Genomes to Life Center for Molecular and Cellular Systems: A Research Program for Identification and Characterization of Protein Complexes

Project Overview

1.0 Project Summary

This proposed project focuses primarily on the objectives outlined in Goal 1 of the Department of Energy's (DOE) Genomes to Life (GTL) program being developed jointly by the DOE Office of Biological and Environmental Research (OBER) and DOE Office of Advanced Scientific Computing Research (ASCR) within the Office of Science (SC). Goal 1 seeks to identify and characterize the complete set of protein complexes within a cell to provide a mechanistic basis of biochemical functions. Achieving this Goal would give us the opportunity to understand cells and their components in sufficient detail to predict, test and understand the responses of a biological system to its environment. Further, Goal 1 forms the foundation necessary to accomplish all of the other objectives of the GTL program, which focus on gene regulatory networks and molecular level characterization of interactions in microbial communities. We propose to establish a Center for Molecular and Cellular Systems that will be anchored at Oak Ridge National Laboratory (ORNL) and Pacific Northwest National Laboratory (PNNL), taking advantage of existing expertise and capabilities for the identification and characterization of protein complexes. We will initially focus on the identification and characterization of protein complexes to elucidate cellular function in two microbial systems, *Shewanella oneidensis* and *Rhodopseudomonas palustris,* to obtain a knowledge base that can provide insight into the relationship between the complement of protein complexes in these microbes and their biological function. The Center will be established employing state-of-the-art techniques for the identification and characterization of protein complexes. We will build toward meeting the high throughput requirements of the GTL program and linking the data obtained to biological function. Therefore, we also propose to establish a dynamic research program that supports the goals of the Center by focusing on the development of new capabilities for sample preparation and complex separations, molecular level identification of the complexes, characterization of the complexes in living cells, and critical bioinformatics and computing.

ORNL: Michelle V. Buchanan, Director, Chemical Sciences Division
Oak Ridge National Laboratory
P.O. Box 2008, MS-6129
Oak Ridge, TN 37831-6129
Phone Number: 865-574-4986
Fax Number: 865-574-4902
E-mail: buchananmv@ornl.gov

PNNL: H. Steven Wiley, Director, Biomolecular Systems Initiative
Pacific Northwest National Laboratory
P.O. Box 999
Richland, WA 99352
Phone Number: 509-373-6218
Fax Number: 509-376-1494
E-mail: Steven.Wiley@pnl.gov

ORNL: Frank W. Larimer
Oak Ridge National Laboratory
P.O. Box 2008, MS-6480
Oak Ridge, TN 37831-6480
Phone Number: 865-574-1253
Fax Number: 865-241-1965
E-mail: larimerfw@ornl.gov

1.2 Introduction

The Human Genome Program has catalyzed the production of information and tools that will provide a new foundation for modern biology, one that is based on understanding the molecular-level processes that govern life at the cell and organism levels. The recent completion of the draft sequence of the human genome was a revolutionary scientific achievement and the sequences of many other organisms continue to be completed [Lander, et al., 2001; Venter, et al., 2001]. However, knowing the sequence of DNA in an organism is only the initial step in understanding how the information held in DNA sequence produces the myriad molecular processes that allow a cell or organism to function. Another key to elucidating the basis of function in an organism is contained within the organism's complement of proteins that are expressed according to the instructions contained in the DNA sequence. The number of proteins present in an organism can be much larger than the number of its genes due to post-translational modifications, alternate gene splicing (primarily restricted to eukaryotic cells), and other alterations [Gesteland and Atkins, 1996]. Furthermore, the collection of proteins within an individual cell is dynamic, changing in amount, types of modifications, and location within the cell in response to a wide variety of external stimuli. Proteins present at any one time within the cell typically form assemblies (protein complexes) that execute a multitude of cellular processes. These "machines of life" embody life's most essential functions, such as growth, metabolism, replication, repair, and response to intra- and extracellular signals. Thus, protein complexes are responsible for a cascade of molecular processes within and between cells, and dictate how a cell or organism interacts with its environment.

The DOE SC has outlined a scientific program, GTL, which focuses on leveraging information obtained from genome sequence to "achieve a fundamental, comprehensive, and systematic understanding of life," from the molecular, to the cellular, and to the organism levels. This program has four major goals:

(1) Identify and characterize the molecular machines of life
(2) Characterize gene regulatory networks
(3) Characterize the functional repertoire of complex microbial communities in their natural environments at the molecular level
(4) Develop the computational methods and capabilities to advance understanding of complex biological systems and predict their behavior.

This project focuses primarily on achieving the objectives outlined in Goal 1 of the GTL program. Achieving the aims of Goal 1 is necessary to accomplish all of the objectives of Goals 2 and 3, which focus on gene regulatory networks and characterization of molecular interactions in microbial communities. Goal 1 seeks to identify and characterize the complete set of protein complexes within a cell to provide a mechanistic basis of biochemical functions. At present, we do not fully understand the size and diversity of all protein complexes, some of which are comprised of dozens of components, nor do we know how many of these complexes exist at a given time, although the number is estimated to be in the thousands. This goal also seeks to elucidate the dynamics of the protein complexes, including changes in concentration, stoichiometry, and spatial localization within a cell in order to obtain a complete understanding of the relationship of these complexes to cellular function. Finally, we also need to be able to follow how changes to the protein components within the complexes (arising from gene alterations, post-translational modifications, and other variations) affect biological function. Achieving this Goal would give us the opportunity to understand cells and their components in sufficient detail to predict, test and understand the responses of a biological system to changes in its environment.

Environmental microbes are key to DOE missions in energy production, environmental restoration, carbon sequestration for mitigation of global climate change, and chemical and biological challenges to national security. These single cell organisms have evolved over nearly 4 billion years, developing a diverse range of biochemical pathways to enable them to survive under even extreme conditions of high temperature, high radiation, and presence of toxic chemicals. The evolution of these biochemical processes has resulted in the

ability of specific organisms to convert carbon dioxide into other carbon forms, reduce metals, and metabolize and detoxify organic chemicals. Protein complexes are key participants in these various biochemical pathways. Once knowledge of these biochemical pathways is obtained, we can establish how these pathways are altered under various environmental conditions, how these pathways differ from one organism to another, and how microbes interact with each other in a microbial community. This knowledge will allow us to harness these molecular processes to meet DOE's missions related to energy, the environment, and national security.

This proposed project focuses on the identification and characterization of protein complexes to elucidate cellular function in two microbial systems, S*hewanella oneidensis*, a heterotrophic gamma-proteobacterium, and *Rhodopseudomonas palustris*, an photoautotrophic alpha-proteobacterium of the purple non-sulfur group. These two organisms are the focus of major DOE Microbial Cell Projects (MCP); this project will leverage the MCP, and in turn the MCP will be hugely advanced by the characterization of complexes proposed herein. In out years, as the high-throughput capacity to process complexes develops, we anticipate that other organisms, particularly those targeted by the MCP and the Microbial Genome Program (MGP), will be characterized. To expedite these studies, we have engaged the support of many of the MGP and MCP investigators in the community (see attached letters of support).

The growth of the non-sulfur purple photosynthetic bacteria can be adapted to fix CO_2 or to use organic molecules as a carbon source for basic metabolism. Energy for these processes can be obtained from photosynthesis or from reduced substrates. In the course of these processes some of these bacteria can degrade lignin and reduce various metals and oxyanions. The regulation of these processes and the protein complexes that perform the functions are of basic interest to the carbon, nitrogen and energy cycles on earth. Significant work has been done on the induction and interactions of these enzyme systems; however, much remains to be done. Many of these processes can be studied in a single organism, *R. palustris*. *R. palustris* is the most commonly isolated photosynthetic bacteria in field samples. When this bacterium is cultured under different conditions, the appropriate carbon, nitrogen and energy handling systems are induced to accommodate the situation. Understanding these processes on the molecular level is vital to downstream goals of energy production and environmental remediation.

Over the last several decades, it has become appreciated that bacteria can adapt to deal with most environmental situations that we call pollution. Pioneering studies in this field include work on camphor degrading bacteria that lead to the description of the mixed function oxidase systems. In addition to the many organic pollutants, contamination of soils with various metals has been recognized as a significant environmental problem. Several bacteria are capable of reducing oxidized metals. Among these organisms, *S. oneidensis*, *MR-1*, can reduce oxidized metals in solid deposits in the soil. The reduction process allows the bacterium to dispose of excess electrons generated by catabolism of carbon sources. The process is unique in that the electrons are ferried from the cell's interior out through the periplasm to the outer membrane where they can be transferred directly to the oxidized metal. Electron transport through complex, multiheme cytochromes mediates this function.

The components of these and all other complexes, and the way they function, are the subjects of study in this proposal. To accomplish this, we will first establish capabilities for production, isolation, and identification of protein complexes using current state-of-the-art capabilities. This work will establish capabilities designed to meet the needs of the longer-term goals of the GTL program that requires approaches for high throughput, genome wide identification and characterization of these complexes. Our dynamic research program will develop new capabilities that will allow us to examine thousands of protein complexes in a single year. The major objectives of this research program are to investigate new approaches for rapid, robust isolation of protein complexes, to analyze protein complexes (including stoichiometry, modifications, levels), and to examine spatial and temporal distributions of the complexes within the microorganism. An important part of this research program is the integration of computational tools that will combine data obtained as part of this project with that from other GTL studies, and structural information being obtained from x-ray crystallography and nuclear magnetic resonance and other sources, to elucidate the relationship between the complement of protein complexes present in the microbe, architecture, and dynamics with biological function and functional mechanism.

