5.0 BIOINFORMATICS AND COMPUTING

Abstract

The goals of this subproject are

1. Identification and characterization of protein complexes in *Shewanella* and *Rhodopseudomonas* through analysis and interpretation of experimental and computational data. To accomplish this, a number of new and improved computational techniques for mass spectral data analysis and interpretation will be investigated and developed.
2. Investigation and development of a suite of computational analysis, prediction, and modeling tools which can analyze and integrate mass spectrometry data of protein complexes with other sources of information, to generate interaction maps of protein complexes and higher-resolution complex structures and their biological functions. These tools will be used for characterization of the organizational structures and biological functions of protein complexes in *S. oneidensis* and *R. palustris*.
3. Development of a problem solving environment for hosting the computational analysis and prediction tools and databases in a distributed, heterogeneous computational resource environment for mass spectral data analysis and interpretation and computational characterization of protein complexes.

We expect that the development of these computational capabilities will significantly improve our current capabilities in structural and functional characterization of protein complexes. These computational capabilities will be deployed for extensive, routine use in the proposed Center, as they become available.

ORNL:
Ying Xu
Oak Ridge National Laboratory
P.O. Box 2008, MS6480
Oak Ridge, TN 37831-6480
Phone: 865/574-7263
Facsimile: 865/241-1965
e-mail: xuy1@ornl.gov

PNNL:
David Dixon
Pacific Northwest National Laboratory
P.O. Box 999/MS K1-83
Richland, WA 99352
Phone: 509/372-4999
Facsimile: 509/375-6631
e-mail: david.dixon@pnl.gov

This page intentionally left blank.

5.2 Background and Significance

The ultimate goal of this subproject is to identify and characterize the "molecular machines of life" – the protein complexes that carry out life's essential tasks, through interpretation of experimental data and computational prediction and modeling. Protein-protein interactions and complex machines are at the heart of biological activities. They constitute the basic components of many biological processes such as signal transduction, cell-cycle control, metabolism, and general cellular machines. The function of a protein, in a cell, is generally performed through specific non-covalent interactions with other proteins by forming large stable or transient structures, referred to as *protein complexes*. To fully understand the biological machinery of a cell, it is essential to know which proteins directly interact with each other in a cell, how these protein structures are docked with each other, and how the organization of a complex relates to its biological function.

Traditionally protein complexes have been studied on an individual basis, driven by goals to understand the biological functions and detailed functional mechanisms of individual biological systems. With the advent of high-throughput analytical technologies for probing structural and functional information of bio-molecules in a cell, we believe that it is not only desirable but also feasible to characterize protein complexes in a more systematic and high-throughput mode. In this subproject, we propose to develop and apply advanced computational techniques to fully utilize the following information sources for characterization of protein complexes in a cell:

(a) information to be generated by the mass spectrometry facilities in the proposed Center,
(b) information that can be extracted from various genomic, gene/protein expression, and protein-protein interaction databases, and
(c) information of protein three-dimensional (3D) structures currently in the PDB [Bernsterin et al., 1977] and to be generated by the NIH Structural Genomics Project [National Institute of General Medical Sciences, 1999].

5.2.1 Characterization of protein-protein interactions and protein complexes

A rough estimate indicates that a genome like *S. oneidensis* (with ~4,800 genes) could have up to 18,000 protein-protein interactions, based on an estimation method [Sali, 1999]. This means that the genome may yield a few thousand stable complexes, assuming that each complex consists of 5 to 10 protein-protein interactions. In addition, complexes may exist in transient forms, where they are formed at a certain stage of a particular biological process and disintegrate after the completion of its biological function. These transient complexes may comprise a significant portion of the regulatory interactions in a cell. Although organizational information about these complexes can be obtained from mass spectrometry, chemical cross-linking and known 3D structures, functional characterization of the complexes will have to rely, to a large extent, on bioinformatic tools, in the near future. It is worth noting that ~30% of protein complexes may be membrane-associated [Jansen & Gerstein, 2000]. Because very little is known about 3D structures of membrane proteins, membrane-associated complexes will need to be dealt with differently. We will initially focus on soluble stable complexes and gradually expand our scope of investigation to transient complexes and membrane-associated complexes, as related technologies mature. Characterization of the protein complexes will be carried out at different levels, depending on the information available:

1) Identification of proteins that belong to the same complex,
2) Identification of all components of a protein complex,
3) Construction of interaction map for each identified complex,
4) Identification of the interaction interface for each pair of protein-protein interactions,
5) Construction of atomic-level complex structures, based on their component structures, and
6) Determination of cellular and biochemical functions of each identified complex.

A key in characterizing protein complexes is to identify pair-wise protein interactions (physical interactions) and the interaction interface. Protein-protein interactions typically exist in two forms: (a) protein docking, where two proteins are locked together in a form similar to glove and fingers [Gschwend et al., 1996]; and (b) "coiled coil"

interaction, where at least one helix from each of the interacting pair, wrap around each other forming a super-helical twist [Burkhard et al., 2001]. Docked proteins may either retain their native 3D conformations, called *rigid-body docking,* or have significant changes from their native conformations, called *flexible docking*. While rigid-body docking problem is generally considered to be solvable when the component structures are known, flexible docking is a much more challenging problem without additional constraints [Gschwend et al., 1996].

5.2.2 Current methods for identification of protein-protein interactions

Several techniques are in use for studying protein complexes. Two-hybrid system represents a major breakthrough in measurement technologies for genome scale biological studies, and provides information of possible protein-protein interactions in a cell [Fields & Song, 1989; Uetz et al., 2000]. Other experimental methods for studying protein-protein interactions include phage display [Rodi & Makowski, 1999], protein "chips" [de Wildt et al., 2000; MacBeath & Schreiber, 2000; Zhu et al., 2000; Reineke et al., 2001], and high-throughput mass spectrometric protein complex identification (HMS-PCI) [Ho et al., 2002]. All these emerging techniques, such as two-hybrid systems and protein arrays, have their limitations. First, the experimental conditions for protein-protein interactions may not be an accurate reflection of the *in vivo* environment where the proteins are located, e.g., two proteins may never be close enough for the interaction to take place, which could lead to false identification of interactions. Second, the proteins have to preserve their native folds while attached to the chip surface (or linked to the hybrid domain) in order to have their interactions captured; this could lead to both false positive and false negative identifications. Third, these methods generally do not provide information about which parts of the component proteins are directly interacting, and are therefore not particularly helpful in deciphering the detailed organization of a protein complex. In addition, these techniques often fail to detect low-abundance proteins from the analysis of the whole cell [Peng & Gygi, 2001].

A number of bioinformatic techniques have been developed for identification of protein-protein interactions, including gene fusion [Marcotte et al., 1999], phylogenetic profiles [Pellegrini et al., 1999], and gene context [Lathe et al., 2000]. These methods generally attempt to identify proteins that may work in the same pathways through functional associations, or that may belong to the same operons or regulons. Hence their interactions could be predicted either physical or genetic. Protein docking prediction is a computational technique for determining whether and how two interacting proteins are physically docked, based on the geometry- and chemistry-complementarity of the interaction interface [Kuntz et al., 1982; Vakser et al., 1994; Fischer et al., 1995]. It provides a highly useful tool for constructing a complex structure when its component structures are known (either through experiments or computational predictions), particularly so when coupled with experimental data. In this subproject, one of the focus areas is the development of new computational methods for the structural characterization of protein complexes that can fully utilize available experimental data.

5.2.3 Study of protein complexes using mass spectrometry and cross-linking

Mass spectrometry has been used to identify components of protein complexes. Mann and co-workers have published several examples of isolating complexes based on affinity techniques [Pandey & Mann, 2000]. Recent work [Ho et al., 2002] has demonstrated the feasibility of large-scale characterization of protein complexes, using mass spectrometry and "bait" techniques. In this published work, the authors used 725 bait proteins and identified 3617 interactions involving 1578 different proteins. In addition to simply identifying component proteins and their interaction map, it is also possible to obtain information regarding the detailed geometries of a complex by partially cross-linking the complex [Rappsilber et al., 2000]. In other words, the fact that two peptides resulting from a proteolytic digest are cross-linked suggests that the regions of the two interacting proteins from which the respective peptides were cleaved must be neighbors in the complex.

Chemical cross-linking is a widely used technique in protein biochemistry, with multiple applications ranging from tethering proteins to surfaces and labeling proteins to studying a protein's three-dimensional configurations and near-neighbor relationships in protein complexes [Wong, 1991]. Chemical cross-linking reagents contain a central spacer arm connecting two reactive groups. Each reactive group is specific to a particular site in a protein, such as the side-chain amine of lysine or the thiol of cysteine. Structural information about proteins can be

obtained by determining the regions of the protein that are cross-linked [Cohen & Sternberg, 1980; Young et al., 2000]. This approach inserts "rulers" into the protein, providing information regarding distances between pairs of amino acids in the 3D structure. Similarly, by determining the regions of two proteins in a complex that are cross-linked, one can get information about the identities of the proteins, as well as the rough geometry of their interaction in the complex, particularly when the individual structure of each protein is known.

Methods based on a combination of mass spectroscopy and cross-linking techniques (MS/CL) overcome some of the limitations described in Section 5.2.2, and create a very attractive avenue for identifying protein-protein interactions. Cross-linking generally takes places in the naturally existing associations, including transient ones. Under reaction conditions that limit the number of cross-links introduced, the method is unlikely to distort the fold conformations of linked proteins. More importantly, the exact location of the cross-linking sites can be derived through the analysis of mass spectrometry data of cross-linked proteins [Rappsilber et al., 2000; Itoh et al., 2001]. This makes it possible to infer/extract detailed spatial characterization.

Using information derived from high-throughput proteomic efforts, a number of protein-protein interaction databases have been created. Among them is the popular Biomolecular Interaction Network Database (BIND) [Bader et al., 2001; http://binddb. org], which is designed to store descriptions of interactions, molecular complexes and pathways. BIND incorporates components of molecular mechanisms including interactions between two objects, where each object can be a protein, DNA, RNA, ligand, molecular complex or an interaction. The interaction description includes subcellular localization, experimental conditions used to observe the interaction, conserved sequence, chemical action, kinetics, thermodynamics, and chemical state. In the early phase of the subproject, data models of BIND will be carefully studied to determine the suitability for our data storage/management/dissemination needs and may be adopted, as appropriate.

