for automated expression, purification and characterization of proteins will be optimized for multiple classes of soluble proteins; membrane proteins; periplasmic proteins; and very small or very large proteins. Cloning, expression, purification, quality assurance (QA) and characterization will be carried out by automated systems directly linked to the Laboratory Information Management System (LIMS). Clones, proteins and affinity tags will be shipped to collaborators at other DOE laboratories, universities and corporate partners for further functional characterization.

Comprehensive protein characterization will include a QA suite to demonstrate that the proper proteins have been produced and to determine the physical state of the proteins. This suite will include mass spectrometry and DNA sequencing for protein identification; dynamic light scattering to assess solubility and ultra violet spectrometry.

Bioinformatics will be used to make an initial assignment of protein function where possible. Further characterization will be designed to elaborate on this assignment and identify additional biophysical and biochemical clues as to protein function. Biophysical characterization techniques to be used include circular dichroism to assess secondary structure; small angle x-ray scattering to determine quaternary structure; x-ray absorption fine structure to determine metal content and the environment of metal ions; wide-angle x-ray scattering to determine the quaternary structure as well as assignment of a structural fold. Biochemical characterizations will include mass spectrometry to identify co-factors and bound ligands; binding assays to the most common small molecule ligands; and high-throughput enzymatic assays for the most common biochemical activities.

The facility will be run by an advanced Laboratory Information Management System (LIMS) that will be capable of predicting the optimum method for production and purification of any protein based on its amino acid sequence and past experimental outcomes. The LIMS will collect experimental data from automated systems as well as manual input from handheld computers in a completely wireless environment. The results of expression, purification, quality assurance and characterization experiments will be automatically fed into a database of protein properties accessible through web-based servers. Results of experiments that are both successes and failures will be available to guide future work.

At least 50% of all proteins are anticipated to pose significant problems for any current production

**Science/Biological and Environmental Research/
05-SC-004/Project Engineering Design (PED),
Facility for the Production and Characterization
of Proteins and Molecular Tags**

FY 2005 Congressional Budget