1.3 The Genomes to Life Center for Molecular and Cellular Systems: A Research Program for Identification and Characterization of Protein Complexes

Goal 1 of the GTL program focuses on the comprehensive identification and characterization of the complete set of protein complexes on a genome-wide scale and establishing the linkages to biological function. One step to accomplishing this goal is to study an organism under a set of standard conditions that will reveal the number of proteins in a given complex and the types and locations of modifications present. A stated goal of the GTL program is to identify 80% or more of the collection of protein complexes within a model microbe per year within the first five years of the program. We have identified a clear approach for accomplishing this task. The heart of the proposed project is the establishment of the **Center for Molecular and Cellular Systems**, which will focus on the goal of identifying and characterizing protein complexes from both *S. oneidensis* and *R. palustris* during the first three years of the proposed project. This will be a highly coordinated Center equipped with state-of-the-art analytical and computational tools required for the analysis of protein complexes. But current technologies will not be sufficient to meet the long term goals outlined in the GTL program requiring the study of thousands of protein complexes per year. To assess what steps need to be taken to reach this goal, we will critically evaluate the best current approaches for isolating and analyzing protein complexes. For this effort, we will target the identification of a limited number of complexes from two organisms that have been sequenced. Complexes from several different cellular compartments will be studied. By examining two microbes, we will gain insight into the general applicability of a variety of different tools, such as statistical inference and cell fractionation. By isolating complexes from different cell compartments, we can identify the technical bottlenecks that may hinder scaling our efforts to meet the ultimate goals of the GTL program. Bioinformatics tools will be applied to assist in identifying protein complexes with particular biological functions, their possible component proteins, and their cellular locations.

Which current approach is "best"? Although many different approaches are currently being used to identify protein complexes, it is unclear whether any single technique is superior in general. The optimal approach would characterize the entire complement of complexes (both strongly and weakly bound), with a minimum of artifacts. Most investigators exploit the particulars of their experimental system, such as the availability of specific antibodies or efficient genetic recombination systems. However, if the aim is to create a high throughput facility to identify all protein complexes, then it is important to determine the best overall strategy that can be applied to the greatest variety of protein complexes.

Recently, two groups attempted the global identification of protein complexes in yeast [Gavin et al., 2002; Ho et al., 2002]. The two groups used similar approaches although the first used tandem affinity tags inserted directly into genomic sequences and the latter used simple affinity tags expressed on extrachromosomal elements. Both used sequential gel electrophoresis followed by mass spectrometry. Of the 589 complex "entry points" examined by Gavin et al. and the 493 examined by Ho et al., 95 were common in both studies and thus provide a point of comparison. Of all of the proteins identified as being in a "complex" with these 95 entry points (a total of 1445), only 10% were identified by both methods. The vast majority of the identified proteins had never been described as being in a complex previously, but both methods could claim to successfully identify proteins previously described in the literature as being complex partners. From these data it is impossible to tell which method is better. Both can claim to identify many specific partners and both studies were carefully controlled and attempted to eliminate spurious results. However, the 90% non-overlap between the proteins identified by the two approaches suggests either that slight variations in experimental conditions can have a profound effect on the identification of protein complexes, or that the current technology has not advanced sufficiently to determine whether any given "identified" protein is actually in a complex.

Because of the uncertainties in the current technologies, we feel that it is essential to have a strong research effort associated with this project. We simply do not know what approach or combination of approaches is "best" for the characterization of protein complexes. It is the intent of our program to establish such criteria and to provide a solid scientific and technological basis for pursuing our studies. Thus we will use a combination of bioinformatics, separation technologies and analytical approaches to critically evaluate the performance of current techniques and develop improved ones for the future.

1.4 Outline of the Center for Molecular and Cellular Systems

The Center for Molecular and Cellular Systems, described in detail in the **Center for Molecular and Cellular Systems Subproject**, will be a distributed center, with the work conducted at both ORNL and PNNL. The dual laboratory approach is being implemented because of the complementary capabilities at both sites and because of the need to evaluate several different approaches simultaneously. For example, ORNL is currently implementing cloning strategies for the *Shewanella* genome and protein crosslinking combined with mass spectrometry whereas PNNL has created active programs in scFv generation and use of accurate mass measurements with high mass resolution mass spectrometry. By jointly comparing and combining the approaches being developed at both national laboratories, we are far more likely to develop an optimal approach for analyzing protein complexes. Common infrastructure requirements, such as in bioinformatics, are being jointly developed across the two laboratories, reducing waste and avoiding unnecessary duplication of efforts.

Our research strategy calls for the critical evaluation of different technical and experimental strategies during the early stages of this program. The optimum approach will then be implemented at both sites. For example, we plan to evaluate endogenously expressed and exogenously expressed tagged proteins as well as single-chain antibodies. As part of our research integration plan, we will evaluate the progress and accomplishments achieved with the different approaches and decide when the evidence favors one approach over the other. The best technique will then be implemented at both sites. Because of our two-year cycle of funding technology development projects (discussed in Management Plan, below), new technical approaches will be continuously evaluated at both sites, greatly accelerating the evolution of more efficient approaches.

As part of this Center, we will clone and express target proteins for complex identification, implement methods for isolation and purification of protein complexes, identify and characterize protein complexes by mass spectrometry, and validate cellular relevance of protein complexes by imaging techniques. Different techniques for isolation and analysis of the protein complexes will be evaluated and over the course of the first two years of this project, approaches will be compared and the optimum methods will be established at both Center sites. During the first year of the Center we will also establish a single Center-wide integrated data management, analysis and delivery system for sample identification/tracking, data acquisition and analysis, and data archival and sharing. This activity is described in detail in Specific Aim 6 in the **Genomes to Life Center for Molecular and Cellular Systems Subproject.** Members of the bioinformatics and computational team will be responsible for procuring commercially available hardware and software for these needs during the first year of the project. They will develop other capabilities that are not available commercially. Having a single system that can be accessed via our internal web page will enable all participants in this proposed project to track a sample throughout the preparation, analysis process, and data archival processes.

As we conduct biological studies to identify and characterize protein complexes within the Center, we will simultaneously establish a research program that addresses the technical bottlenecks identified and leads to the development and implementation of new capabilities for the sample preparation and analysis that will enable the identification and characterization of thousands of protein complexes per year. We have devised a plan for a research program that will support the goals of the Center by developing and implementing new separations, analysis, and computational methods that will enable the robust, high-throughput identification and characterization of protein complexes. Initial proposed research projects are described in four subprojects that address specific technology needs of the Center. The **High Throughput Complex Processing Subproject** focuses on new approaches for the rapid and parallel isolation of protein complexes from microbial cells and efficient separation of the individual complexes for analysis. The **High Throughput Molecular Characterization Subproject** focuses on the development of new mass spectrometry-based techniques for the high throughput analysis of protein complexes. The **Bioinformatics and Computing Subproject** focuses on development of new computational tools for the analysis and interpretation of mass spectral data, characterization of organizational structures, and establishing biological functions of proteins complexes. The **Molecular and Cellular Imaging Subproject** focuses on the validation of the location and function of protein complexes in living cells. The relationship between the various subprojects is illustrated in **Fig. 1.4.1**. The capabilities to be

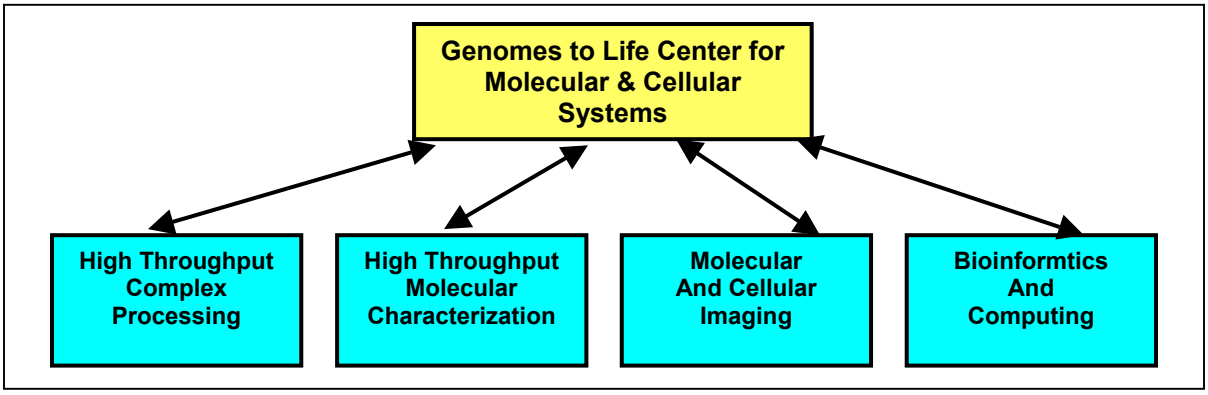

Fig. 1.4.1. Relationship of Subprojects
The Center for Molecular and Cellular Systems is the core of the proposed research project, focusing on the identification and characterization of protein complexes and the elucidation of the relationship between the complement of protein complexes in a microbe and their biological function. The four research subprojects are driven by the needs of the Center. As new capabilities are developed in the research subprojects, they will be incorporated into the Center. Although not shown in this diagram, the four research subprojects will be closely linked to coordinate synergy between capabilities as they are developed.

incorporated into the Center and the research tasks are described in the separate documents for each subproject. However, a brief overview of approaches for isolation, identification, and characterization of the complexes are given below to help understand the interactive nature of these subprojects and how they will contribute to the overall goal of this proposed project.

1.4.1 High Throughput Complex Processing

Current approaches to isolating protein complexes. Three basic classes of methods have been used to define protein complexes: (1) direct isolation, (2) genetic "two-hybrid" screens, and (3) affinity separation approaches. The direct isolation approach relies on physical separation techniques, such as ultracentrifugation, to separate specific complexes. To accomplish this, the complexes must have some physical characteristic that distinguishes them from other complexes within cells, and there must be a specific assay for the complex. Examples of complexes isolated using this approach include ribosomes and RNA polymerase holoenzyme. Physical isolation techniques display low specificity for biological molecules and usually must be combined in series to be effective. The additional requirement for specific assays makes this approach very slow and laborious.

In the genetic two-hybrid screening approach, a known gene is cloned into the "bait" vector, which places it next to the DNA-binding domain from some transcription factor. Separately, a second gene (or a library of cDNAs encoding potential interactors, or "prey") is cloned in-frame adjacent to an activation domain of a different transcription factor. When interacting prey and bait genes are expressed in the same cell, this allows transcription of another gene essential for growth. Recovery of the cells provides identification of the prey gene. Although proven useful for developing protein association maps, this approach is sequential in nature, requiring any potential bait protein to be evaluated extensively before use. Complications with the two hybrid system include false-positive results and inherent problems with interpreting the behavior of two fusion proteins in the context of a foreign yeast cell.