5.2.4 From experimental data to complex structures/functions: our vision

Using well-designed cross-linkers, the MS/CL method can provide the information regarding residues of adjoining proteins of a complex. This information, coupled with the 3D structures of individual component proteins, can possibly lead to a detailed and accurate characterization of structural organization of a complex, using computational techniques like data-constrained docking (see **Research Design and Methods**). Such structural information can provide significant clues about the molecular- and cellular-level functions of protein complexes and even their functional mechanisms. The prerequisites for such detailed characterization of structural organization are the 3D structures of the component proteins. The PDB database currently has ~17,600 x-ray crystallographic or NMR protein structures (release of March 2002), and the current rate of submission to the PDB is about 100 structures per week (http://www.rcsb.org/pdb). This rate is expected to go up significantly as the NIH Structural Genomics Centers ramp up their structure productions [National Institute of General Medical Sciences, 1999]. Although we do not expect to have the 3D experimental structures for most of the proteins in *S. oneidensis* and *R. palustris* for at least a few years, we do expect that a vast majority of the soluble proteins in those two genomes (and other microbial genomes) will have either close or remote homologs in the PDB. It is estimated that a soluble protein has about 50 – 70% probability of having a close or remote homolog in the PDB [Montelione et al., 2000]. Our recent experience has indicated that this number is even higher as the size of the PDB has more than doubled since the article was written. These homologs could be identified through either sequence-based search methods like PSI-BLAST [Altschul et al., 1997] or structure-based search methods like protein threading [Jones, 1999; Xu & Xu, 2000]. The true value of MS/CL method is that it does not just provide the information about which proteins are physically interacting, but that it also provides us with a framework for exploiting the structural information, generated by the Structural Genomics Projects, for more detailed characterization of protein complexes.

It is foreseeable that the MS/CL method, especially in the early phase of the project, may not always detect adjoining proteins or may provide false cross-linking data, resulting in incomplete coverage of a complex or false linking of multiple complexes into one, or combinations of both. We believe that other data sources can provide additional information to resolve the missed and false identifications of cross-linked proteins/residues. Genome-level protein interaction maps, based on two-hybrid system, have been constructed and are publicly available for

the yeast genome [Fields & Song, 1989; Uetz et al., 2000]. Through homology search, parts of the interaction maps could be used as possible templates when constructing interaction maps in *S. oneidensis* and *R. palustris.* Protein-protein interactions can also be predicted using other (independent) approaches, e.g. through predictions of operon structures [Stephanopoulos et al., 1998], analysis of micro-array gene expression data [Jansen et al., 2002], or gene fusion [Marcotte et al., 1999]. Each of these techniques has its strengths and limitations. One challenging problem in this subproject (see **Research Design and Methods**) is to develop computational techniques for most effectively combining these prediction results for derivation of interaction maps that are most "consistent" with the available data.

The functional annotation of the identified protein complexes will initially focus on functional inference of individual proteins and inferring possible functions of each individual complex, based on functions of its component proteins. As our technology for structural characterization of protein complexes improves and more data, like gene/protein expression data and pathway information, become available, we will shift our focus to derivation of the functionality of a complex as a system, in the context of regulatory networks and metabolic pathways.

Validation of our computational results through rational design of experiments will be an important part of this Center project. We will work closely with the experimentalists in the proposed Center in the rational design of experiments and interpretation of experiment results.

As outlined in the Overview to this proposal, the projects described are the initial research projects being proposed to support the Center for Molecular and Cellular Systems. These projects will be reviewed annually as outlined in the Management Overview for progress and continued relevance to the Center. Normally no project will extend beyond two years. It should be noted that techniques developed as part of this project will be incorporated into Center sites at both ORNL and PNNL.

5.3 Preliminary Studies

This ORNL/PNNL bioinformatics team and our collaborators at SNL have extensive experience and strong track records in the analysis/interpretation of mass spectrometry data, computational prediction and characterization of protein/complex structures and functions, construction and management of large biological databases, and high performance (HP) computing for biological applications. Our achievements, in past ten years, include hundreds of scientific publications, a number of R&D 100 awards for the development of bio-software/databases, and a long list of software packages for biological applications. The combined computing environment at ORNL, PNNL and SNL is among the best in the world. We have highlighted several studies, closely relevant to this proposed project to illustrate our general capabilities.

5.3.1 Mass spectrometry data analysis capabilities

The overall capabilities in mass spectrometry for proteins at PNNL and ORNL are described in the **High Throughput Molecular Characterization** subproject of this Center proposal. Here we provide one recent development, directly related to our data analysis capabilities.

5.3.1.1 Data interpretation for cross-linked protein

In a recent experiment, bovine ribonuclease A was crosslinked using the reagent bis(sulfosuccinimidyl) suberate (BS^3) and subjected to tryptic digestion. A crosslinked pair of peptides was tentatively identified by a signal at m/z 4650.04 in the MALDI-TOF mass spectrum of the digest from the crosslinked protein. Subsequent electrospray MS-MS experiments on an ion trap mass spectrometer yielded sufficient partial sequence information to confirm that a crosslink had been inserted between lysines 37 and 104, and further that the crosslinked peptides were Asn34-Arg39 and Tyr92-Val124, as illustrated in **Fig. 5.3.1**. The triply-charged ion from the crosslinked peptide was fragmented, yielding an MS-MS spectrum showing y_8 (m/z 871.3), y_{11} (m/z 1230.5), and y_{14} (m/z 1530.7) ions [Roepstorff, 1984], corresponding to fragments containing the C-terminus of

the larger Tyr92-Val124 peptide. An ion at m/z 1951.8 (+2 charge state) marked 92-124+XL is consistent with the large peptide plus the crosslinker; this ion results from fragmentation at the amide bond connecting the crosslinker and the ε-amine on the internal lysine of the smaller peptide. An ion at m/z 1870.3 (+2 charge state) represents the fragment labeled y_{32}+XL in **Fig. 5.3.1**, corresponding to loss of the N-terminal tyrosine from the 92-124+XL ion. This y_{32}+XL ion was isolated in the ion trap and further fragmented to yield the MS-MS-MS spectrum shown in **Fig. 5.3.1**. The MS-MS-MS spectrum shows extensive y- and b-ion series [Roepstorff, 1984] that confirm the amino acid sequence of the large peptide fragment. In this example, it was not possible to observe sequence-informative ions from the smaller Asn34-Arg39 peptide of the crosslinked pair due to its relatively small size. Nevertheless, its presence could be inferred from the intact m/z measured for the parent ion, and by characteristic mass shifts observed in sequence ion series from the larger peptide. Recent examples from the literature indicate that it is possible to obtain sequence information from each of the two crosslinked peptides [Chen, 2001; Pearson, 2002].

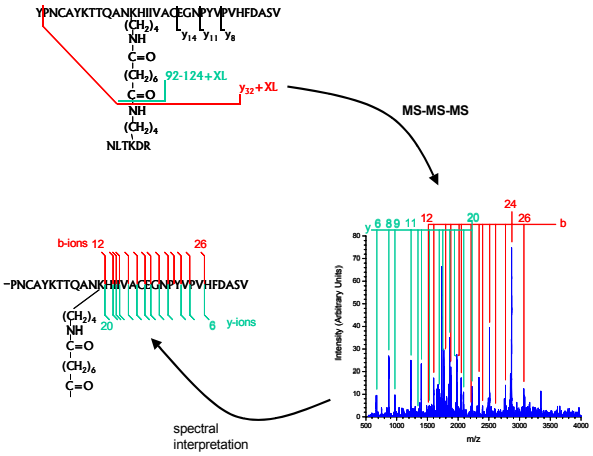

Fig. 5.3.1. Tandem mass spectrometry of a crosslinked peptide pair. *Top*: fragments observed in MS-MS spectrum. *Lower right*: MS-MS-MS spectrum from y_{32}+ XL ion. *Lower left*: Fragments observed in MS-MS-MS spectrum.

This example illustrates several points relevant to using mass spectrometry to characterize crosslinked proteins. The MS-MS fragmentation of crosslinked peptides appears similar to that observed in non-crosslinked peptides; the amide bonds between the amino acid residues do fragment to yield sequence information. However, amide bonds between the crosslinker and the peptides also appear to fragment. Rules for interpreting tandem mass spectra of peptides are therefore still valid for crosslinked peptides. However, the interpretation is more complex, and will require the development of sophisticated computational tools if high throughput is to be achieved.

5.3.2 Computational inference of protein complexes

5.3.2.1 Protein and complex structure prediction by PROSPECT and other tools

PROSPECT (PROtein Structure Prediction and Evaluation Computer Toolkit) [Xu & Xu, 2000; Xu et al., 1998a, Xu et al., 1998b; http//compbio.ornl.gov/structure/prospect/] is a computer software system that we have developed for protein structure prediction at ORNL. For each structure prediction, PROSPECT provides a confidence value, based on which a user can decide if to trust the prediction or not [Xu et al., 2002]. In CASP4, PROSPECT's performance was ranked the sixth out of 123 participating teams in the category of protein fold recognition [Xu et al., 2001c]. We have recently implemented an automated structure prediction pipeline, based on PROSPECT and other prediction and analysis tools, on the IBM/SP 3 supercomputer at ORNL, and we have made structure predictions for 750 soluble proteins of *R. palustris* (unpublished results).

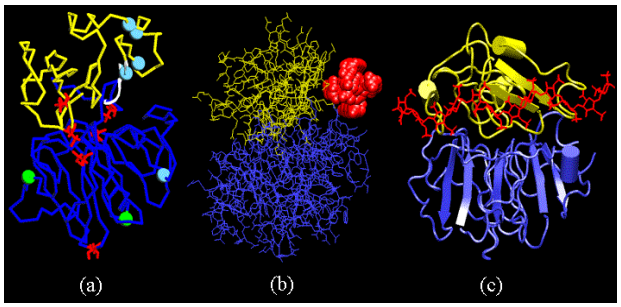

Fig. 5.3.2. (a) the docked structure between the central (blue) and C-terminal (yellow) domains. (b) and (c) show the predicted docking conformation between the heparin (in red) and the combined structure of the central (blue) and C-terminal (yellow) domains from two different perspectives.

We now highlight one prediction work we did on the protein vitronectin through collaboration with Dr. Cynthia Peterson at the University of Tennessee [Xu et al., 2001a]. Vitronectin is a multifunctional glycoprotein found in blood and in the extracellular matrix. No structure has been solved for vitronectin or any of its close homologues. Previous studies have suggested that the 459-residue protein contains three structural domains and a long unstructured linker between the N-terminal and central domains. Fold recognition and sequence-structure threading were performed using PROSPECT for each of three structural domains, i.e., the N-terminal, the central, and the C-terminal heparin-binding domains. The atomic structure of each domain was generated using MODELLER [Sali & Blundell, 1993] based on the alignment obtained from threading. Docking experiments between the central and C-terminal domains were conducted using the program GRAMM [Vakser & Aflalo, 1994], with constraints from a known inter-domain disulfide bridge. The docked structure has a large inter-domain contact surface and defines a putative heparin-binding groove (see **Fig. 5.3.2 (a)**). We also docked heparin together with the combined structure of the central and C-terminal domains using GRAMM (see **Fig. 5.3.2 (b,c)**). The predictions from the threading and docking experiments are consistent with the experimental data on purified plasma vitronectin pertaining to protease sensitivity, ligand-binding sites, and buried cysteines.

