Genomes to Life: Technology Assessment for Mass Spectrometry

December 10-11, 2001, Washington, D.C.

Executive Summary

Over the past decade, mass spectrometry (MS) has become the most widely used analytical tool for characterizing proteins and other biomolecules. MS undoubtedly will play an important role in the Department of Energy's Genomes to Life (GTL) program developed by the Office of Science offices of Biological and Environmental Research (BER) and Advanced Scientific Computing Research (ASCR). The GTL program has four goals, all focused on understanding the composition and function of the biochemical networks and pathways that carry out the essential processes of living organisms. Although MS will be utilized in Goals 1, 2, and 3, it is particularly relevant to Goal 1, which is to identify life's molecular machines, the multiprotein complexes that carry out the function of living systems. In addition, MS will require and, in turn, drive the development of the computational tools described in Goal 4.

Because MS is a key enabling tool for meeting GTL long-term goals, a workshop was held to assess the status of MS for robust, high-throughput analysis of proteins and protein complexes and to identify where technology improvements are needed. Scientists from across the country, including experts from universities and national and industrial laboratories, were invited to participate. In addition, current capabilities and needs were assessed for preparing samples from microbial systems before MS analysis. Finally, computational tools for data interpretation, handling, archiving, and modeling were evaluated. Invited speakers and workshop participants were encouraged to discuss current capabilities, "warts and all," so areas for improvements could be identified to meet GTL long-term goals.

MS-based techniques for characterizing and even quantitatively measuring proteins were generally thought to be relatively mature. The most common of these use a "bottom-up" approach in which proteins are separated by 2D gel electrophoresis and subsequently digested to form peptides that are analyzed by MS. Recently, labor-intensive, time-consuming gel electrophoresis has been replaced by liquid chromatography, greatly improving total analysis time. A mass spectrum ("peptide map") from the enzymatic digestion of a protein can be compared to a database to identify the protein, or tandem mass spectrum, which provides partial amino acid sequence of a particular peptide. Both approaches use data obtained at the peptide level to identify the proteins. Another variation for peptide-level analysis employs very high mass resolution Fourier transform ion cyclotron resonance (FTICR) MS to identify components from peptide digests based on the exact masses of peptide fragments called "accurate mass tags" (AMTs). Although this technique requires appreciable time to establish a suite of AMTs for each organism, it potentially can be used to rapidly detect proteins from specific organisms without requiring tandem MS spectra to identify the peptides. Using either method, labeling with stable isotopes can provide a means of quantitatively measuring the number of proteins present.

Recently, these bottom-up approaches have been augmented by a top-down method in which accurate molecular weight measurement of the intact protein provides information on existing modifications. Combining these two methods is valuable for obtaining a more complete picture of the proteome and for characterizing protein complexes. Identification and quantitative measurement of these modifications are critical for understanding the function of protein complexes and central to meeting GTL goals.

Techniques for MS analysis of protein complexes were judged to be much less developed than those for individual proteins. The best strategies and methods are not yet known for isolating protein complexes in a robust manner that preserves all-important interactions and does not isolate biologically irrelevant artifactual, contaminating polypeptides. This is the key challenge in analyzing protein complexes. Although MS has been used to analyze some protein complexes directly, this approach was thought to have only limited application to the full range of protein complexes in a cell. One successfully employed approach uses engineered tags to modify target proteins to allow "pulling down" of complexes from mixtures. Some promising techniques use cross-linking reagents to stabilize delicate complexes, and others use a combination of cross-linking and affinity tags to help identify complexes while minimizing sample preparation.

Workshop participants noted that the major bottlenecks in analyzing proteins and protein complexes are not within the MS analysis itself but rather in sample preparation before analysis and in data handling and interpretation. New and automated procedures will be required for efficiently preparing samples from intact cells and introducing them into the mass spectrometer. In the case of protein complexes, sample-handling steps are critical to minimize both the dissociation of weakly bound complexes and the formation of artifacts. Gavin et al. (*Nature* **415**, 141–47, 2002) reported detection of complexes at 15 copies per cell, but this number varies quite a bit among different labs. Attendees strongly agreed that computational tools are needed for virtually every aspect of Goal 1 to reduce demands on analytical technology and to increase throughput and information content. Examples include tools for MS data interpretation that build on information provided by the gene sequence, creation of databases and tools for data mining, and a wide range of modeling tools. Finally, attendees noted that the most efficient way to approach the high-throughput analysis of proteins and protein complexes would be to establish centralized facilities with many instruments and integrated computational resources dedicated to this task.