Affinity separation approaches rely on the ability of selective matrices to bind one protein in a complex, thus allowing isolation of all complexes containing that protein. The basis of the matrix affinity can be an antibody raised against a specific protein, or it can be a genetically encoded affinity "tag". Antibodies have been very effective in isolating protein complexes, but raising a full complement of suitable antibodies using traditional approaches is time consuming and expensive. This has prompted the more recent adoption of the genetically encoded affinity tag approach. This technique involves fusing the gene of interest with a specific genetic sequence specifying a peptide tag. The tagged protein can then be recovered using a solid matrix with an affinity for the tag. Proteins associated with the tagged proteins can then be identified by sensitive analytical techniques, such as gel

electrophoresis and mass spectrometry. We propose two possible solutions to these limitations. The first is to establish an automated system to produce the scFv's in a high throughput manner. The second is to rely on sequence tags inserted by high throughput cloning methods to isolate the complexes with a universal affinity approach.

Many variations of the affinity approach have been described, such as the use of tandem affinity tags, which allows a two-step purification protocol that can increase the specificity of the overall process. All of these variations have the same drawbacks, such as potential disruption of the protein structure by the fusion peptide, difficulties in recovering low-affinity partners, and the need to express the fusion protein in the target cells by gene transfection. The last drawback has been addressed by using exogenously expressed tagged proteins that can be mixed with cell lysates. If the exchange rate of proteins within a complex is significant, this approach can be very useful. It does, however, present its own difficulties, such as ensuring that the expressed tagged protein has the correct conformation, is exchanged rapidly, and that the incubation conditions are appropriate for complex formation.

Almost all two-hybrid and affinity tag approaches can potentially disrupt normal protein function and interactions. They are also sequential in nature because the genes must be individually modified and individually expressed in target cells. This can be done in a relatively high-throughput fashion, but each cell containing a tagged gene or protein must be isolated, carried, and propagated separately. This raises issues of gene-drift and the effect of the transfection and expression techniques on cell responses and phenotype. Each cell carrying a tagged gene can be used to identify a single "complex," which could be heterogeneous. To explore the assembly of complexes under a variety of different experimental conditions, each set of cells must be subjected to all of the conditions being studied and used individually to evaluate the formation of the complexes—an extremely laborious and inefficient process. In addition, these approaches are not even applicable unless the organism is genetically tractable and can be cultured.

A statistical approach, sometimes referred to as "guilt by association," has also been used to infer the identity of proteins in a complex. It assumes that proteins that are co-expressed and are observed as fusion proteins in phylogenetic homologs are likely to exist in complexes. This approach could potentially be applied to hard-to-culture organisms, but it is indirect and must eventually be verified by direct data.

A systematic approach to isolating protein complexes. The fundamental technology needs of a large scale effort for isolation of protein complexes are challenging and will require a highly coordinated effort to produce, isolate, and analyze the many hundreds of protein complexes in an individual cell. In general, the rate-limiting step in identifying protein complexes is <u>not</u> collecting the data, rather preparing the sample and analyzing the resulting data. It will be critically important that the analysis of the protein complexes be tightly integrated in a systems-based approach for understanding cellular function, including gene expression, protein expression profiling, imaging, and analysis of functional intracellular networks. These data can provide supporting evidence for which putative protein complexes actually exist in the organism. This type of supporting data is currently being generated in the MCP program for *S. oneidensis* and *R. palustris* and others will be generated as a matter of course in the overall GTL program. Bioinformatics tools developed as part of this project will build an integrated data base of this information and will be employed for the analysis, management, and sharing of data. Advanced computational tools will be employed to link protein complexes to regulatory and metabolic pathways/networks, modeling protein interaction and network dynamics, and to obtain new insights about how organizational structure of a protein complex relates to its biological functions.

There are several common technical issues that must be addressed to isolate, identify, and validate the presence of specific protein complexes from cells, regardless of the approach used. The most important issues involve the physical breakage of cells; separation into particulate, soluble, and membrane fractions; stabilization of

complexes; identifying nonspecific interactions; and minimizing protein contaminants. The Genomes to Life Center for Molecular and Cellular Systems will address these fundamental issues. Starting with technical approaches that represent the current state-of-the-art, we will compare and contrast different techniques for isolating and characterizing protein complexes. We will isolate complexes from different cellular fractions and use different approaches for affinity isolation and complex characterization. This work is essential for any systematic attempt to isolate and identify protein complexes from cells.

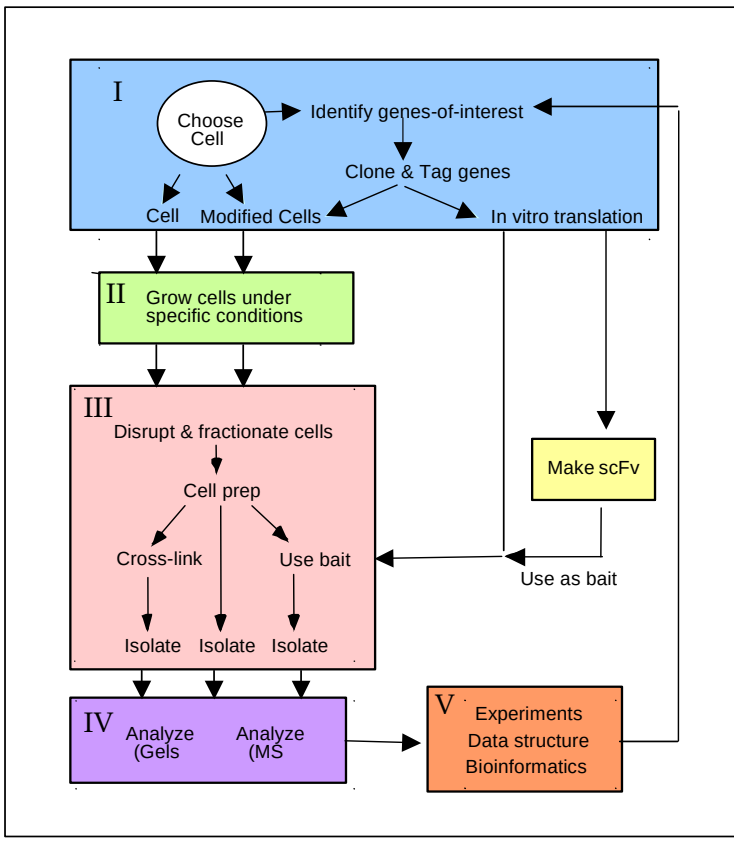

Fig.1.4.1.1. General Flow Diagram for Initial Approaches for Isolating and Analyzing Protein Complexes

Shown in **Fig. 1.4.1.1** is a flow diagram of our initial general approach for isolating protein complexes. Step I: Genes of interest are identified either by prior experimental data or using a statistical, bioinformatics approach. The genes are isolated by PCR and inserted into a universal vector. From this construct, the gene can be inserted into a bacterial expression vector in-frame with an appropriate affinity tag and inserted into the target cell. Alternately, it can be placed in a T7 vector for in vitro transcription/ translation. Step II: Normal cells, or those expressing the affinity tagged gene, are grown under experimental conditions under which the complexes are thought to occur. Step III: Cells are disrupted, fractionated and the complexes are isolated by affinity isolation approaches. A crosslinking step can be used to stabilize the complexes. Step IV: Proteins in the isolated complexes are characterized using a separation technology (such as gels or high-pressure LC) in combination with molecular identification (e.g., mass spectrometry). Step V: Data from this basic approach is then fed into the bioinformatics and computing component of the Center for analysis. Other technologies, such as scFv development and/or crosslinking are expected to complement or eventually replace the approaches being used in the Center as techniques are developed in the research subprojects (see below).

New high-throughput approaches for isolation of complexes. A fundamental problem of most current approaches is that they are intrinsically low throughput. To identify a large percentage of the protein complexes within a given cell in a reasonable amount of time will require new high-throughput approaches. The specific intent of the **High Throughput Complex Processing Subproject** component of our program is to develop these new approaches. One of these approaches will be to convert the normal, sequential process of tagging and expressing genes into a highly parallel, automated process. The basis of this approach will be the large-scale, in vitro expression of tagged proteins. This technology has been recently developed by the Yokoyama group at RIKEN Yokohama Institute. These expressed tagged proteins can then be used to directly isolate protein complexes from cell lysates in a massively parallel fashion. Alternately these expressed tagged proteins will be used for the generation of a panel of single-chain antibodies (scFv) in a high-throughput manner.

We have recently created a high-diversity scFv library displayed on the surface of yeast that allows the high-throughput, multiplexed selection of scFv using high-speed flow cytometers. This technical breakthrough eliminates most of the problems associated with traditional approaches to antibody production, such as high expense and restrictions against self-antigens. The scFv “affinity tags” also allow the isolation of many different

complexes simultaneously from a single cell extract. The direct protein tag and scFv approaches both allow a large batch of genetically unmodified cells to be used as the single source of extract, thus obviating problems with cell-cell and inter-experimental variations. Use of the scFvs also eliminates problems associated with modification of the protein being studied. Individuals at our laboratories have successfully demonstrated each part of this strategy.

We propose to apply these approaches in a coordinated fashion to create a high-throughput complex isolation approach. We will investigate incorporation of many of the isolation and sample preparation into a format that will minimize sample handling. Specifically, we will optimize methods for separation of protein complexes and their components in microfluidic "lab-on-a-chip" devices that have been pioneered at ORNL (see description in the **High Throughput Complex Processing Subproject**). This approach has been demonstrated to provide high resolution, rapid separation of peptides from tryptic digests [Gottschlich, 2000] and has been interfaced directly to a mass spectrometer [Lazar, 1999]. In the longer term, it may be possible to sort cells, lyse them, isolate protein complexes, and separate them on a single microfluidic chip and interface the chip directly to the mass spectrometer for high efficiency sample preparation and analysis.