5.3.2.2 Identification of operons/regulons

Interacting proteins are often co-expressed from a single mRNA. In this case, they are located adjacent to each other and in the same direction in microbial chromosomes in “operons”. Sets of operons are often coordinately expressed in “regulons” by means of shared control elements in the DNA that bind particular regulator proteins. We have developed computational/information technology, using Java, to support the biologist in extracting potential interactions between elements of the molecular machinery that are encoded in the genome at multiple levels. Our strategy is based upon organizing large-scale genomic information in such a way that a context is provided for visualizing patterns that enhance the ability of the biologist to detect them.
We now describe an application of our technique that is a logical extension of the Similarity Box analysis [Sofia et al., 2001] to the question of operons. We have performed a large-scale operon analysis based on the newly characterized Radical SAM superfamily, which now consists of over 1100 proteins from all three divisions of life. We extracted 5092 putative operon sequences from the microbial genome tables at NCBI based on the 858 Radical SAM proteins in 55 completed microbial genomes using the approximation that operon proteins are adjacent and encoded on the same strand. To examine the major patterns, we used a global BLAST comparison to extract those proteins that occur in the overall set 10 or more times at an Expect value better than 10^{-4}. We then applied hierarchical clustering (complete linkage, Euclidean distance) to these 788 putative Radical SAM operon proteins, and produced a Similarity Box visualization of the results to quickly identify the diverse protein groups in the overall set. There are 47 different types of operon proteins associated with Radical SAM proteins that occur at least 10 times. Recurring operon patterns become easily detectable when these protein classifications are associated with the putative operon structures.

5.3.2.3 Characterization of amino acid/peptide transport pathway

Amino acid and peptide transport in yeast *S. cerevisiae* occurs through a number of transport proteins, including Gap1p, Agp1p, and Ptr2p [Island et al., 1991]. By sensing extracellular conditions of amino acids and peptides, receptors on the cytoplasmic membrane transduce signals to intracellular molecules. Among the receptors, Ptr3p plays a crucial role as a switch for regulating expression of the di/tri-peptide transporter [Barnes et al., 1998; Klasson et al., 1999]. It is suggested that a signal transduction pathway will be activated between Ptr3 and the transcription factors of the amino acid and peptide transporters. We have conducted computational prediction of the protein structures and the pathways, through collaboration with an experimentalist, Dr. Jeffery Becker at the University of Tennessee.

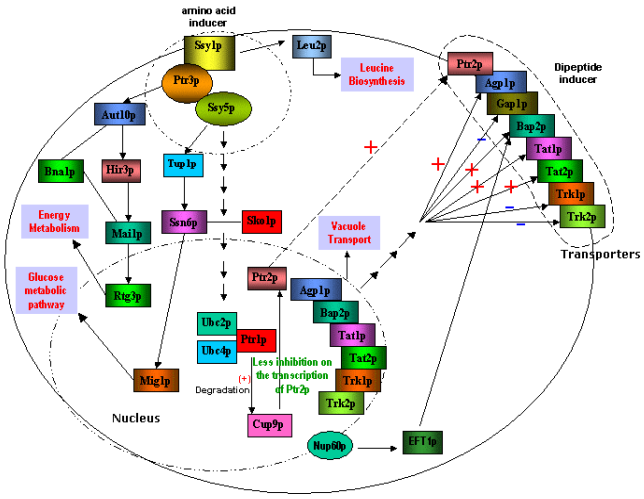

Fig. 5.3.3. A pathway model for peptide transport.

We first made a structure prediction of Ptr3p. Ptr3p does not have a close homolog in the PDB. Our prediction, using PROSPECT, shows that the 678-residue protein has two major domains, the N-terminal domain and the C-terminal domain. Our structure model for the N-terminal domain contains a zinc finger motif, supporting our prediction that Ptr3p has a nuclear sub-cellular localization (Becker's lab is carrying out experiments to verify this prediction). Our prediction for the C-terminal domain is that it shares the structural folds of 1qqg (PH-PTB targeting region of IRS-1) and 1tbg (the heterotrimeric G-protein transducin), both being a seven-bladed β-propeller and being signal transduction proteins. Based on this prediction, we hypothesize that the C-terminal domain may be involved in a protein-protein interaction since many of the β-propeller structures have known functions in governing protein-protein interactions [Beisel et al., 1999; Wall et al., 1998].

We have then constructed an interaction map for the Ssy1p-Ptr3p-Ssy5p complex and transcription factors that control proteins in the related pathways, using information including data of DIP [Xenarios et al., 2002], BIND [Ho et al., 2002] and gene expression data [Forsberg et al., 2001; Zhu et al., 2000]. We have also constructed the pathways connecting the complex and the glucose metabolic pathway as well as the energy metabolism pathway, as shown in **Fig. 5.3.3**. We found that Ssy5p interacts with Tup1p, which is a transcription factor. Tup1p works with several other transcription factors, including Ssn6p, which activate Mig1p. Mig1p is known to be the repressor for several proteins in the glucose metabolic pathway, including Suc1p, Suc2p, Suc4p, Cyc1p, and Ena1p, all of which share similar gene expression profiles and binding motifs. This pathway model is in agreement with the observation that Ptr3p induces the amino acid/peptide transport pathway while it represses the glucose metabolic pathway.

5.3.3 Bio-computing at ORNL

5.3.3.1 Biological databases

We have substantial experience in the development and management of large biological databases. The Genome Genes and Protein Information Warehouse (GGPIW) was developed to store comprehensive sequence annotations for human, mouse, and completed microbial genomes. The role of the GGPIW is to support a consistent annotation process that uses ORNL's Genome Analysis Pipeline and Genome Analysis Toolkit and Server for generating up-to-date annotations. The GGPIW is based on the sequences available from NCBI and includes the NCBI annotation as well as the computationally derived annotations from various gene-modeling tools. The database is implemented using Oracle 8i and contains over 29 million records. GGPIW data is publicly accessible through the Genome Channel (http://genome.ornl.gov) and the ORNL DAS server (http://genome.ornl.gov/das).

MuTrack is an electronic data system for collaborating laboratories in the Tennessee Mouse Genome Consortium (TMGC) (http://www.tnmouse.org/mutrack/index.php) we have recently developed, which helps large, geographically distributed teams to collaborate and achieve synergistic efficiencies. The MuTrack database currently tracks mice as they move throughout the TMGC institutions and labs, helps in automating processes, uses statistics to highlight "interesting" data for the attention of researchers, and generates periodic task reports for the researchers and technicians. We have also developed sample-tracking databases. The Genosensor Information Management System (GIMS) was developed to track DNA clones on ORNL microarrays used in mRNA expression analysis. GIMS supports cDNA processing and microarray printing.

5.3.3.2 High-performance computing for biological applications

The genomic Integrated Supercomputing Toolkit (GIST) is a toolbox for distributed computing in a heterogeneous computing environment, developed at ORNL. The computational infrastructure consists of the centerpiece IBM SP3, several SGI SMP machines, a DEC Alpha Workstation cluster, and a Linux PC cluster. The initial tools in GIST provide a framework of high performance biological application servers including massively parallel BLAST codes (versions of BLASTN, BLASTP, and BLASTX), Pfam [Bateman et al., 1999], and PROSPECT [Xu & Xu, 2000]. The GIST resources are extensively used in the genome annotation at ORNL. It runs in a transparent fashion, permitting the gradual introduction of new algorithms and tools, without jeopardizing existing operations. Due to the logical decoupling of the query infrastructure, an infrastructure with both excellent scaling abilities and many fault-tolerant characteristics has been produced.

5.3.4 High-performance computing for molecular science at PNNL

PNNL has several ongoing efforts in developing Problem Solving Environments (PSEs), collaboratories, and large-scale data management systems. These efforts focus on meeting the needs of specific scientific and engineering domains and we have developed systems tailored for the specific community to address their specific requirements for security, computation, and data scaling. Over the past five years, we have also been engaged in an effort to develop generalized components for use in developing PSEs in new domains. This research provides us with a basic platform on which to build new software technologies.

A multidisciplinary team of scientists and computer experts at PNNL's Environmental Molecular Sciences Laboratory (EMSL) developed the Molecular Sciences Software Suite, MS^3. MS^3 is an integrated suite of comprehensive software that enables scientists to study complex chemical systems by coupling advanced computational chemistry techniques with high-performance, parallel computing systems. It enables modeling and simulations of environmentally relevant chemistry, with models more complex and more accurate than heretofore possible. The MS^3 consists of three components: 1) the Extensible Computational Chemistry Environment (*Ecce*), a domain-encompassing PSE composed of a suite of tools (Jones et al., 1999); 2) the Northwest Computational Chemistry Software (*NWChem*), advanced computational chemistry techniques; and 3) Parallel Software Development Tools (*ParSoft*), efficient and portable libraries and tools that enable NWChem to run on a wide variety of parallel computing systems. Among them, Ecce assists chemists with many tasks, including the management of projects and calculations, construction of complex molecules and basis sets, generation of input decks, distributed execution of computational models, real-time monitoring, and post-run analysis.

5.4 Research Design and Methods

The overall goal of this bioinformatic subproject is to characterize the structural organization and biological functions of targeted protein complexes in *S. oneidensis* and *R. palustris*, through investigation, development, implementation and applications of advanced computational techniques. The focus areas of the R&D work will initially be

(1) investigation, development and application of highly effective computational techniques for interpretation and analysis of mass spectrometry data of protein complexes, which can handle proteins with post-translational modifications (PTMs) and mutations, chemical cross-links, and novel proteins;

(2) investigation and development of highly effective and efficient computational techniques for characterization of protein complexes, based on interpreted mass spectrometry data and information extracted from other sources; and applications of these tools to a selected set of complexes in *S. oneidensis* and *R. palustris*, as a proof of principle;
(3) development of a computational problem-solving environment to host the computational tools listed in (1)-(2) on a collection of heterogeneous computing facilities and to support access to all our databases through a unified interface. These computational tools will be designed and implemented with flexibility, scalability and long-term stability in mind.

To achieve these goals, we will implement the following three tasks.

5.4.1 Task 1: Mass spectrometry data analysis and interpretation

Our capability for interpreting mass spectral data for identification of proteins and protein-protein interactions on a genome-scale is essential for the success of the proposed Center. The primary goal of this task is to investigate and develop new and improved analysis and interpretation capabilities for mass spectrometry data of protein complexes, for genome-scale protein complex characterization. There are a number of existing computer software systems for both peptide identification using MS/MS data and protein identification using MS data including SEQUEST [Eng et al. 1994; http://fields.scripps.edu/sequest/], PepFrag (http://prowl.rockefeller.edu/), ProteinProspector (http://prospector.ucsf.edu/), and GPMAW (http://www.welcome.to/gpmaw), ProFound [Zhang & Chait, 2000; http://prowl.rockefeller.edu/] and many others. Our initial data analysis and interpretation package will consist of these tools (see **Molecular and Cellular Imaging subproject**). The new and improved computational tools will be added to this package as they become available. The ultimate outcome of this effort will be to significantly enhance the current capability of

(1) interpreting mass spectral data of proteins that are possibly linked by chemical cross-linkers,
(2) interpreting mass spectral data of proteins with PTMs and mutations, and
(3) interpreting mass spectral data of novel proteins.