Introduction

The DOE Office of Science, including BER and ASCR, has developed a roadmap for a scientific program that builds on the foundation of whole-genome sequences to achieve a fundamental, comprehensive, and systematic understanding of life (see http://DOEGenomesToLife.org/). The GTL program, which will use a "systems biology"approach, will require highly integrated, multidisciplinary teams of scientists to address the issues outlined in the roadmap. GTL's four goals are to (1) identify life's molecular machines, the multiprotein complexes that carry out the functions of living systems; (2) characterize the gene regulatory networks that control these molecular machines; (3) characterize the functional repertoire of complex microbial communities at the molecular level and in their natural environments; and (4) develop computers and computational capabilities needed to model the complexity of biological systems.

Early in the GTL planning process, participants recognized that a systems approach requires reliable, high-throughput technologies. The evolution of current and development of new technologies, along with their marriage to automation and advanced computational tools, will be critical to generating the data required to meet GTL goals. Mass spectrometry was identified by DOE staff as key to GTL, especially in identifying proteins and protein complexes as part of Goal 1 and in support of Goals 2 and 3. As a result, DOE sponsored a workshop on mass spectrometry on December 10–11, 2001, in Washington, D.C. More than 40 scientists from DOE, academic, and industrial laboratories, as well as DOE representatives, gathered to appraise MS capabilities for analyzing proteins and protein complexes and to examine GTL technology needs. Speakers were requested to realistically assess the current state of the art and to identify needed improvements in MS instrumentation, sample preparation, and computation and bioinformatics. Principal emphasis was placed on issues related to Goal 1. The workshop agenda and list of attendees are in Appendix A.

Overview of Goal 1

Marvin Frazier (DOE OBER) provided introductory comments about GTL and outlined meeting objectives. Ray Gesteland (University of Utah) helped place Goal 1 into the context of GTL's larger vision. He stated that directly analyzing proteins and protein complexes is key to GTL. mRNA-expression chips do not provide information on global protein modifications, so analysis of intact proteins is critical to understanding post-translational modifications, alternate gene splicings, protein abundance, and the impact of these and other alterations on biological function. More than 200 modifications are possible for most proteins; furthermore, analyzing peptides from protein digests alone will not capture this information, so analyzing intact proteins is required. As part of GTL, we will identify new proteins, protein complexes, protein-degradation products, and even new amino acids. We do not know how complicated protein complexes are and how many exist at any given time in the cell—probably thousands. Being able to identify correct protein complexes will be critical to GTL, so we have to make sure that the complexes studied represent what is actually in a cell. We need to understand how changing one or more genes will impact these molecular machines, so we must be able to follow the changes in our analytical schema. Further, we need to know the location of complexes in a cell and understand the dynamics of their formation. Information from other tools such as imaging thus must be correlated with what we learn from MS about the machines. Gesteland concluded that technology requirements for this goal are immense. While 2D gel electrophoresis has been the workhorse for protein analysis, it no longer meets our needs because it is too slow and not amenable to the wide range of proteins in cells. Technologies are needed for robust, high-throughput analysis of intact proteins, peptides from digests, and protein complexes. Gesteland also posed the question of whether these capabilities should be available in every lab or established in national centers such as the sequencing centers that were part of DOE's Human Genome Program.

Barbara Wold's (California Institute of Technology) presentation was given by Elbert Branscomb (Lawrence Livermore National Laboratory). He continued the overview of GTL, focusing on Goal 1, and stated that multiple protein complexes are key connectors to gene function. The goal is to catalog all protein complexes, and MS is key to reaching this goal. He cited Gerry Rubin's doctrine that there is a finite number of these complexes, a significant core set of which is thought to be similar across evolution. This is basically a "theme plus variations" concept—variant and highly conserved portions within complexes. Branscomb/Wold defined a number of key issues for Goal 1. We need to know the minimum defining stability for a complex, its prevalence threshold, and its dynamics. The least stable complexes might be the most interesting, and a way must be devised to preserve and analyze the unstable complexes. Cross-linking is one approach that may prove valuable, although others may be identified. Distinguishing "real" complexes from those that may be produced artifactually by laboratory techniques will be challenging and will require the comparison of results to other available data and the use of computational tools. Two remaining issues of critical importance to Goal 1 are how to achieve high-throughput analysis of the molecular machines and how to manage and share data.