1.4.2 High Throughput Molecular Characterization

Current capabilities of Mass Spectrometry for Protein Complex Identification. Mass spectrometry (MS) is a rapidly advancing family of experimental tools that have in common the measurement of molecular masses of intact and fragmented molecules. In the context of characterization proteins and protein complexes, MS is uniquely capable of identifying the presence of modifications on the protein, including post-translational, mutational, and others, which can affect biological function. MS is inherently sensitive, being able to detect proteins at the picomole level and below and can be employed to quantitatively measure the detected proteins. For these reasons, MS has been identified in the GTL program as a key tool for the identification of proteins and protein complexes. Advances since the late 1980's have made MS applicable to larger biomolecules such as proteins, nucleic acids, and their complexes. The key to forming gas-phase ions from these larger molecules, a prerequisite for MS analysis, lies in using electrospray (ES) [Fenn et al., 1989] and matrix-assisted laser desorption/ionization (MALDI) [Hillenkamp et al., 1991]. With these ionization techniques, MS can be used as a new type of high-resolution "readout" for many existing biological and biochemical procedures, providing biological information in the form of molecular weight. For instance, the advent of modem MS techniques has greatly expanded the utility of the venerable proteolytic digest. Accurate MS measurement of the masses of the proteolytic peptides yields a "mass map" that can be used to identify a protein, and is a foundation technology of the rapidly-expanding field of proteomics [Pandey and Mann, 2000]. Furthermore, tandem mass spectrometry (MS/MS) enables full or partial sequencing of the proteolytic peptides, in some cases without prior separation [Yates, 2000], providing an additional route to protein identification. Approaches have been described for the high-throughput MS analysis and identification of the suite of individual proteins in a cell or tissue [Gygi et al., 1999; Pasa-Tolic et al., 1999; Pandey and Mann, 2000], as have methods for studying isolated protein complexes [Loo, 1997; Rappsilber et al., 2000]. Exploitation of the capabilities of MS for high-resolution molecular weight measurement as well as structural interrogation offers a powerful tool for elucidating protein complexes in high-throughput fashion.

Mass spectrometry has been used to study intact noncovalent protein complexes [Loo, 1997; Miranker, 2000], yielding information on the stoichiometry of the complex and the molecular masses of the component proteins of the complex [Loo, 2000], as well as identification of the components of multiple-protein [Pandey and Mann, 2000]. Methods have been described for isolating protein complexes from cells by expressing one component of the complex as a fusion protein with affinity tags [Rigaut et al., 1999; Gavin et al., 2002; Ho et al., 2002]. The tagged component, along with co-purifying proteins present in complexes with the tagged component, is then purified. The components of the complex are then identified using 2D-PAGE/MS analysis with database searching. In addition to simply identifying the component proteins in the complex, it is also possible to obtain information on their relative geometries in the complex by partially crosslinking the complex after the affinity purification [Rappsilber et al., 2000].

State-of-the-Art Mass Spectrometry in the Center for Molecular and Cellular Systems. As we initiate the Center for Molecular and Cellular Systems, proteins from complexes isolated using techniques described above will be identified by an approach that combines "top-down" analysis (measurement of intact protein molecular masses) and "bottom-up" analysis (measurement of peptide molecular masses in enzymatic digests). The major advantage of this combination is identification of post-translational modifications via intact molecular mass measurements that could be missed by a peptide-only method. Stoichiometries of proteins in complexes will be identified by whole protein analysis in conjunction with purified expressed proteins, which will serve as standards to relate the signal observed in the mass spectrometer to the relative abundance of the observed protein. The peptides produced by the "bottom-up" approach will be analyzed by two different approaches, the accurate mass tag (AMT) mass spectrometry methodology being used at PNNL and the integrated multi-dimensional chromatography combined with MS methodology being used at ORNL. Comparison of data obtained using different approaches at ORNL and PNNL will help to validate identification of protein complexes and determine the best approach ultimately used in the Center.

Gel electrophoretic separation followed by mass spectrometry will be applied initially at ORNL to the affinity-isolated complexes. Recent literature methods [Gavin et al., 2002; Ho et al., 2002] for identifying the components of protein complexes have involved separating the components of an affinity-isolated complex using denaturing one-dimensional gel electrophoresis (1D-GE). The resulting bands, each of which represents a single protein or a small number of proteins, are excised from the gel, subjected to in-gel tryptic digestion, and the peptides are extracted and subjected to mass spectrometric analysis and database searches to identify the protein. Although this approach is rather time consuming and labor intensive, it is important to have available this technology at the outset because it is the current benchmark technique and will provide a basis for validating results from other non-gel-based approaches at ORNL and PNNL for identifying components of protein machines.

ORNL has demonstrated multi-dimensional chromatography interfaced with electrospray mass spectrometry in several different configurations, which will be applied to analysis of *R. palustris* proteins and initial *S. oneidensis* complexes. These techniques avoid the problems associated with gel-based separations. The first approach is anion exchange FPLC fractionation followed by mass spectrometric analysis of each fraction. This analysis is done either with the intact proteins ("top down") using high performance Fourier transform ion cyclotron resonance (FTICR) MS or after tryptic digestion ("bottom up") with a capillary liquid chromatography (LC) with a quadrupole ion trap MS. The analysis of the intact proteins allows modifications to be readily identified. The tryptic digestion step produces tens to hundreds of peptides (and modified peptides) from each protein. Another approach that will be evaluated is a two-dimensional LC/LC-MS/MS technique developed by Yates and co-workers [Washburn et al., 2001; Wolters et al., 2001]. Two chromatographic columns with different separation properties are connected serially and interfaced with the electrospray inlet of the Finnigan LCQ instrument. We have already demonstrated this multi-stage LC-MS technique at ORNL in the characterization of the proteome of *S. oneidensis* [Verberkmoes, 2002].

The second strategy to be applied for protein complex analysis is based upon modification of the combination of instrumental and methodological approaches being developed at PNNL for proteomics that provide broad coverage, high sensitivity and the capability for greatly increased throughput. Protein complexes are cleaved into peptide fragments (e.g. using trypsin). This complex peptide mixture is then analyzed by capillary LC combined directly with FTICR MS capabilities developed at PNNL for the characterization of proteomes, due to its capability to simultaneously achieve high sensitivity, dynamic range, mass measurement accuracies (MMA) of low to sub part-per-million (ppm), and high resolution ($>10^5$), and which provides an exceptional capability immediately applicable for the characterization of protein complexes. Given sufficient mass measurement accuracy, a peptide mass measurement can often be confidently attributed to a single protein within the constraints provided by a single genome sequence and its predicted proteome, and serve as an accurate mass tag (AMT).

As described further in **Center for Molecular and Cellular Systems Subproject**, these mass spectrometry approaches will be evaluated for the characterization of the protein complexes being generated from both *S. oneidensis* and *R. palustris*. We will also work with compuatational scientists to establish protocols for uniform data collection and analysis at both Center sites.

Advances in Mass Spectrometric Techniques for Protein Complex Analysis. A central challenge in Goal 1 of the DOE's GTL program is to establish capabilities to identify all protein complexes in an organism in a single year within the next three to five years. In the **High Throughput Molecular Characterization Subproject**, we outline a series of research tasks that will address new capabilities for the analysis of protein complexes by mass spectrometry that will help us reach the goals of the GTL program. Some of these advances will build on demonstrated technologies for characterizing complex mixtures of proteins, while others offer novel approaches to addressing the issue of high throughput for identifying the suite of protein complexes in an organism. To meet the needs of this project, the identification of protein complexes by MS-based techniques will be closely coordinated with developments in complex isolation to permit integration of these isolation and chromatographic isolation steps directly with the mass spectrometer. In addition, we will take advantage of computational tools to provide new approaches for "reducing the complexity" of the analyses by incorporating novel data interpretation and prediction techniques. These approaches, described in the **Bioinformatics and Computing Subproject**, could be used to minimize steps in the analytical protocols.

In one area of research, PNNL will focus on improving the sensitivity and dynamic range of MS measurements. These improvements will increase the throughput of measurements and also greatly reduce the sample size requirement, an important requirement for protein complexes having lower abundances. The efforts will also develop approaches for obtaining improved protein identifications, better absolute protein abundance measurements and improved assessments of data quality. These efforts will address the important area of the confidence for protein complex characterizations and the issues of data quality (e.g. of relative and absolute protein abundances) that are crucial for data use in understanding and modeling aspects of cellular systems. While initial AMT studies will employ FTICR MS for the characterization of protein complexes, efforts under **High Throughput Molecular Characterization Subproject** will also explore the utility of another instrumental approach using less expensive instrumentation based upon commercially available quadrupole-time of flight (Q-TOF) mass spectrometers that could apply the AMT approach with even greater throughput and reduced cost. The efforts in this subproject will apply artificial neural nets and statistical methods to aid the identification of proteins and also increase the quality and utility of data by e.g. using elution time information from separations more effectively and to extract more quantitative information from mass spectrometric results.

In proteomics applications of mass spectrometry, revolutionary advances are resulting from making a leap from "one-at-a-time" measurements that involve, for example, cutting individual protein spots from 2D gels, digesting, and MS analysis, to "shotgun" measurements that involve digesting an unseparated mixture of proteins, and then analyzing the resulting peptides via chromatographic separation with MS detection [Washburn, 2001; Wolters, 2001]. An individual peptide eluting from the chromatographic system and analyzed by MS (both single-stage and tandem) can often provide information to identify the protein from which that peptide originated [Eng, 1994; Mann, 1994]. Compiling this information for all peptides gives a picture of the proteome that rivals or surpasses, in both speed and quality, that obtained by the 2D gel approach. At ORNL, research will be focused to assess the feasibility of a similar "leap" for analyzing protein complexes. The initial approach to be taken in the **Center for Molecular and Cellular Systems** will be to isolate individual complexes, based on established affinity techniques–an approach will provide necessary initial Goal 1 successes. However, to progress beyond the single-complex-at-a-time scale, approaches are required for treating many complexes simultaneously, without using reagents specific to each complex. A combination of affinity-labeled chemical crosslinkers with modern proteomics techniques offers the potential for such an approach. The crosslinking reagents, which will incorporate an affinity label, will stabilize the protein complexes to allow isolation from the organism. The isolated complexes can then be digested for MS analyses, with the crosslinker attached to the pair of peptides from the associated proteins in the complex. The affinity tags will be used to isolate the crosslinked pairs from the complex mixtures of non-linked peptides and analyzed by MS. This approach will take advantage of computational tools to analyze the data and identify the proteins in the complex and potentially provide insight into how the proteins are associated in the complex. This approach is described in detail in the **High Throughput Molecular Characterization Subproject.**

It is important when designing high throughput approaches to consider opportunities for performing operations in parallel. ORNL will initiate research in the latter stages of the proposed work that will explore the use of arrays of protein affinity capture devices interfaced with electrospray mass spectrometry that could be combined with

microfabricated arrays of mass analyzers that could potentially perform mass analysis on many samples simultaneously. This approach would take advantage of complex isolation techniques being developed as part of the research program described above, including scFvs, or other "baits."