5.4.1.1 Mass spectrometry data analysis of cross-linked proteins

The goal of this sub-task is to develop and implement new algorithms for identification of cross-linked proteins and precise identification of residues that are cross-linked. A peak in a single stage mass spectrum could represent one of four possibilities: (a) the mass (actually, mass-charge ratio) of a peptide without any cross-linkers, (b) a peptide with cross-linkers attached to it with their other ends unattached due to hydrolysis, (c) two peptides linked by one cross-linker, and (d) multiple peptides with multiple cross-linkers linking them together. We will initially consider only the first three possibilities. Since the fourth scenario has a relatively low probability of occurring, it will be dealt with in a later phase of the subproject.

The first step will be to identify which peaks in mass spectra represent peptides linked or labeled by cross-linkers. For this purpose, we plan to do LC-MS-MS experiments, so that there will be many mass spectra from a given crosslinking experiment, since the LC separation tends to vary from sample to sample. Comparing these spectra against uncross-linked control spectra will lead to the identification of cross-linking information. Peaks that correspond to species with multiple charge states in the MS spectra, that appear in more than one consecutive LC scan, and that have a recognizable isotope pattern will be carried forward for further analysis. After identifying these candidate peaks representing cross-linked or labeled peptides, we will then begin the assignment procedure. Essentially, assignment of the peaks involves determining what peptide(s) linked by a cross-linker have a total mass consistent with the observed peak. This can possibly be done using two different approaches, borrowing from the standard techniques for interpreting (non-crosslinked) protein mass spectral data: (1) through database search to identify a pair of peptides whose combined mass, plus the mass of a cross-linker, is consistent with the spectral peak; or (2) through identifying each single amino acid of the peptide from the mass spectral data, referred to as the *de novo* sequencing method [Dancik et al., 1999].

As we are interested in dealing with proteins of genomes with complete sequences (e.g., *S. oneidensis*, *R. palustris)*, our proposed method will mainly rely on database search techniques with some information incorporated from *de novo* sequence methods. For simplicity of discussion, we assume that our cross-linker is designed to link to lysines. We will construct a database of all Lys-containing sequence-unique peptides (given a protease and a maximum number of missed cleavages) of all the predicted protein sequences of our target genome (e.g., *S. oneidensis*, *R. palustris)*. We first find all labeled and intra-molecularly cross-linked pairs of peptides whose total mass (plus the mass of cross-linker) is consistent with the spectral peak within a given ppm error. This set contains all possible peptides to be considered in the next step. Then we will score each peptide and peptide pair based on its consistency with the tandem mass spectral (MS-MS) data. Initially, we will test to see if current programs such as SEQUEST [Eng et al., 1994], Polygraph (under development at PNNL) or MS2Assign (developed at SNL) can be adapted for these purposes. If not, a scoring scheme will be developed based on the consistency information and also taking into account the possibilities of (a) noise in the observed spectrum and (b) non-unique matches of some mass with the peptide-pair. Ideas from *de novo* sequencing algorithms, like matching against a spectrum graph [Dancik et al., 1999], will also be used in developing the scoring scheme. We will use mass spectral data of known cross-linked peptides to test, validate and refine our scoring scheme, and also develop a technique for match-score threshold determination for identification of the peptide pair. The pairs with match scores above the selected threshold will be considered as identified peptides. It is possible that a cross-linker may link two residues of the same protein. Such intra-protein cross-links will be removed from further consideration.

We estimate the computational complexity of our method to be $O(N^2 S)$ for each spectral peak of a possible cross-linked peptide, where N is the number of peptides in our database and S is the computing time to score each pair against the tandem spectral data as discussed above. The actual compute time could be much smaller than this since we only need to search all peptides with their masses no larger than the mass indicated by the spectral peak.

5.4.1.2 Identification of PTMs/mutations

Post-translational modifications (PTMs) play an important role in stabilizing protein structure and controlling intracellular protein function [Graves et al., 1994]. Over 400 types of PTMs have been identified, up to this date, as constituents of different proteins in different organisms [Krishna & Wold, 1998, and see http://prowl.rockefeller.edu/aainfo/deltamassv2.html]. As PTMs change the mass of the amino acids, they substantially complicate the interpretation process of mass spectral data. Although PTMs in microbes (GTL's focus) are less common than in eukaryotes, it is crucial to identify PTMs as they provide highly useful information for studying protein functions and biological pathways.

While identification of PTMS remains a highly challenging and unsolved problem in general, some *ad hoc* methods have been developed. For example, FindMod [Wilkins et al., 1999; http://www.expasy.org/tools/findmod/] is a program for *de novo* discovery of protein PTMs. It attempts to identify 22 most common types of PTMs with a reasonable level of accuracy, including acetylation and biotinylation. Pevzner and co-workers have developed an interesting framework for identifying PTMs, called *spectral alignment* [Pevzner et al., 2001]. Peaks in the spectra are represented as elements in a matrix, and alignments between the two spectra are paths taken through this matrix. Using the technique of dynamic programming [Smith & Waterman, 1981], one can find the best match between two spectra, given some number of allowed mismatches caused by PTMs. The limitation of this approach is that it requires that the number of PTMs be small.

In the initial phase of the subproject, we will implement this spectral alignment algorithm, using *de novo* sequencing information, as our first attempt to deal with PTMs. One of the extensions we will make is that we will incorporate predicted PTM information, using existing programs such as NetPhos [Blom et al., 1999], into the scoring scheme of the spectral alignment algorithm. This should increase the sensitivity and specificity of the prediction capability of PTMs. We will try to find the optimal weights for the predicted PTMs in this scheme, based on known PTMs in known proteins. Because the specific information about PTMs will be incorporated, we expect that this generalized spectral-alignment algorithm will fare better in dealing with proteins with higher

number of PTMs. In addition, algorithmic improvements will be implemented to make the algorithm applicable to a broad range of protein identification problems with PTMs.

A significant challenge facing proteomic analysis is the integration of multiple data sources into a coherent picture of a protein's functional role in the cell. A new paradigm is required that can integrate multiple measurements of a protein to provide a more wholistic picture, not only of the gene from whence the protein derives, but of the full set of mechanisms by which it came to be in its *in vivo* state. We have developed a "data-driven" proteomics approach that attempts to address this challenge, based on a concept of iterative refinement, using multiple complementary measurements made on a protein or group of proteins to obtain a statistically robust picture of a protein's character *in vivo.* It will combine the multiple measurements using an agent-based bioinformatics backbone titled the protein inference engine, or PIE.

A concrete example might serve, based upon the approach used in the Gesteland/Atkins labs. The identification step consists of using data from tryptic-digestion of liquid chromatography (LC) separated proteins through peptide-mass fingerprinting, sometimes with the addition of tandem MS analysis. This allows the drilling down to a specific gene identification. Another measurement taken is that of the full-size protein, obtained via ESI-MS. This is then used on the outward projection step to more fully characterize the protein, in particular, looking for common post-translational modifications that may have occurred.

The specific technologies that have been developed to-date for these purposes include genome fingerprint scanning (GFS) and the ProteomeMods software. The GFS works on similar principles to other peptide-mass fingerprinting approaches, with the difference that it is designed to scan the raw, uninterpreted genomic data, to determine a genomic origin based on location from whence the protein is derived. Though GFS works well by itself, it is clear that combining this with another identification method such as tandem MS will provide an improvement in the accuracy of the results. This highlights a key to the data driven approach, which is taking as many measurements as are reasonable and using them all to fill out the picture of a protein in its native environment.

The ProteomeMods software is a second-phase component, designed to help determine whether any common post-translational modifications are present on a protein. Eventually this will be driven directly from genomic data; currently it starts with a set of ORF's defined by peptide mass fingerprinting as likely candidates for producing the protein observed. Starting with these, it searches a depth-first tree, considering common post-translational modifications. It also considers possible proteolytic cleavages. These tests are matched against measured full size masses. The tree-search is theoretically infinite, necessitating a means of bounding the search heuristically. A simple fuzzy-logic based system for considers at each node in the tree whether to continue down that branch or to backtrack. Currently the bounds are based on depth-in-tree, delta-mass between target and prediction, and likelihood of modification. More intelligent rules involving domain knowledge present in the literature, will be incorporated, e.g., common phosphorylation motifs or transit-tag predictions.

During the term of this project, an approach similar to the ProteomeMods program will be usede as the basis for our protein inference engine (PIE). The PIE will consist of multiple "agents", considering alternative pathways from gene to protein. An important component of this will be incorporating new strategies involving measurements additional to the full-size mass. This kind of additional information can be utilized to more fully characterize a protein or complex - each bit of data like this fills out one more piece of the characterization puzzle.

5.4.2 Task 2: Protein complex characterization

Organizational characterization of protein complexes can be done at different resolution levels. Through bioinformatic analysis of genomic and protein-protein interaction data, we can possibly derive a rough interaction map of a protein complex. This information can help the experimentalist identify which proteins are linked and which may be observed in the mass spectrometer. In conjunction with function predictions of proteins, such a capability can be used to suggest an initial target (complex) list for the experimentalist to work on. As more detailed information about interaction geometry is revealed through mass spectrometry and chemical cross-

linking, we will conduct more detailed characterization of the structural organization of the complexes, by taking advantage of the protein structure information in the PDB and that are being generated by the Structural Genomics Project. As we have demonstrated in our preliminary studies (see Section 5.3.2.3), detailed (not necessarily highly accurate) structural information can provide a great deal of functional information about a protein complex.

5.4.2.1 Identification of protein complexes

We will develop a computational capability for building protein-protein interaction maps in our target genomes (initially, *S. oneidensis*, and *R. palustris,* and later other DOE-relevant genomes), based on the experimental data and publicly available outside information (see **Specific Aim 1** of the **Molecular and Cellular Imaging subproject** for the detailed list of the information). Part of this task will be implemented in the **Molecular and Cellular Imaging subproject,** and the more R&D oriented work is proposed and budgeted in this subproject.

Information from various sources, including protein-protein interaction information from two-hybrid systems, conserved gene neighborhood information, and information from gene-expression data (see Task 5.4.2.3 for details), will be used to construct an initial map of interactions of protein complexes. We will represent all proteins and their interactions, identified based on this information, as a graph -- each protein represented as a node and each predicted pair-wise interaction as an edge connecting two proteins. Each edge has a distance, which will be defined as the inverse of the likelihood of the two proteins being interactive. A statistical model will be developed to define the likelihood, based on the overall prediction accuracy of each prediction method used. To achieve this, we will conduct thorough analyses on each of the methods listed in Task 5.4.2.3 to estimnate its prediction accuracy, based on a selected set of known protein-protein interactions and protein complexes. Prediction accuracies, in terms of sensitivity and specificity under different conditions, will be carefully documented. We will find the optimal weights for each of the prediction methods when combining them in the statistical model. Intuitively, a protein complex will form a connected subgraph with its edge density relatively high (compared to the edge density connecting this subgraph with the rest of the graph) and its edges having relatively short distances. We will formulate the problem of identifying a protein complex as an "interaction cluster" identification problem. We have recently developed a general clustering capability, based on graph-theoretic approaches [Xu et al., in press]. We will apply this general clustering technique to identify the "interaction clusters". These "interaction clusters" will be used as our initial prediction of protein complexes. These interaction maps will be further refined using the information from interpretation of mass spectral data of cross-linked proteins.

5.4.2.2 Characterization of organizational structure of a complex

We will characterize detailed structural organization based on structures of component proteins (monomers) using cross-linking information from simulated data or from experiments in Task 1. The structural information of component proteins could either exist in the PDB or be predicted using a computational method like PROSPECT. We will use a predicted structure only if the confidence score of PROSPECT's prediction is above certain threshold. Protein-protein docking will be our main technique to model the detailed complex structures. Existing docking algorithms are generally not adequate for large-scale applications due to the lack of prediction accuracy and high computational complexity [Aloy et al., 2001]. What makes our proposed approach unique is that we will use cross-linking information up-front as modeling constraints (*versus* treating constraints in the post-processing stage to filter the ones violating the constraints – the problem there is that the initial set may not even contain the correct docking structure). In addition, we have found that electrostatic surface calculation can provide useful information for restricting the portions of two protein structures that can interact. It is unlikely that two negatively or two positively charged surfaces will directly interact with each other. Thus simple charge surface matching will tell us which portions of two proteins could interact with each other, thus restricting search space to a small portion of the docking space. We will perform the electrostatic calculations with NWChem. This information will be used as a filter in each of the following processes for complex-structure construction.

We will consider four different cases when constructing a complex structure from individual structures: (a) coiled-coil interaction of soluble proteins, (b) rigid-body protein docking of soluble proteins under constraints,

(c) flexible docking of soluble protein under constraints, and (d) membrane-associated protein complexes. For a complex with more than two components, its structure will be constructed hierarchically through a series of two-body docking operations.

Coiled-coil structures: Coiled-coil proteins are characterized by a repeating sequence of seven amino acids, $(abcdefg)_n$, in which the a- and d-position residues are hydrophobic, while the e- and g-position residues are usually polar or charged [Wagschal et al., 1999]. This sequence pattern makes it possible to predict the occurrence of coiled coils. It was estimated that 5% of all proteins in *S. cerevisiae, C. elegans, A. thaliana and D. melanogaster* contain a coiled-coil region. Similar number should hold for our two target genomes. It is likely that most of these coiled coils mediate protein-protein interactions or oligomerization. The interaction between two coiled-coil proteins can be predicted relatively easily, since the intertwined packing between the two coiled-coils has a well-defined conformation [Wagschal et al., 1999]. With the help of cross-link information, the packing should be predicted accurately, using the algorithm of Wagschal et al., which we will implement.

Constrained rigid-body docking of two soluble proteins: The objective of this sub-task is to develop a computational method for rigid-body protein docking under the constraints of residue-residue links identified in Task 1, i.e., the docked structures should be consistent with the distances specified by the residue-residue links. Rigid-body docking covers a substantial subset of all docking problems.

We plan to build our data-constrained docking procedure based on some well-established and well-tested docking programs. We have carefully studied a number of existing docking methods and have decided to build our docking procedure based on an algorithm developed by Nussinov and co-workers [Norel, et al., 1994]. This algorithm has proven to be reliable and fast on a large test set. The original algorithm consists of the following six steps: (1) computation of molecular surface for each of two proteins based on their structures using Connolly's algorithm [Connolly, 1986]; (2) representation of protein surfaces by a set of sparse "interest points" and their normals; (3) matching compatible interest points and normals of the two proteins; (4) clustering closely related docking conformations; (5) evaluation of each docking conformation; (6) Ranking the survived docking conformations. Our modification will be done mainly in step 3 and step 6, to incorporate the cross-linking information. In step 3, instead of going through all possible pairings, we will examine only the sets of interest points from the two proteins, which are consistent with the residue-residue links. In step 6, we will also include the cross-linking information in the final ranking of the docked structures. To deal with possible inconsistency issues among identified residue-residue links (possibly due to experiment problems or problems in our data analysis codes developed in Task 2), e.g., two identified cross-links cannot be simultaneously satisfied, we will build a weighting scheme for identified cross-links, based on their associated structural regions and possibly other information. A voting procedure will be established, based on this weighting scheme, to overcome the possible inconsistency issue.

As the technology outlined above matures, we will refine docked structures through energy minimizations and molecular dynamics simulations. Explicit water molecules will be added during the refinement to account for solvent effects. Since the refinement process is computationally intensive, we will carry out the calculations on high performance parallel machines using NAMD [Nelson et al., 1995], which is applicable in parallel computing environment.

Flexible docking with distance constraints: Some protein structures may undergo significant changes during complex formation. To deal with such flexible docking problem, we will develop data-constrained docking procedures that will allow for structural changes upon complex formation. Considering the current reliability of flexible docking, we will initially focus on low-resolution (i.e., considering only C_α atoms) models. As the technology matures, we will gradually move into all-atom models in the out years of the project.

The development of the low-resolution approach will be especially important for two types of systems: large protein-protein complexes that induce significant structural rearrangements upon binding and protein-protein complexes of significant complexity (involving many subunits). The reasons for doing this are clear: (1) dramatic binding rearrangements cannot be captured by rigid docking schemes and (2) for large protein systems involving 20-40 proteins (e.g., some membrane complexes) the low resolution approach could be the only option. We will

use an approach similar to the one outlined below to modify the scoring fitness function derived from the characteristic distributions of residue specific backbone atom densities.

Our approach is to refine the prediction results obtained by the Rosetta-like technology [Bonneau et al., 2001] by introducing the constraints generated from the MS/CL data. Rosetta-like technology provides constraints for local conformation based on mini-threading. It produced extremely good structures with sparse constraints [Bowers et al., 2000]. To refine/improve results obtained from Rosetta, we will convert MS derived data into distance constraints and integrate them into the Rosetta method to rank/filter the results predicted by the Rosetta. Distance constraints will be derived from the information provided by Mass-Spec data as outlined in the Task 1. Two approaches to constraints integration will be tested. In the first approach, the simulations will be conducted with the existing Rosetta program and MS-constraints will be used as an output filter for the obtained structures before clustering and ranking of the filtered solutions. In the second approach the constraints will be integrated into an energy function used in Rosetta simulations. We will also utilize high performance computing: (1) to generate much bigger and more representative ensembles of structures, and (2) to keep complete information about each particular simulation in order to collect entropy-type information about the vicinity of the ground state conformation.

Structure modeling of membrane-associated complexes: Integral membrane protein structures are difficult to solve. We propose to model the 3D structures of membrane-associated complexes by adapting a technique called MS3D to specifically address integral membrane proteins. MS3D uses distance constraint information derived from intramolecular crosslinking, proteolysis, and mass spectrometry experiments to construct a three-dimensional model of a protein structure [Young et al., 2000]. We plan to use a similar approach to derive distance constraints within and between proteins in membrane-associated complexes. These distances, in conjunction with theoretical information such as helical packing preferences and genomic information, will be integrated to produce model structures that are consistent with experimental data. We are currently validating MS3D on a well-characterized integral membrane protein of known structure, bacteriorhodopsin (BR) in the light-activated rhodopsin-transducin complex.

Prediction of the structures of transmembrane (TM) domains, in particular, transmembrane helices, has proven to be more feasible than for soluble proteins [Bohm, 1996]. The two dimensional nature of the lipid bilayer places considerable constraints on the structure of TM protein segments, thus reducing the possible topological arrangements to a manageable number. Models for individual helical transmembrane proteins will be constructed by first defining the transmembrane domains by using neural networks [Rost et al., 1996], hydropathy profiles [Claros et al., 1994], hidden Markov models [Sonnhammer et al., 1998], sequence analyses [Persson & Argos, 1997], or using data from mutagenesis, labeling, and proteolysis experiments [Zelenski et al., 1999]. Once the helices are defined, the transmembrane domain can be constructed in one of two ways – either through selection of the correct domain structure from a pregenerated library, or by direct determination of the structure. The first approach involves a one-time-only computationally intensive calculation to generate the library of possible helix configurations (or templates) for each system (as in [Bowie, 1999]), followed by fast elimination of templates that are inconsistent with the experimental data. The other approach to transmembrane domain model building is to construct the experimentally consistent template(s) directly. By representing each transmembrane segment as a perfect alpha helix, the crosslinking data as through-space distance constraints, and adding in helix-helix connectivity constraints commensurate with the lengths of intermediate loops, we can use either distance geometry or simulated annealing to initially position the helices. Relaxation of the template can then be performed by energy minimization calculations using X-PLOR [Brunger, 1995], AMBER [Case et al., 2002], or SYBYL (http://www.tripos.com/). Once the structure of the transmembrane domain has been determined, we will model the structures of the extracellular and cytoplasmic loops using energy minimization and constrained molecular dynamics. The previously mentioned software packages, X-PLOR or AMBER, can be used for this step.

Models for protein complexes will be constructed hierarchically. The structures of the protein subunits, if not solved, will be modeled. Membrane protein models will be constructed using the methods described above; soluble protein models will be generated using constrained threading methods such as PROSPECT. The full model of the complex can be constructed from intermolecular crosslinking-based constraints by using distance

geometry to orient the subunits as rigid bodies relative to one another. The model can then be further refined using constrained energy minimization calculations.

5.4.2.3 Functional inference of protein complexes

Based on our experience with microbial genome annotation at ORNL, 30-50% of predicted genes can have some level of functional assignments using sequence-based methods like PSI-Blast [Altschul et al., 1997]. Using structure-based search methods like threading, we can probably assign an additional 20-30% of the genes to some structural folds in the PDB, which may provide some clues about their possible functions (see Section 5.3.2.3). Additional function information can also be derived through applications of the procedures outlined in the following:

Protein-protein interaction information: The classical "two-hybrid system" allows genome-wide characterization of protein networks by identifying all detectable binary interactions between proteins encoded by the genome. For the bacteria genomes that we are going to study, two-hybrid data may be limited. Nevertheless, the yeast interaction map may provide a useful reference template when constructing maps for our target genomes through homology search. In addition, databases like DIP [Xenarios et al., 2002] and BIND [Bader et al., 2001] also contain protein interaction data derived from additional experimental approaches, such as mass spectrometry [Ho et al., 2002].

Information of conserved genetic neighborhood: Proteins with conserved genetic neighborhood (a group of genes are arranged in tandem in one genome and also appear in a similar fashion in its related genomes) tend to interact with each other to form complexes. Operons represent one such conserved gene context. Identification of operons or "conserved gene contexts" can provide clues about which set of proteins may form a complex. Identification of operons is often done through a combination of discovering conserved gene context across multiple related genomes and the sequential arrangement of these genes in the genomic sequence. There are a number of existing algorithms for identification of operons [Lathe et al., 2000].

Information from gene expression data: In prokaryotic bacteria genomes, co-expressed genes may represent co-regulated genes. Hence identification of co-expressed genes could provide indication of interacting proteins and protein complexes. Co-expressed genes can possibly be identified through clustering of genes based on their expression profiles. We have developed a gene expression data analysis and clustering software for this purpose, called EXCAVATOR [Xu et al., 2001b, Xu et al., in press].