Finally, validation of data on protein complexes by comparing results from different methods will be important for the entire GTL Goal 1 effort. ORNL will investigate alternative approaches to mass spectrometric characterization of complexes, including separation of complexes preserved through anion exchange fractionation of cell lysates, and split-effluent, simultaneous high-resolution and MS/MS measurements.

1.4.3 Molecular and Cellular Imaging for Complex Characterization

Aim 2 of Goal 1 of the GTL program seeks to determine the cellular and subcellular locations of protein complexes and establish the dynamic relationships between the complexes. Imaging technologies are ideally suited for these studies because they can confirm and identify the existence and function of individual protein complexes in living cells. Further, imaging techniques can provide unique insights into the function of protein complexes. For example, information on spatial and temporal localization of complex formation, information on the formation and dissociation of protein complexes, and information on the environmental conditions that influence complex formation will be available from advanced imaging technologies. This information will be essential for defining metabolic relationships between protein complexes and for validating complexes identified by mass spectrometry-based methods.

Observing protein complexes in their native environment will be essential for describing cell function in terms of the molecular-level information obtained from genomics and proteomics techniques. Unfortunately, retaining a molecular level description, while observing on the length scale of a biological cell, presents a significant challenge. Microbes present additional challenges due to their small size, minimal organizational substructure and impervious cell walls (compared to eukaryotes). Current imaging approaches employ various optical/fluorescent microscopy techniques coupled with appropriate labeling strategies [Opus, 1999; Cook, 2002]. However, conventional microscopy techniques are limited in their ability to spatially resolve the sub-wavelength-scale complexes we seek to characterize. Fluorescence resolution elements are typically too large to provide detailed positional information.

To circumvent this limitation, we will examine several imaging technologies, including near field optics, non-linear optics and multiphoton excitation [Squier, 2001; Weiss, 1999; Mukamel, 1995]. Improvements in the design and use of fluorescent labeling strategies will also be explored. These techniques can reduce the spatial dependency imposed by the diffraction limit and increase the information content of imaging-based studies. Further augmenting these imaging techniques with information on cell state (metabolite concentration, pH, ion concentration, temperature, etc.) will also be necessary. Currently, these techniques are not commonly applied to cellular imaging studies but, with further development, could provide an unprecedented view of protein complexes and cellular metabolism within a live cell.

Clearly, new and innovative methods will be required to provide the level of detail required to validate complexes in living cells. In addition to optical based approaches for cell imaging, we will also examine scanning probe techniques, which also offer unique advantages [Engell, 2000; Allison, 2002]. The raster scanning of a molecular-scale probe can provide detailed information on the structure and location of molecular complexes. Advanced versions of scanning probe microscopes can also provide molecular identification and force measurements. Such instruments are compatible with live cell imaging and can be further combined with optical techniques. The integration of imaging modalities presents a powerful approach to characterizing protein complexes. The establishment and routine application of such tools will be pursued.

Efforts at both ORNL and PNNL will focus on developing advanced optical tools, and advanced scanning probe microscopes for characterizing protein complexes. These tools will be applied to complexes that occur intracellularly, within the cell membrane and within simpler systems that attempt to mimic the complex's natural environment. In all cases, we will strive for real-time, molecular level analyses. Our goal will be to provide

imaging tools that allow for the routine characterization of cellular events at the molecular scale. Details of this research plan are outlined in the **Molecular and Cellular Imaging Subproject.**

1.4.4 Computational Characterization of Protein Complex Structures and Functions

Computational tools are increasingly important in modern biology. These tools will be exceptionally important to this proposed project. In the near term, tools are needed to support the Center to collect, interpret and store data obtained. However, computational tools are needed to help analyze the data and provide models that can be used, ultimately, to predict the structure, function, stability, and localization of complexes in a dynamic cellular environment. In the course of the research described herein, it will be necessary to cope with the deconvolution of a huge volume of mass spectral data, integrate disparate lines of evidence informing on protein complexes, accommodate equally diverse sources of structural, proteomic and transciptomic data, and develop a predictive capacity for describing the behavior and influence of protein complexes in cellular function.

The availability of high-throughput biological data, including genomic sequence data [Lander et al., 2001; Venter et al., 2001], gene/protein expression data [DeRisi et al, 1997; Chu et al., 1998; Zhu et al., 2000], protein-protein interaction data [Fields and Song, 1989; Uetz et al., 2000; Rodi and Makowski, 1999; de Wildt et al., 2000; Ho et al., 2002], protein 3D structures (National Institute of General Medical Sciences, 1999), is gradually transforming the wet-lab based biological science into a new information-based system biology, thanks to the revolutionary breakthroughs in measurement technologies in the past 15 years. Computation is becoming an essential part of this new biology. Computers have not only made it possible to assemble the whole Human genome from millions of pieces of sequenced clones, within months' time [Venter et al., 2001], significantly speeding up the process of completing the Human genome sequence, but also led to the development of a suite of new and powerful tools for biological studies. These tools include (1) BLAST homology search program [Altschul et al., 1997], one of the most cited works in the biological literature in the past 10 years; (2) GRAIL gene finding program [Uberbacher and Mural, 1991], its creation considered by Science as one of the major milestones in Human Genome Project (Science 2001, Feb. 16; 291); (3) biological databases like Genbank for genomic sequences (http://www.ncbi.nih.gov/Genbank/), which made it possible to identify the biological function of an unknown gene within hour's time. Running computational tools and doing database searches through the Internet have become a part of the standard procedure in biological research! Driven by the goals of the GTL project and the availability of an even wider spectrum of biological data, we expect that a new generation of bioinformatic tools will soon emerge, which will play even more significant roles in understanding molecular machines in a systematic manner.

Through analysis of mass spectrometry data of chemically crosslinked proteins *in vivo*, we can derive detailed information about which proteins interact with each other in a cell and which residues of the interacting proteins are in close contact. By putting all such information together for different interacting proteins, we can possibly obtain an interaction map of all proteins in a cell. Through analyzing this interaction map, though possibly with false and missing links, we can possibly derive a collection of all the protein complexes and their component proteins. The missing and noisy data from our experiments (e.g., mass spectrometry with crosslinking) can possibly be compensated by other sources of information, e.g., protein-protein interaction data from two-hybrid systems [Fields and Song, 1989; Uetz et al., 2000]. By combining these independent sources of information, we expect to derive the interaction maps with good accuracy. Our predicted protein complex maps, coupled with the function predictions using tools like Pfam [Bateman et al., 1999], can be used as a way of making target (protein complex) selections when the Center is in full production. Experimental and computational procedures will work hand-in-hand in a cycle of prediction-validation-refinement, leading to the full characterization of the interaction map of a complex.

Mass spectral data interpretation at the genome-scale poses a number of new and challenging problems. The first is how to deal with proteins with post-translational modifications (PTMs) and mutations. Existing software capabilities are clearly inadequate in dealing with this problem [Pevzner et al., 2001]. Identification of PTMs not only is necessary for accurate identification of a protein from mass spectral data, but also provides invaluable information for understanding the biological function and functional mechanism of a protein complex. The next

generation of mass spectrometry data analysis tools will need to include both the capabilities of predicting PTMs and identifying them from the mass spectral data. The second challenge is how to identify crosslinked peptides from mass spectra, possibly coupled with PTMs. As little data is available for crosslinked proteins, our current understanding about the relationship between crosslinking and properties of involved proteins is very limited. Whereas brute-force methods (by guessing all possibilities of crosslinked peptides of all possible proteins in a whole genome) may be able to solve some crosslinking cases, thorough solution for such a problem, which a production Center like this one will need, requires some new thinking and development of novel methods.

One of the aims specified as part of Goal 1 of the GTL Program is focused on correlating information about multiprotein complexes with structural information in the NIH Protein Structure Initiative to permit understanding of the geometry and function of the protein complexes. An advantage of the combined approach of mass spectrometry and chemical crosslinking is that it provides detailed information about which parts of adjoining proteins are interacting. When coupled with 3D structure information of interacting proteins, we can possibly derive the detailed structural organization of a protein complex. Currently the PDB database [Bernsterin et al., 1977] has close to 18,000 protein structures and the PDB submission rate in recent months is about 100 structures per week (http://www.rcsb.org/pdb). As the Structural Genomics Centers (National Institute of General Medical Sciences, 1999) ramp up their productions to full speed, we are going to process significantly more proteins within the next few years. Though many of the proteins in *S. oneidensis* and *R. palusris* may not have their structures solved, we expect that a vast majority of the soluble proteins will have close or remote homologues in PDB. As estimated by G. Montelione in 2000 [Montelione et al., 2000], 50-70% of all soluble proteins have homologues in PDB. This number should have increased as the total number of proteins structures in PDB has almost doubled since the article was written. It is well established that protein homologues have similar structures and related functions. We can use the homologous structures of the component proteins to model the protein complex structure, providing detailed structure models and possibly binding sites, through which we can possibly predict the functions and functional mechanisms. There are mature computational techniques for identification of close homologs [Altschul et al., 1997; Hughey and Krogh, 1996] and modeling their structures [Sali and Blundell, 1993]. For remote homologues, threading represents the main techniques [Jones, 1999; Xu and Xu, 2000]. Though threading is not yet fully reliable, crosslinking information can provide a big help in making the structure prediction highly reliable as we have previously demonstrated [Xu et al., 2001a; Xu et al., 2001b]. ORNL has a well-established and recognized capability in threading techniques for protein structure. We will employ the PROSPECT (Xu and Xu, 2000] threading program, developed at ORNL, to accomplish data-constrained structure predictions for proteins of *S. oneidensis* and *R. palustris* with remote homologues in PDB. This combination of mass spectrometry + crosslinking and 3D structures will make a unique capability for protein complex characterization.