Information of co-occurance of genes in genomes: The phylogenetic profiling approach [Gaasterland and Ragan, 1998; Pellegrini et al., 1999] is based on the assumption that proteins functioning together in a pathway or structural complex are likely to evolve in a correlated fashion. During evolution, all such functionally linked proteins tend to be either preserved or eliminated in a new species. A phylogenetic profile can be used to represent the presence or absence of a protein in every known genome. It is shown that proteins having similar profiles tend to be functionally linked. The method of phylogenetic profiling will be used to establish the probability of two proteins interacting with each other and to predict the function of uncharacterized proteins.

Predicted protein interaction map will also be used to help make functional assignments. The idea is "guilt-by-association": If a hypothetical protein interacts with two or more proteins with the same function, it is likely that the hypothetical protein also has this function. If we have determined a hypothetical protein in a regulon and other genes in the regulon have the same known cellular role, then this hypothetical protein may have the same cellular role as well. In addition, we can study genes with similar expression profiles to the hypothetical protein. If there is a consensus among the functions of these genes, it may suggest the function of the hypothetical protein. We will implement an automated procedure to accomplish these. In the initial phase of the project, we will be content with this level of functional assignments. Methods towards understanding functions of a protein complex as a system will be investigated and applied in the out years of project.

5.4.2.4 Experimental validations and applications

For selected protein complexes (see the **Overview** of the Center proposal), we will apply our data analysis and computational modeling programs for the initial structural and functional characterization of these complexes. Each prediction (interaction map, structural organization or biological function) will be validated through an iterative process of computation and rationally designed experiments. The detailed validation protocol is described in the **Molecular and Cellular Imaging subproject**.

5.4.2.5 Towards understanding mechanisms of biological functions

In the out years of this subproject, we plan to investigate a number of computational techniques towards understanding the functional mechanisms of protein complexes. No budgetary requests are made in this proposal for these possible developments. We are listing them here only to provide a roadmap about the directions we may take two years from now. Additional budget may be requested a year or two from now if we feel we are ready to move to the next stage if this project is funded. These may include

Investigation of disordered regions of proteins and the binding mechanisms of protein complexes: Dr. Keith Dunker of Washington State University has published a series of pioneering work linking disordered regions [Romero et al., 1998; Romero et al., 2001; Dunker et al., 2001] with protein binding interface. When we have a sufficiently large number of identified protein complexes (in this project or elsewhere), we plan to conduct a systematic study on issues regarding how the protein-complex formation process can be possibly explained in terms of the concept of disordered regions, possibly shedding light on the binding mechanism of protein complexes. This work may be conducted through collaboration with Dr. Dunker, with whom we have had numerous discussions on this topic.

Studies of biochemical function for protein complex: Although the cellular function of a protein complex can be assigned using the method described in 5.4.2.3, understanding the mechanism of its biochemical function often represents a much more difficult task. We plan to use computational approaches to study biochemical functions of protein complexes and make suggestions for further experiments. The computational methods include (1) continuum electrostatics [Gilson et al., 1985; Juffer et al., 1991] using DELPHI [Honig & Nicholls, 1995] and GRASP [Nicholls et al., 1991], (2) molecular dynamics simulations to study the role of conformational changes in protein function [Ayton et al., 2002], and (3) simulated mutations to study the role of an individual residue in energetics and dynamics [Fanelli et al., 2001]. These computational studies will provide suggestions for mutagenesis (to study the role of an individual residue), neutron scattering (to study protein conformational changes), and calorimetry (to study energetics). An iterative approach using computational methods and experimental approaches can be very effective in elucidating the biochemical mechanism of a protein complex.

Long-time scale simulation of protein interaction dynamic: The simulation time-scale by existing codes is usually not long enough to allow biological functional analysis and interpretation. We will investigate data-constrained simulation techniques to significantly constrain our simulation trajectory and hence greatly increase the simulation time-scale, by using data collected from the novel cellular imaging techniques to be developed in the **Molecular and Cellular Imaging subproject** . 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. The constraints will be incorporated to accelerate the protein interaction process in simulation by using steered molecular dynamics [Lu & Schulten, 1999; Isralewitz et al., 2001].

Linking protein complexes to biological pathways: We will build a Bayesian hypothesis test that examines the proteins in a complex and determines whether those proteins are involved in either protein regulatory networks or gene-regulatory networks. The test will examine the statistical relationships between proteins using protein expression data or mRNA expression obtained, for instance, from the *Shewanella* Microbial Cell Projects. In a Bayesian network model, the relationship between variables (genes) is determined by the joint probability

distributions of the genes that are consistent with the independence assertions in the graph. Two genes **C** and **D** are independent if the probability of the expression of both genes can be factored into the probability of expression of each gene individually, P(**C,D**)= P(**C**)P(**D**). If this is true then there is no path connecting the genes in the graphical model. For a directed graph with two nodes **B** and **C** and an edge **B**->**C**, the parent of **C** is **B**. In a Bayesian network, the joint probability distribution of the genes in the network can be factored into a product of probability distributions in which each gene is conditioned on its parents.

5.4.3 Task 3: Protein complex database and collaborative computational environment

The process of identifying and characterizing multi-protein complexes produces diverse data sets distributed across a number of contributing institutions. To facilitate the development of new bioinformatic methods and to support computational analysis, this data must be easily accessible, managed and mined as if it resided in one repository. Data ingest, analysis and dissemination must be managed as a single task across all elements of the project that will be creating or using the data sets. We are developing a data management strategy and the associated tools to handle these diverse distributed data sets and creating a computing environment that supports the biologist through the scientific discovery process. Two parallel problem solving environments will be developed at ORNL and PNNL, considering the related on-going R&D works at both organizations and their current expertise.

5.4.3.1 Protein complex data

The Center's Protein Complex Database will contain all the experimental and computational data, and prediction/modeling results generated in this project. This includes (a) identified proteins, (b) identified PTMs and mutations of each protein, (c) identified protein-protein interactions, (d) interaction maps of protein complexes, (e) complex structures, in low- and high-resolutions, and biological functions, at different levels, and (f) links to related protein databases. The initial data model for this database will be based on 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. The database development work will be split between this project and the **Molecular and Cellular Imaging subproject**. The R&D work on data model construction and search capabilities will be done under this project and the implementation work of the databases will be under the Core.

A number of search and analysis capabilities will be built on top of this Protein Complex Database to facilitate information extraction by outside users. These 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.

5.4.3.2 Computational cell environment (CCE) at PNNL

Computational Cell Environment (CCE) is a problem-solving environment (PSE) that is currently under development at PNNL. The CCE will manage data such as the sample preparation information, proteomic experimentation, and annotation going to and from ORNL and PNNL databases. The CCE will provide a unified, integrated interface to data residing in multiple databases, in essence creating a single, virtual data repository. This has many underlying requirements: (1) the data must be interoperable and the pedigree of the data be tracked; (2) the analytical tools used must be able to support the fusion of multiple data streams into a coherent visual environment and provide the ability to hypothesize; (3) the interfaces with which a scientist interacts must be useable by the wide-range of users – experimentalists, computational biologists, and theorists; and 4) the tools must be able to provide easy ways for researchers to share data and information. The broad goal of the CCE is to provide these capabilities as a web-based biological PSE. The portion of the CCE that will be addressed in this proposal includes the provision of an environment for data storage and retrieval for the data that is collected and

analyzed and the web interface needed to support data retrieval. We do not imply that we will be recreating any existing databases; instead, this work will provide an integrated environment for data mining and use.

This flexible access to scientific data can be achieved by implementing a virtual data repository through the Distributed Authoring and Versioning (DAV) protocol [Whitehead & Wiggins, 1998] This data repository is a representation of the data as it exists in multiple locations, but does not have to be a physical data repository placed in one location. The data repository is designed to contain meaningful information of a cross-disciplinary nature that can be stored, queried, and readily manipulated for use in data mining. The DAV-enabled Web server is a multi-tiered, component based data object model that supports arbitrary metadata using the DAV protocol.

The basis for the CCE architecture is the formatting of metadata using the Extensible Markup Language (XML). XML is a powerful language for encoding the definition of technical terms in a human- and machine-readable form. XML's expressive power, together with the availability of technologies for manipulating it – authoring, parsing, validating, translating, etc., have made it a *de facto* standard for information exchange in new systems. DAV is an Internet Engineering Task Force (IETF) standard set of extensions to the HTTP/1.1 protocol to support basic data management over the web with a unifying approach to data representation and access. This includes storage and retrieval of typed, opaque data files/objects, content locking, hierarchical collections and annotation of the data with arbitrary metadata to make collections and objects semantically meaningful. DAV provides operations for managing the metadata accompanied by a sister protocol DAV Searching and Locating (DASL) that adds methods for server-side search capabilities. DASL provides a basic search grammar and can be extended with additional grammars, e.g. XML Query. The CCE schema will be based on an XML representation of the data. Several XML schemas have been defined for biological information primarily focused on gene annotation. We will base the CCE schema on one or several of these schemas, extending the CCE schema as needed in order to represent the full community of biological information from gene annotation to proteomic experimentation.

5.4.3.3 Mass spectrometry tools (MST) environment at ORNL

A large number of analysis tools will be required for the computational analysis of the Mass Spectrometry data that will be generated in the Center's laboratories at ORNL and PNNL. These tools will be deployed on the ORNL and PNNL high performance, massively parallel supercomputers, as well as on Unix workstation clusters at both laboratories. MST Environment will provide a unified, integrated interface to this distributed deployment of tools, while internally managing the distribution of analysis requests on the available computational resources at both laboratories. Communication protocols will be established between ORNL MST and PNNL CCE for analysis transactions, to enable access to specific tools deployed within either facility. This will provide flexibility of independent tools and system development at the two laboratories, at the same time facilitating collaboration in tools and computational resource sharing. Individual tools may need to access data repositories, for which they will be able to use PNNL CCE database access interface. Similarly, CCE data repository components may need to use analysis tools, for which they will be able to use ORNL MST interface.

The MST Environment will consist of four main components: **ServiceRegistry** server will provide information about all available tool services, and detailed interface specifications for each service. These specifications can be used by a client to formulate and submit valid analysis service requests. **RequestServer** will accept service requests from clients, authenticate (when appropriate) and validate them and issue request ID tickets. **ResultServer** will provide status information for individual request ID tickets and will return the analysis results on completion of individual analysis tasks. These three components will provide the external interface to the system. Access to all three servers will be via TCP socket connection or Web CGI request, using predefined XML (Extensible Markup Language) message specifications. The use of XML as a data exchange mechanism provides many benefits including data format standardization, robust data parsing and validation, data translation and merging and portability across diverse computer architectures. The fourth component, **TaskManager**, will internally coordinate and manage task queuing and distribution on available resources, perform system status monitoring and fault detection and queue migration, and time estimation for individual requests.

The development of such a system will involve implementation of a service layer abstraction. The service layer will make extensive use of XML for creating a precise and comprehensive service specification for each tool,

including tool version, input and output data formats and options, required and optional parameters, and default parameter values. Every attempt will be made to use existing XML representations developed within the biological community. Relevant parts of XML specifications for related tools will be standardized to provide a consistent overall interface. Each tool's service layer will implement data format translators to convert between standardized service formats and data formats that the tool itself may require. This will be especially useful when incorporating third party tools for which source code may be unavailable or modification of the code may be unwieldy. In some cases, a third party tool itself may not be available for local installation, and access via Web CGI request may be required. Such tools will also be incorporated in MST Environment by capturing the tool specifications from the tool's form URL, and implementing Web Agents for request processing.

We will explore the use of new and emerging technologies for deployment of MST Environment capabilities as Web services. We will consider using technologies such as Web Services Description Language (WSDL) for service specification, Universal Description Discovery and Integration (UDDI) for service registry implementation and Simple Object Access Protocol (SOAP) for communication between system components and with other collaborating systems.

5.6 Literature Cited

Abagyan, R., and Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol *235*, 983-1002 (1994).

Aloy, P., Querol, E., Aviles, F. X., and Sternberg, M. J. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol *311*, 395-408 (2001).

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-402 (1997).

Ayton, G., Smondyrev, A. M., Bardenhagen, S. G., McMurtry, P., and Voth, G. A. Calculating the bulk modulus for a lipid bilayer with nonequilibrium molecular dynamics simulation. Biophys J *82*, 1226-38 (2002).

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., and Hogue, C. W. BIND--The Biomolecular Interaction Network Database. Nucleic Acids Res *29*, 242-5 (2001).

Bairoch, A., and Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res *28*, 45-8 (2000).

Barnes, D., Lai, W., Breslav, M., Naider, F., and Becker, J. M. PTR3, a novel gene mediating amino acid-inducible regulation of peptide transport in Saccharomyces cerevisiae. Mol Microbiol *29*, 297-310 (1998).

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D., and Sonnhammer, E. L. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res *27*, 260-2 (1999).

Beisel, H. G., Kawabata, S., Iwanaga, S., Huber, R., and Bode, W. Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab Tachypleus tridentatus. Embo J *18*, 2313-22 (1999).

Beliaev, A. S., Saffarini, D. A., McLaughlin, J. L., and Hunnicutt, D. MtrC, an outer membrane decahaem c cytochrome required for metal reduction in Shewanella putrefaciens MR-1. Mol Microbiol *39*, 722-30 (2001).

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol *112*, 535-42 (1977).

Bleasby, A. J., Akrigg, D., and Attwood, T. K. OWL--a non-redundant composite protein sequence database. Nucleic Acids Res *22*, 3574-7 (1994).

Blom, N., Gammeltoft, S., and Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol *294*, 1351-62 (1999).

Blundell, T. L., Burke, D. F., Chirgadze, D., Dhanaraj, V., Hyvonen, M., Innis, C. A., Parisini, E., Pellegrini, L., Sayed, M., and Sibanda, B. L. Protein-protein interactions in receptor activation and intracellular signalling. Biol Chem *381*, 955-9 (2000).

Bohm, G. New approaches in molecular structure prediction. Biophys Chem *59*, 1-32 (1996).

Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E., and Baker, D. Rosetta in CASP4: Progress in ab initio protein structure prediction. Proteins *45*, 119-26 (2001).

Bowers, P. M., Strauss, C. E., Baker, D. De novo protein structure determination using sparse NMR data. J Biomol NMR. *18*, 311-8 (2000).

Bowie, J. U. Helix-bundle membrane protein fold templates. Protein Sci *8*, 2711-9 (1999).

Brunger, A. T. Solution of a Fab (26-10)/digoxin complex by generalized molecular replacement. Acta Crystallogr A *47*, 195-204 (1991).

Brunger, A. T. X-PLOR, version 3.1. A system for X-ray crystallography and NMR, New Haven: Yale University Press, 1992.

Burkhard, P., Stetefeld, J., and Strelkov, S. V. Coiled coils: a highly versatile protein folding motif. Trends Cell Biol *11*, 82-8 (2001).

Butt, A., Davison, M. D., Smith, G. J., Young, J. A., Gaskell, S. J., Oliver, S. G., and Beynon, R. J. Chromatographic separations as a prelude to two-dimensional electrophoresis in proteomics analysis. Proteomics *1*, 42-53 (2001).

Cai, K., Itoh, Y., and Khorana, H. G. Mapping of contact sites in complex formation between transducin and light-activated rhodopsin by covalent crosslinking: use of a photoactivatable reagent. Proc Natl Acad Sci U S A *98*, 4877-82 (2001).

Case, D., Merz, K., Cheatham, T., Simmerling, C., Darden, T., and Pearlman, D. AMBER: http://amber.ch.ic.ac.uk/doc7/index.html. (2002).

Chen, T., Jaffe, J. D., and Church, G. M. Algorithms for identifying protein cross-links via tandem mass spectrometry. J Comput Biol *8*, 571-83 (2001).

Claros, M. G., and von Heijne, G. TopPred II: an improved software for membrane protein structure predictions. Comput Appl Biosci *10*, 685-6 (1994).

Cornell, W. D. A second generation force field for the simulation of proteins and nucleic acids. Journal of the American Chemical Society *117*, 5179-5197 (1995).

Costanzo, M. C., Crawford, M. E., Hirschman, J. E., Kranz, J. E., Olsen, P., Robertson, L. S., Skrzypek, M. S., Braun, B. R., Hopkins, K. L., Kondu, P., Lengieza, C., Lew-Smith, J. E., Tillberg, M., and Garrels, J. I. YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. Nucleic Acids Res *29*, 75-9 (2001).

Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., and Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. J Comput Biol *6*, 327-42 (1999).

de Wildt, R. M., Mundy, C. R., Gorick, B. D., and Tomlinson, I. M. Antibody arrays for high-throughput screening of antibody-antigen interactions. Nat Biotechnol *18*, 989-94 (2000).

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. Intrinsically disordered protein. J Mol Graph Model *19*, 26-59 (2001).

Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., and Cherry, J. M. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res *30*, 69-72 (2002).

Eng, J. K., McCormack, A. L., Yates, J. R. I. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom *5*, 976-989 (1994).

Fanelli, D., Merafina, M., and Ruffo, S. One-dimensional toy model of globular clusters. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics *63*, 066614 (2001).

Fernandez-de-Cossio, J., Gonzalez, J., Satomi, Y., Shima, T., Okumura, N., Besada, V., Betancourt, L., Padron, G., Shimonishi, Y., and Takao, T. Automated interpretation of low-energy collision-induced dissociation spectra by SeqMS, a software aid for de novo sequencing by tandem mass spectrometry. Electrophoresis *21*, 1694-9 (2000).

Fields, S., and Song, O. A novel genetic system to detect protein-protein interactions. Nature *340*, 245-6 (1989).

Finley, R. L., Jr., and Brent, R. Interaction mating reveals binary and ternary connections between Drosophila cell cycle regulators. Proc Natl Acad Sci U S A *91*, 12980-4 (1994).

Fischer, D., Lin, S. L., Wolfson, H. L., and Nussinov, R. A geometry-based suite of molecular docking processes. J Mol Biol *248*, 459-77 (1995).

Forsberg, H., Gilstring, C. F., Zargari, A., Martinez, P., and Ljungdahl, P. O. The role of the yeast plasma membrane SPS nutrient sensor in the metabolic response to extracellular amino acids. Mol Microbiol *42*, 215-28 (2001).

Gaasterland, T., and Ragan, M. Phyletic and Functional Patterns of ORF Distribution Among Prokaryotes. Journal of Microbial and Comparative Genomics *3*, 199-217 (1998).

Gilson, M. K., Rashin, A., Fine, R., and Honig, B. On the calculation of electrostatic interactions in proteins. J Mol Biol *184*, 503-16 (1985).

Graves, D. J., Martin, B., and Wang, J. H. Co- and post-translational modification of proteins: Chemical principles and biological effect ,N.Y., N.Y: Oxford Press, (1994).

Grundling, A., Blasi, U., and Young, R. Biochemical and genetic evidence for three transmembrane domains in the class I holin, lambda S. J Biol Chem *275*, 769-76 (2000).

Gschwend, D. A., Good, A. C., and Kuntz, I. D. Molecular docking towards drug discovery. J Mol Recognit *9*, 175-86 (1996).

Herzyk, P., and Hubbard, R. E. An atomic model of transmembrane domain of bovine rhodopsin from experimentally derived structural restraints. Biophys J *72*, TUAM4-TUAM4 (1997).

Herzyk, P., and Hubbard, R. E. Automated method for modeling seven-helix transmembrane receptors from experimental data. Biophys J *69*, 2419-42 (1995).

Herzyk, P., and Hubbard, R. E. Using experimental information to produce a model of the transmembrane domain of the ion channel phospholamban. Biophys J *74*, 1203-14 (1998).

Hines, W., Falick, A., Burlingame, A., and Gibson, B. Pattern-Based Algorithm for Peptide Sequencingfrom Tandem High Energy Collision-Induced Dissociation Mass Spectra. J Am Soc Mass Spectrom *3*, 326-336 (1992).

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180-3 (2002).

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. The PROSITE database, its status in 1999. Nucleic Acids Res *27*, 215-9 (1999).

Honig, B., and Nicholls, A. Classical electrostatics in biology and chemistry. Science *268*, 1144-9 (1995).

Huynen, M., Snel, B., Lathe, W., 3rd, and Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res *10*, 1204-10 (2000).

Iraqui, I., Vissers, S., Bernard, F., de Craene, J. O., Boles, E., Urrestarazu, A., and Andre, B. Amino acid signaling in Saccharomyces cerevisiae: a permease-like sensor of external amino acids and F-Box protein Grr1p are required for transcriptional induction of the AGP1 gene, which encodes a broad-specificity amino acid permease. Mol Cell Biol *19*, 989-1001 (1999).

Island, M. D., Perry, J. R., Naider, F., and Becker, J. M. Isolation and characterization of S. cerevisiae mutants deficient in amino acid-inducible peptide transport. Curr Genet *20*, 457-63 (1991).

Isralewitz, B., Gao, M., and Schulten, K. Steered molecular dynamics and mechanical functions of proteins. Curr Opin Struct Biol *11*, 224-30 (2001).

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A *98*, 4569-74 (2001).

Itoh, Y., Cai, K., and Khorana, H. G. Mapping of contact sites in complex formation between light-activated rhodopsin and transducin by covalent crosslinking: use of a chemically preactivated reagent. Proc Natl Acad Sci U S A *98*, 4883-7 (2001).

Jansen, R., and Gerstein, M. Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. Nucleic Acids Res *28*, 1481-8 (2000).

Jansen, R., Greenbaum, D., and Gerstein, M. Relating whole-genome expression data with protein-protein interactions. Genome Res *12*, 37-46 (2002).

Jones, D. T. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol *287*, 797-815 (1999).

Jones, P. C., Sivaprasadarao, A., Wray, D., and Findlay, J. B. A method for determining transmembrane protein structure. Mol Membr Biol *13*, 53-60 (1996).