Much of our bioinformatics and computing efforts during the early phases of this project will necessarily be focused at establishing capabilities to support the Center core efforts with data collection, interpretation, and archival. We will place increasing emphasis on developing computational tools to meet our longer-term goal of taking full advantage of data generated from this proposed project to obtain a detailed understanding of the function of the complement of protein complexes in a cell. For example, simulation of interaction dynamics of protein complexes is planned for the out years of the project. The simulation time-scale by existing codes is generally not long enough to allow biological functional analysis and interpretation. With the data collected using the novel cellular imaging techniques to be developed in the **Molecular and Cellular Imaging Subproject**, we can significantly constrain our simulation trajectory using steered molecular dynamics [Lu and Schulten, 1999; Isralewitz et al., 2001] and hence greatly increase the simulation time-scale. The methodologies of single-molecule, real-time topological/optical imaging and spectroscopy that are developed in that subproject will provide highly useful spatial constraints for our simulation code, which could possibly lead to better understanding of the general interaction mechanisms of biological complexes, e.g., the transcription, replication, and signal transduction networks involved in many key biological functions. Functional prediction and characterization of the identified protein complexes represent another aspect of the **Bioinformatics and Computing Subproject**. Our planned research will fully utilize the derived structural information and later dynamics information, along with other publicly available information including sequence data and expression data, for functional characterization. Our strategy is to take advantage of the close teamwork between experimental and computational efforts as established as part of the Center for Molecular and Cellular Systems to

provide a wealth of information that can be used by biologists to guide their experiments, to cross-validate our predictions, and by computational biologists to mine the data that is much richer than just genomic sequence data and single structure data for new biological discoveries.

1.4.5 Impact of the Center for Molecular and Cellular Systems

The potential impact of the proposed project and the capabilities being developed as part of the Center may be best understood by considering the following example. For a microbe that reduces and sorbs metal ions (e.g., uranyl ions), *mass spectrometry* data could be obtained that could identify the key reductase enzyme, its subunits, and its partners within a complex. *Imaging techniques* could be used to locate the protein within the cell and determine whether this protein is physically/spatially linked to the sites of uranium sorption and precipitation. Computational tools could be used in conjunction with *imaging tools* and *mass spectrometry data (especially using crosslinker with known lengths* and via analysis of cysteine bonds) to provide insight into the structure of the proteins in a complex and elucidate how the proteins interact with each other so we can understand the interfacial chemistry that occurs between the protein partners. This multi-tool approach may also identify other proteins that are overexpressed in conjunction with this reductase, whether directly as part of this complex or independently with *computational tools combined with MS data.* This new protein can be identified via its sequence to a specific gene via *bioinformatics tools*, and details about post-translational and other modifications can be obtained with *mass spectrometry.* By examining the protein complexes from various states, we may be able to infer whether this protein is constitutive or tightly regulated. These data may also allow one to determine if a regulated pathway of multiple enzymes (such as benzoate degradation in *R. palustris*) exists as a protein complex or is truly a collection of "free" soluble enzymes. The tools developed in this project may ultimately allow identification of protein and protein complexes from environmental extracts of microorganisms that are difficult to culture. For example PCB degradation occurs in microorganisms that are generally difficult to culture. For bioremediation, it would be critically important to understand the pathways involved in PCB degradation and how the protein complexes that perform this degradation differ from those involved in the typical aromatic degradation. By examining other microbes, this data could allow pathways within these different organisms to be compared. Ultimately, this data would provide information required for improving microbial processes that are critical to DOE missions in energy, and the environment.

We have designed this Center to create a dynamic scientific program that anticipates long-term growth of the GTL program and can adapt as new knowledge is acquired and new biological, analytical, and computational capabilities are developed. As the Center is established and new capabilities for increasing through-put are implemented, we envision the unique capabilities of the Center will be made available as a Goal 1 resource for other GTL projects and Microbial Cell Program projects within the DOE community, as guided by DOE program managers. We have already held discussions with other laboratories submitting proposals as part of this GTL solicitation, including Lawrence Livermore National Laboratory (LLNL), Argonne National Laboratory (ANL), Los Alamos National Laboratory (LANL), who have expressed interest in collaborating with us. We have worked in concert with PNNL to help develop a Goal 3 proposal entitled "An Integrated, Multi-Level Approach for Understanding and Modeling Functional Networks in Microbial Consortia and Complex Communities." In addition, we will work with several MCP projects, specifically leveraging the federated *Shewanella* MCP's, two of which are based at ORNL and PNNL, and the *Rhodopseudomonas* MCP, in which ORNL, PNNL, and other laboratories (including Prof. Robert Tabita at Ohio State University and Prof. Caroline Harwood at the University of Iowa) are participating. A number of MGP and MCP laboratories, representing interests in *Nitrosomonas europaea, Ralstonia metallidurans, Procholorococcus marinus, Synechococcus,* and *Methanococcus maripaludis*, have also expressed their desire to interact with the Center to further the understanding of these organisms (see letters of support in the **Center for Molecular and Cellular Systems Subproject**).

We have established a set of milestones for this project and these are included in the individual subproject proposals. We have three overarching goals for the project. The first is to complete the characterization and comparison of an initial set of protein complexes in *S. oneidensis* and *R. palustris* by the end of the second year of this project, demonstrating the efficacy of our concepts. The second goal is to evaluate technical bottlenecks in the isolation and characterization of protein complexes during these initial biological studies. The third goal is to

develop and implement new technologies required for meeting the specified GTL goal of characterizing thousands of protein complexes in a single organism per year by the end of five years. A strong, dynamic research program is a critical part of the proposed project. It will allow us to develop the computational and analytical tools required to meet the long term goals of the Center. However, the research program will also develop concepts that will enable even greater capabilities in the Center that are not defined at the present time. Thus, we intend to build a synergistic relationship between the Center and these research subprojects.

1.5 Management Plan

To meet the goals of this proposed project, as described above, we propose to establish a multidisciplinary team that includes biologists, chemists, engineers, physicists, and computational scientists to conduct the biological studies outlined above. This team is composed of scientists at five DOE laboratories, ORNL, PNNL, Sandia-Livermore National Laboratory (SNL), ANL, and Brookhaven National Laboratory (BNL) and two university laboratories, the University of Utah and the University of North Carolina–Chapel Hill. As outlined below, we expect to have additional collaborators as the project progresses and new research tasks are identified to support the Center. However, during the first three years of this project, we will focus our efforts on establishing the core capabilities of the Center for Molecular and Cellular Systems.

The Center is anchored by ORNL and PNNL and both of these laboratories have expertise relevant to the outlined project, including biology expertise working with the microbial systems as part of the Microbial Cell Program and other projects [Verberkmoes, 2002]. Together these two laboratories have the widest range of analytical tools for the analysis of proteins and protein complexes within the DOE laboratory system with special world-leading capabilities and expertise in MS. Other distinguishing capabilities at these two laboratories include a wide range of imaging tools, high performance separation techniques (including pioneering work in lab-on-a-chip technologies), and extensive computational biology expertise and state-of-the-art computational facilities that are critical to the success of this project. For example, the Genome Analysis group at ORNL is actively engaged in integrating MCP data from the *Shewanella* and *Rhodopseudomonas* projects into the widening annotation of these genomes; biological insight from these, and the dozens of draft genomes processed, represents an unrivalled resource for comparative genomics and proteomics. Although not specifically described in this proposal, ORNL is also home to unique neutron scattering facilities and is establishing a small angle neutron scattering (SANS) user facility, Center for Structural Molecular Biology (CSMB), funded by DOE BER (completion in FY03). SANS is well suited for the study of protein complexes and could play a valuable role in providing additional insight into protein interactions. PNNL is home to an extensive collection of high field nuclear magnetic resonance spectrometers within the Environmental Molecular Sciences Laboratory, and these instruments will also be available to support the proposed project. Both of these techniques could provide valuable insight into the interactions between components of protein complexes once they have been identified. Collaborators from other national laboratories and university laboratories have joined this project to bring special expertise required to meet the scientific goals of this project. These collaborators include Malin Young (Sandia, mass spectrometry and computing), Carol Giometti (Argonne National Laboratory, 2D-gels), Joseph Wall (Brookhaven National Laboratory, electron microscopy), Raymond Gesteland (University of Utah, mass spectrometry), and Michael Giddings (University of North Carolina, computing related to mass spectrometry). The specific roles of all participants in this proposed project are outlined in the individual subproject descriptions. As described below, additional collaborators will contribute to this project as specific needs of the Center are defined and research programs evolve to focus on these needs.

1.5.1 Organizational Structure

An organizational structure has been devised to facilitate coordination of activities among all participants to insure meeting the goals of this project (see **Fig. 1.5.1.1**). To manage this project, Dr. Michelle Buchanan (ORNL) will serve as Scientific Director. The Director will be responsible for overall direction of the project, including setting priorities for the Center, monitoring technical progress of the activities within the Center and the various research subtasks, and overseeing financial, capital, and staff resources. She is accountable to the DOE program managers at DOE Office of Science responsible for the GTL program. A major role of the Director will

be to maintain good communication with all project participants and DOE program managers. The Director will seek advice from the Center Advisory panel (described below), which will be appointed with approval of DOE program managers. The Director will also coordinate the activities of the Center with other GTL projects and with existing Microbial Cell Program projects and facilitate communication and data sharing with the scientific community.