Juffer, A. H., Botta, E. F. F., Keulen, B. A. M., Ploeg, A., and Berendsen, H. J. C. The Electric Potential of a Macromolecule in a Solvent: A Fundamental Approach. Journal of Computational Physics *97*, 144-171 (1991).

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. The KEGG databases at GenomeNet. Nucleic Acids Res *30*, 42-6 (2002).

Karplus, K., Barrett, C., and Hughey, R. Hidden Markov models for detecting remote protein homologies. Bioinformatics *14*, 846-56 (1998).

Klasson, H., Fink, G. R., and Ljungdahl, P. O. Ssy1p and Ptr3p are plasma membrane components of a yeast system that senses extracellular amino acids. Mol Cell Biol *19*, 5405-16 (1999).

Kontoyianni, M., DeWeese, C., Penzotti, J. E., and Lybrand, T. P. Three-dimensional models for agonist and antagonist complexes with beta 2 adrenergic receptor. J Med Chem *39*, 4406-20 (1996).

Krishna, R. G., and Wold, F. Posttranslational Modifications. In Proteins: Analysis & Design, R. Angeletti, ed. (San Diego, CA: Academic Press), pp. 121-206 (1998).

Kuhlbrandt, W., and Williams, K. A. Analysis of macromolecular structure and dynamics by electron cryo-microscopy. Curr Opin Chem Biol *3*, 537-43 (1999).

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. J Mol Biol *161*, 269-88 (1982).

Lathe, W. C., 3rd, Snel, B., and Bork, P. Gene context conservation of a higher order than operons. Trends Biochem Sci *25*, 474-9 (2000).

Lennon, J. J., and Walsh, K. A. Locating and identifying posttranslational modifications by in-source decay during MALDI-TOF mass spectrometry. Protein Sci *8*, 2487-93 (1999).

Lu, H., and Schulten, K. Steered molecular dynamics simulations of force-induced protein domain unfolding. Proteins *35*, 453-63 (1999).

MacBeath, G., and Schreiber, S. L. Printing proteins as microarrays for high-throughput function determination. Science *289*, 1760-3 (2000).

Mann, M., and Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem *66*, 4390-9 (1994).

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. Science *285*, 751-3 (1999).

Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S., and Weil, B. MIPS: a database for genomes and protein sequences. Nucleic Acids Res *28*, 37-40 (2000).

Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C., and Szyperski, T. Protein NMR spectroscopy in structural genomics. Nat Struct Biol *7 Suppl*, 982-5 (2000).

Mrowka, R. A Java applet for visualizing protein-protein interaction. Bioinformatics *17*, 669-71 (2001).

National Institute of General Medical Sciences Pilot Projects for the Protein Structure Initiative (Structural Genomics). http://www.nih.gov/grants/guide/rfa-files/RFA-GM-99-009.html *June*, RFA GM-99-009 (1999).

Nelson, M., Humphrey, W., Gursoy, A., Dalke, A., Kale, L., Skeel, R., and Schulten, K. NAMD- A parallel, object-oriented molecular dynamics program. J. Supercomputing App. *10*, 251-268 (1995).

Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P., Sleeman, J., Lamond, A., and Mann, M. Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. Nat Genet *20*, 46-50 (1998).

Nicholls, A., Sharp, K. A., and Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins *11*, 281-96 (1991).

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst *8*, 581-99 (1997).

Norel, R., Fischer, D., Wolfson, H. J., and Nussinov, R. Molecular surface recognition by a computer vision-based technique. Protein Eng *7*, 39-46 (1994).

Pandey, A., and Mann, M. Proteomics to study genes and genomes. Nature *405*, 837-46 (2000).

Pearson, K. M., Pannell, L. K., and Fales, H. M. Intramolecular cross-linking experiments on cytochrome c and ribonuclease A using an isotope multiplet method. Rapid Commun Mass Spectrom *16*, 149-59 (2002).

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A *96*, 4285-8 (1999).

Peng, J., and Gygi, S. P. Proteomics: the move to mixtures. J Mass Spectrom *36*, 1083-91 (2001).

Persson, B., and Argos, P. Prediction of membrane protein topology utilizing multiple sequence alignments. J Protein Chem *16*, 453-7 (1997).

Pevzner, P. A., Mulyukov, Z., Dancik, V., and Tang, C. L. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. Genome Res *11*, 290-9 (2001).

Rappsilber, J., Siniossoglou, S., Hurt, E. C., and Mann, M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. Anal Chem *72*, 267-75 (2000).

Reineke, U., Volkmer-Engert, R., and Schneider-Mergener, J. Applications of peptide arrays prepared by the SPOT-technology. Curr Opin Biotechnol *12*, 59-64 (2001).

Rice, L. M., and Brunger, A. T. Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. Proteins *19*, 277-90 (1994).

Rodi, D. J., and Makowski, L. Phage-display technology--finding a needle in a vast molecular haystack. Curr Opin Biotechnol *10*, 87-93 (1999).

Roepstorff, P., and Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom *11*, 601 (1984).

Rohdich, F., Hecht, S., Gartner, K., Adam, P., Krieger, C., Amslinger, S., Arigoni, D., Bacher, A., and Eisenreich, W. (2002). Studies on the nonmevalonate terpene biosynthetic pathway: metabolic role of IspH (LytB) protein. Proc Natl Acad Sci U S A *99*, 1158-63 (2002).

Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., Garner, E., Guilliot, S., and Dunker, A. K. Thousands of proteins likely to have long disordered regions. Pac Symp Biocomput, 437-48 (1998).

Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. Sequence complexity of disordered protein. Proteins *42*, 38-48 (2001).

Rost, B., Fariselli, P., and Casadio, R. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci *5*, 1704-18 (1996).

Sali, A. Functional links between proteins. Nature *402*, 23, 25-6 (1999).

Sali, A., and Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol *234*, 779-815 (1993).

Schwikowski, B., Uetz, P., and Fields, S. A network of protein-protein interactions in yeast. Nat Biotechnol *18*, 1257-61 (2000).

Smith, T. F., and Waterman, M. S. Identification of common molecular subsequences. J Mol Biol *147*, 195-7 (1981).

Snel, B., Bork, P., and Huynen, M. Genome evolution. Gene fusion versus gene fission. Trends Genet *16*, 9-11 (2000).

Sofia, H. J., Chen, G., Hetzler, B. G., Reyes-Spindola, J. F., and Miller, N. E. Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods. Nucleic Acids Res *29*, 1097-106 (2001).

Sonnhammer, E. L., von Heijne, G., and Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol *6*, 175-82 (1998).

Steiner, K., and Malke, H. relA-Independent amino acid starvation response network of Streptococcus pyogenes. J Bacteriol *183*, 7354-64 (2001).

Steinhoff, H. J., Mollaaghababa, R., Altenbach, C., Hideg, K., Krebs, M., Khorana, H. G., and Hubbell, W. L. Time-resolved detection of structural changes during the photocycle of spin-labeled bacteriorhodopsin. Science *266*, 105-7 (1994).

Stephanopoulos, G. Metabolic engineering. Biotechnol Bioeng *58*, 119-20 (1998).

Taylor, J. A., and Johnson, R. S. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom *11*, 1067-75 (1997).

Tjandra, N., and Bax, A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science *278*, 1111-4 (1997).

Tucker, C. L., Gera, J. F., and Uetz, P. Towards an understanding of complex protein networks. Trends Cell Biol *11*, 102-6 (2001).

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623-7 (2000).

Vakser, I. A., and Aflalo, C. Hydrophobic docking: a proposed enhancement to molecular recognition techniques. Proteins *20*, 320-9 (1994).

Vaughan, R. A., and Kuhar, M. J. Dopamine transporter ligand binding domains. Structural and functional properties revealed by limited proteolysis. J Biol Chem *271*, 21672-80 (1996).

Venkateswaran, K., Moser, D. P., Dollhopf, M. E., Lies, D. P., Saffarini, D. A., MacGregor, B. J., Ringelberg, D. B., White, D. C., Nishijima, M., Sano, H., Burghardt, J., Stackebrandt, E., and Nealson, K. H. Polyphasic taxonomy of the genus Shewanella and description of Shewanella oneidensis sp. nov. Int J Syst Bacteriol *49 Pt 2*, 705-24 (1999).

VerBerkmoes, N. C., Bundy, J.L., Hauser, L., Asano, K.G., Razumovskaya, J., Larimer, F., Hettich, R.L., Stephenson, J.L. Integrating "top-down" and "bottom-up" mass spectrometric approaches for proteomic analysis of *Shewanella oneidensis*. J. Proteome. Res. *Submitted (2002)*.

5.8 Milestones

As outlined in the Overview to the Center proposal, the tasks described in the subprojects are the initial research projects being proposed to support the Center for Molecular and Cellular Systems. These tasks will be reviewed annually as outlined in the Management Overview for progress and continued relevance to the Center. This subproject will receive relatively low budget in FY02 and FY03 due to limited data available from the Center. Substantial budget increase is planned in FY04 and thereafter, since bioinformatics will play a more critical role in the Center as massive data are generated. It should be noted that techniques developed in this subproject will be incorporated into the both Center sites at ORNL and PNNL for the high throughput analyses of protein complexes.

Deliverables in FY03 include new algorithms for predicting protein complex interaction maps, new algorithms for interpreting mass spectrometry data of cross-linked proteins, a prototype system for a problem-solving environment, and new algorithms for data-constrained rigid docking. Deliverables in FY04 include new algorithms for identifying PTMs from mass spectrometry data, improved algorithms and software for interpreting mass spectrometry data of cross-linked proteins, a fully developed problem-solving environment (including implementation of all the analysis tools on supercomputers), new algorithms for constructing protein complex structures, and new algorithms for inference of functions of protein complexes. In FY05 and thereafter, we will develop more sophisticated analysis, prediction and modeling methodologies for protein complex characterization, further test our algorithms and the problem-solving environment using experimental data generated from the Center, and apply our development in large-scale studies of protein complexes. The following table summarizes the tasks and schedule.

Tasks	FY03	FY04	FY05
5.4.4.1 Data analysis of cross-linked proteins	x	x	x
5.4.1.2 Identification of PTMs/mutations	x	x	x
5.4.2.1 Identification of protein complexes		x	x
5.4.2.2 Characterization of organizational structure of complex	x	x	x
5.4.2.2a Coiled-coil structures		x	x
5.4.2.2b Constrained rigid-body docking of two soluble proteins	x	x	
5.4.2.2c Flexible docking with distance constraints		x	x
5.4.2.2d Structure modeling of membrane-associated complexes	x	x	x
5.4.2.3 Functional inference of protein complexes		x	x
5.4.2.4 Experimental validations and applications			x
5.4.2.5 Understanding functional mechanism			x
5.4.3 Collaborative computational environment	x	x	x
5.4.3.1 Protein complex database	x	x	
5.4.3.2 Computational cell environment (CCE) at PNNL	x	x	
5.4.3.3 Mass spectrometry tools environment at ORNL	x	x	