Deputy Directors, Dr. Frank Larimer (ORNL) and Dr. Steven Wiley (PNNL) will assist the Director with management of the Center. The Deputy Directors will oversee the daily activities of the Center at their respective institutions and coordinate these activities with the other Center site. The Deputies will assist the Director with setting priorities and tracking Center and research activities. Each Deputy will maintain close communication with his counterpart at the other Center site.

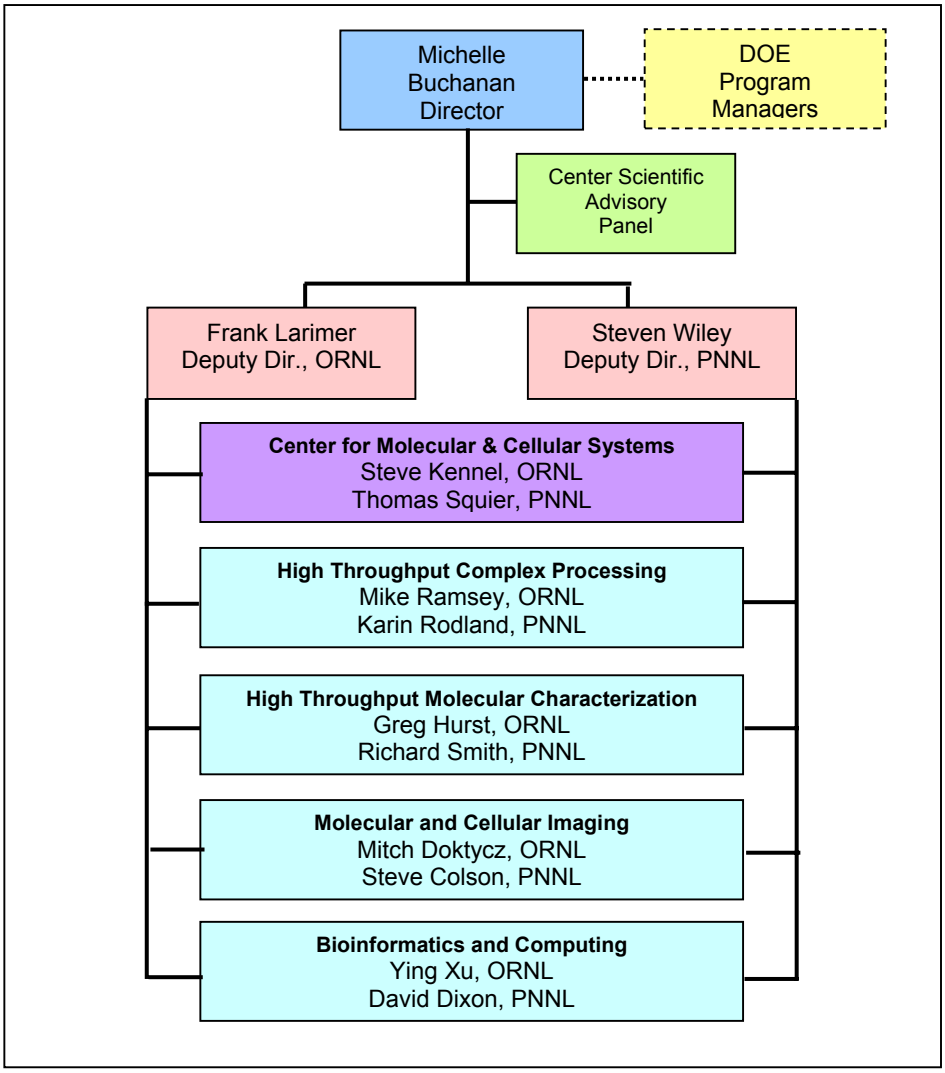

Fig. 1.5.1.1. Organizational Structure of the Proposed Project. The Director and Deputy Directors will work together to set priorities and monitor progress in all subproject areas. Each subproject will have co-team leaders, one at ORNL and one at PNNL, who will jointly oversee progress in their area. Co-Team Leaders for the four research subprojects are shown in the light blue boxes and the Co-Team Leaders for the Center are shown in the purple box.

Each subproject has assigned co-Team Leaders at ORNL and PNNL, reporting to the site Deputy Directors. The Director and two Deputies will work with the Team Leaders to establish schedules to meet the goals outlined for this project. The Team Leaders share responsibilities for oversight of activities within the subproject and will maintain regular communications with each other and with the team members. Activities at the Center will be

coordinated by close communications between the two Team Leaders to insure established goals are met, as described in the **Center for Molecular and Cellular Systems Subproject**. Team Leaders of the research subprojects will maintain regular communications with their team members, including outside collaborators, to set goals, monitor progress, and exchange ideas. The teams for each of the subprojects are listed in the individual subproject proposals. In many cases, individuals are members of both the Center and a research subproject, which further helps integrate activities between the subprojects.

Center Scientific Advisory Panel. We will establish an Advisory Panel to provide input to the overall direction of the Center for Molecular and Cellular Systems and to assess the quality of scientific program. The Advisory Panel will also assist the Director in selecting research projects to be funded in support of the Center. This Panel will meet together annually with the Center management team. In addition, individual members will participate in reviews of individual activities throughout the year and provide advice to the Director and Center participants. Nominations for the Advisory Panel members will be submitted to the Director, who will approve after consulting with DOE program managers. The term will normally be for three years and we intend to stagger the appointments so about one third of the panel rotates off annually. We have approached five individuals to ask if they would be willing to serve on our Advisory Panel if approved and these five individuals have expertise representative of each of the five subproject areas. Roger Brent (Molecular Sciences Institute, Berkeley, CA), Jim Fredrickson (PNNL, Head of MCP Shewenella Federation; microbiology), Alan Marshall (Florida State University, Head of NSF FTICR Center; mass spectrometry); John Wright (University of Wisconsin, Department Chair, optical and analytical techniques); Temple Smith (Boston University). We have received verbal agreement from all of these individuals to be considered; letters from some of the nominees are included in this section.

1.5.2 Management of Research Subprojects

The initial research projects conducted as part of this proposed project have been selected based on relevance to the established goals of the Center. Each project will be reviewed annually, for progress and continued relevance to the Center by a panel selected by the Director consisting of members of the Center Scientific Advisory Panel and other subject matter experts. Each research project will be required to submit a written report to the Director for this review. The written documents will be submitted to the DOE program managers, along with recommendations for a second year of funding. Projects will normally span a maximum of two years, although additional funding may be provided to transition the developed capability into the Center. New research activities will be initiated within these Subproject areas as Center needs are identified. Research activities within these four subprojects will be selected by a panel appointed by the Director consisting of members of the Center Scientific Advisory Panel and other subject matter experts. Recommendations for funding will be forwarded to the DOE program managers for final funding decisions. These projects will be subject to the same annual reviews as described above.

The levels of funding for the research tasks are not equal and reflect the initial priorities identified for the Center, with the largest needs (and therefore the largest budgets) associated with new techniques for the isolation and analysis of the protein complexes (activities within **High Throughput Molecular Characterization and High Throughput Complex Processing Subprojects**), which have been identified as the primary "bottlenecks" associated with the analysis of the protein complexes. Research subprojects in bioinformatics/computing and imaging technologies will be initiated at lower funding levels. However, after the initial periods of the proposed project, data on the identities of the microbial protein complexes will become increasingly available. This will enable the research in the **Bioinformatics and Computing** and **Molecular and Cellular Imaging Subprojects**, which will, in turn, provide the next level of capabilities for the Center, namely the ability to associate function with the identified protein complexes. Over the longer term, these specific subproject areas may evolve and incorporate other analytical and computational tools as new needs are identified in the Center. These research projects will include collaborators from national laboratories, universities, and private industry. We believe that establishing the research program in this manner insures that the research tasks are focused on the needs of the Center and will insure that the longer-term goals of the proposed project within the GTL program are met.

1.5.3 Communications and Outreach

Internal communications between all participants in this proposed project will be of key importance to the success of the proposed project. The Director and Deputy Directors will maintain close communication and will hold phone conferences at least weekly. Biweekly telephone (or web conferences) will be held with the Director, Deputies, and Team Leaders to discuss issues and track progress. These discussions will be especially important to insure that the Center and research Team Leaders are aware of developing needs of the Center and developing capabilities arising from the research projects. The Director, Deputies, and Team Leaders will meet twice a year, once at ORNL and once at PNNL, to review progress with teams within the Center and research subprojects at those sites. Collaborators outside of these two institutions will attend at least one of these meetings per year, normally at the site with whom they are most closely associated. To facilitate communication between team members, a secure internal website was established in January 2002 and is already being used by our team members to exchange information. We also have extensive video conference facilities available to facilitate communications between the various participants in this proposed project.

External communications will be accomplished through several mechanisms. We have established an external website (htttp://www.ornl.gov/GenomestoLife/) that will be used to communicate with the outside scientific community, including sharing of data from the Center. ORNL has extensive experience with sharing data with the outside scientific community through web-based systems like the Genome Genes and Protein Information Warehouse (GGPIW) (see the Genome Channel at http://genome.ornl.gov), which was developed to store coprehensive sequence annotations for human, mouse, and completed microbial genomes. This and other databases we have established for the biological community are described in the **Bioinformatics and Computing Subproject.** This website will also used to post highlights of recent accomplishments and abstracts from journal publications and scientific talks. We will also communicate the scientific accomplishments of the Center via presentations at scientific meetings and publications in peer-reviewed journals. We will also organize and participate in symposia and workshops that focus on issues relevant to the Genomes to Life Program.

The Human Genome Management System (HGMS) at ORNL, is an established DOE BER public outreach communication instrument that includes the website for the Genomes to Life Program (http://www.doegenomestolife.org/). HGMS is active in maintaining communications and outreach activities for the DOE OBER and we have arranged to participate in this established system as part of the Center's communication and outreach activities.

The Center is structured to train young scientists by having them contribute to research activities throughout all subprojects. In addition to post-doctoral associates that are included in each of the research projects (from the various DOE and university laboratories), we will also include students at other stages of training. For example, many of the ORNL participants in this Center are adjunct professors at the University of Tennessee (UT) and/or faculty members of the ORNL/UT Graduate Program in Genome Science and Technology. This latter graduate program trains students in biology, bioinformatics and computing, and analytical technologies and currently has 30 Ph.D. candidates enrolled. Many of these students will participate in the proposed project as part of their thesis project. Graduate students from our university partners' laboratories (as well as those from MCP collaborators) will be encouraged to spend time at the Center to train with Center staff and have access to state-of-the-art analytical and computing resources. Both ORNL and PNNL have active educational programs to bring undergraduates into research programs during academic and summer semesters. ORNL has an active program with Historically Black Colleges and Universities that provides research experience for professors and students. We will take advantage of these existing programs to train the next generation of scientists.

1.5.4 Data and Information Management Plan

Large sets of data will be generated by this proposed Center and these will be made available to the scientific community via a queryable web interface. These data will include (a) sample data and their processing-status information, (b) mass spectrometry data, (c) interaction maps of identified complexes, and (d) detailed structural and functional information of identified protein complexes. The data will be stored locally at ORNL and PNNL

and will be integrated by web applications accessible through the Center's web portal. This component of the facility will consist of (1) a sample tracking system, (2) a laboratory information management systems (LIMS) for mass spectrometry data production, and (3) a protein complex database, initially utilizing existing well-defined data models such as BIND. The database will be implemented as an Oracle database with applications developed using a combination of Perl, PHP, and Oracle 9i Application Server or SQLServer databases. Search tools will be implemented in the early phase of the project and be made accessible through the Web Interface of the Center. Interfaces will be provided to enable authorized users to extract data from the database, process it through the ORNL toolkit or the PNNL Problem Solving Environment, and optionally store the results back into the Protein Complex Database. Additionally, as completed data become available, applications will be developed for Protein Complex Database data to be submitted to or pooled with related data in community databases like BIND. Scientists mining the Center's heterogeneous data sets will require simplified access from many different computing platforms. This access will be provided through a community data portal that will integrate and deliver data and information management, analysis, and sharing capabilities through a web-based interface accessible via a standard web browser.

1.5.5 Project Initiation

When this proposed project is funded, a meeting with the Scientific Director, Deputy Directors and Team Leaders will be held at ORNL to outline a prioritized project plan prior to initiation of the project The Team Leaders will obtain input from their respective teams and their collaborators to assist in devising this plan. The project plan will outline a detailed schedule with performance measures (taking into consideration any guidance from DOE). In addition, aspects of administrative controls, procurement and subcontracts, and expenditure authorization will be established. All work will be evaluated with respect to environmental safety and health using Integrated Safety Management controls. Finally, we will establish common standards for quality assurance and records and data management. All aspects of this project plan will be transmitted to the DOE program managers for approval.

1.5.6 Performance Measurement

Performance will be evaluated against established milestones and DOE expectations. As described above, the Center Scientific Advisory Board will assess project performance annually and a written report will be provided to DOE. Programmatic progress will also be assessed by scientific peer review as required by DOE. The Director will prepare regular, quarterly progress reports to DOE, focusing on technical progress and listing relevant publications and presentations, outreach activities, activities associated with communication with the scientific community, and related areas.

Project Personnel:

Listed below are key personnel that will be initially involved in this proposed project. Technicians and other support people are not listed, nor are numerous post-doctoral associates and graduate students who will be working with our staff.

Key Staff	Primary Role/Responsibilities
Michelle V. Buchanan (ORNL)	Principle Investigator and Scientific Director of Project
Frank Larimer (ORNL)	Deputy Director at ORNL; will coordinate overall activities at ORNL
H. Steven Wiley (PNNL)	Deputy Director at PNNL; Will coordinate the overall activities at PNNL
Center for Molecular and Cellular Systems (Core)*	
Steve Kennel (ORNL)	Co-Team Leader - Center for Molecular and Cellular Systems; In charge of coordinating the isolating and analyzing protein complexes

Key Staff	Primary Role/Responsibilities
Thomas C. Squire (PNNL)	Co-Team Leader – Center for Molecular and Cellular Systems (CORE); In charge of coordinating the isolating and analyzing protein complexes
Eric J. Ackerman (PNNL)	Responsible for expression and isolation of proteins to be used as bait for complex isolation and for antigen preparation for scFv isolation
Brian Davison (ORNL)	Responsible for culturing of microorganisms under controlled conditions at ORNL
Ms. Linda Foote (ORNL)	Insertion of GST coding sequence into vectors for transfection of cells, selection and characterization of clones
Yuri A. Gorby (PNNL)	Responsible for culturing of microorganisms under controlled conditions & cell isolation/preparations at PNNL
Robert Hettich (ORNL)	Responsible for MS analysis of proteins at ORNL, FTICR of intact proteins.
Greg Hurst (ORNL)	Co-Team Leader for MS analysis of proteins at ORNL
Ms. Patricia Lankford (ORNL)	Cell lysis; protein tagging, crosslinking and prep.
Mary S. Lipton (PNNL)	Responsible for sample digestion and analysis by MS.
Denise Schmoyer (ORNL)*	Responsible for maintaining protein complex database and the sample tracking system.
Yison Wang (ORNL)	Responsible for molecular biology and construction of vectors for transfection of target bacteria
Richard C. Zangar (PNNL)	Responsible for cell fractionation and isolation of intact complexes by affinity baits or by scFvs
*Note: Not all bioinformatics and computational staff involved in establishing data acquisition, interpretation, and archival software are listed here; refer to activities below.	
High Throughput Complex Processing	
J. Michael Ramsey (ORNL)	Co-Team Leader at ORNL; Responsible for coordinating overall activities; microfluidic separation devices
Karin D. Rodland (PNNL)	Co-Team Leader at PNNL; Responsible for coordinating overall activities.
Michael Feldhaus (PNNL)	In charge of the scFv identification and screening facility. Expertise in flow cytometry and scFv screening.
Robert Foote (ORNL)	Expertise in flow cytometry and scFv screening
Robert W. Siegel (PNNL)	Responsible for affinity maturation of scFvs. Also responsible for expression and purification of identified scFvs
Cynthia Bruckner-Lea (PNNL)	Responsible for microcolumn designs, implementations, automation and evaluation efforts.
High Throughput Molecular Characterization	
Greg Hurst (ORNL)	Co-Team Leader at ORNL; Responsible for coordinating overall activities; protein crosslinking with MS
Richard D. Smith (PNNL)	Co-Team Leader at PNNL; Responsible for coordinating overall activities and AMT-based technologies

Key Staff	Primary Role/Responsibilities
Gordon Anderson (PNNL)	Software development for MS
M. Belov (PNNL)	Dynamic range and mass accuracy improvements for MS
Phil Britt (ORNL)	Synthesis of biotinylated chemical crosslinkers
Raymond Gesteland (U. Utah)	Coordination of efforts at U. Utah on MS approaches for protein complex analysis using tags
Robert Hettich (ORNL)	FTICR MS characterization of intact proteins at ORNL
L. Kangas (PNNL)	Improved quantitation and data quality
T. W. Lane (SNL)	Preparation of proteins at SNL
L. Pasa Tolic (PNNL)	Instrumentation improvements for AMT approaches
E. Strittmatter (PNNL)	Data management and informatics tools
H. Udseth (PNNL)	Improved sensitivity of electrospray
Gary Van Berkel (ORNL)	Development of ES interface for protein capture arrays
William Whitten (ORNL)	Miniaturization of ES and Mass analyzers
Malin Young (SNL)	Crosslinking with MS
Bioinformatics and Computing	
Ying Xu (ORNL)*	Co-Team Leader; Responsible for coordinating overall activities and tasks related to crosslinked protein data analysis and modified protein MS interpretaton
David A. Dixon (PNNL)*	Co-Team Leader; Responsible for coordinating overall activities at PNNL
William R. Cannon (PNNL)*	Responsible for inferring network structures, predicting ORFs and protein subunit expression
Andrey Gorin (ORNL)*	Responsible for correlating protein interactions with experimental constraints
Debbie A. Payne (PNNL)*	Responsible for the bioinformatics, database structure, data visualization infrastructure at PNNL
Nagiza Samatova (ORNL)*	Responsible for functional inference of protein complexes
Heidi J. Sophia (PNNL)*	Bioinformatics, predicting protein function, extrapolating homologies of protein families
Ed Uberbacher (ORNL)*	Protein docking and ms data interpretation algorithms
Tjerk P. Straatsma (PNNL)	Protein structure prediction. 3D simulation of binding interfaces
Dong Xu (ORNL)*	Development of understanding function of complexes and relationship to structure
Michael Giddings (U. North Carolina)	Informatics associated with MS of protein complexes
*Also involved in establishing data acquisition, interpretation and archival as part of the Center core activities	
Molecular and Cellular Imaging	
Mitchell Doktycz (ORNL)	Co-Team Leader at ORNL, Responsible for coordinating overall activities; force microscopy
Steven D. Colson (PNNL)	Co-Team Leader at PNNL, Responsible for coordinating overall activities; optical microscopy
David Allison (U. Tenn.)	AFM, modified canteliever tips
Michael Barnes (ORNL)	High sensitivity imaging; novel tags for optical and force microscopies
Gary Holtom (PNNL)	Optical imaging
Peter Hoyt (U. Tenn.)	Force, optical, electron microscopies

Key Staff	Primary Role/Responsibilities
H. Peter Lu (PNNL)	Responsible for imaging of individual complexes
Joseph Wall (BNL)	Electron microscopy

Overall Project Budget

To enhance coordination of activities and flexibility in prioritizing activities, it is anticipated that funding of all DOE Laboratory partners receiving greater than $250K over the specified period of performance will be directed to ORNL and then distributed via an Interoffice Work Order (IWO). This funding mechanism enables transfer of funds between government entities *without* incurring any subcontracting or overhead fees. Funding will be allocated only after guidance and approval from DOE Program managers. This mechanism will be particularly important to allow the research subprojects to have flexibility as the needs of the Center are defined and research emphasis areas change. It will also allow priorities in the Center to be continuously evaluated and activities shifted as deemed necessary to reflect these priorities. Funding to university partners would be direct from DOE to avoid subcontracting fees.