Proteomics

John Yates (Scripps Research Institute) presented an overview of the comprehensive proteomics of complexes, cells, and tissues. He identified needs for protein identification (and localization), identification of post-translational modifications, and quantification. Although many are using 1D and 2D gel electrophoresis techniques, Yates stated that we must move away from these techniques to chromatography-based approaches using liquid chromatography (LC), capillary electrophoresis (CE), affinity chromatography, and multidimensional chromatographic separations. In a 2D LC analysis of about 2000 yeast proteins, Yates analyzed proteins of low abundance, both basic and acidic proteins,

and membrane proteins, all of which are difficult to resolve by 2D gel methods. He calls his approach "shotgun proteomics," analyzing whole cellular fractions using 2D LC and tandem mass spectrometry (MS/MS). His lab currently processes about 10^6 MS/MS spectra per week, makes extensive use of computational tools (e.g., SEQUEST) to interpret data, and uses quadrupole ion trap mass spectrometers exclusively. [Note: SEQUEST, developed in the Yates laboratory, is now commercially available.] Tandem affinity purification (TAP) tags [Rigaut et al., *Nature Biotechnol.* **10**,1030 (1999)] are employed to pull down protein complexes for analysis with Multidimensional Protein Identification Technology (called MudPIT) [Washburn et al., *Nature Biotechnol.* **19**, 242 (2001)]. Yates pointed out, however, that this approach requires multiple elutions that might break up weak complexes.

Yates concluded by identifying challenges that need to be addressed. The first is improved sensitivity that will allow visualization down to10 copies of a protein per cell. The second challenge is to improve dynamic range, which now is 1 in 10,000. At least a 10-fold improvement is required, however, to match the range of protein concentrations observed in biological systems. Improved, robust methods to identify post-translational modifications and for quantification also are needed. Techniques such as isotope coded affinity tags (ICAT) that employ isotope labeling [Gygi et al., *Nature Biotechnol.* **17**, 994 (1999)] according to Yates are "tricky and expensive." A final challenge was defined as devising ways in which MS can contribute to "real-time" analysis, thus minimizing sample preparation and handling.

Pat Griffin (Merck Research Laboratories) led a general discussion of proteomics approaches, beginning with an outline of Merck's perspective, which they refer to as "molecular profiling." Many companies are working in the area of molecular profiling which, in addition to MS, includes informatics, databases, and samples [e.g., tissues, annotations, single-nucleotide polymorphisms (SNPs), gene modification (antisense, knockouts), gene analysis (expression patterns, SNP discovery], and protein monitoring (chips, antibodies, reagents and other tools, including chemical tags and affinity hooks)]. Griffin noted that differential analysis of proteins is a growing area of importance and requires the obtainability of statistically significant differences among populations. In general, improved computational and bioinformatics methods are critically needed for this entire process, including statistical analysis of replicated data, database correlations, de novo algorithms, and faster search times. He also identified the need for improved algorithms that will identify unknown components in mixtures, rather than techniques that can work only with known components. With respect to instrumentation for proteomics and molecular profiling, he noted that across the full range of analysis, which includes sample preparation, chromatographic separations, MS, and data analysis, the developed instrumentation must be robust, low cost, and fully automated.

Richard Smith (Pacific Northwest National Laboratory, PNNL) described a proteomics method being developed in his laboratory that employs MS/MS and FTICR. Although MS/MS is typically slow, use of accurate mass tags (AMTs) circumvents this limitation. He identified the "warts" of his approach as the FTICR instrument, which is "…big and expensive, not known to be robust, finicky, and not high throughput." Although the process for establishing AMTs takes considerable up-front time, the advantage of this approach is that protein peptides can be identified quickly. This procedure uses a single, high-pressure LC separation of protein digests, enhancing the separation of peptides and reducing loss of sample. Potential mass tags (PMTs) are generated on LC-quadrupole ion trap and quadrupole time-of-flight (TOF) instruments. Smith noted the need for software to support PMT assignment, including data-management and -analysis tools. Once PMTs are known, the masses of peptides are determined to 1 ppm on the FTICR. Smith pointed out that close to 100% confidence in AMT identifications is critical, and many fractionations and measurements are required to build a reliable database. He described how this approach has been used to study *Deinococcus radiodurans,* in which more than 60% of predicted proteins were identified. A concept called DREAMS (dynamic

range enhancement applied to MS) is being developed in his laboratory to improve detection of low-abundance peptides and increase the number of peptides observed. In DREAMS, an initial spectrum is evaluated, and the most abundant ions are ejected in the next measurement to observe the less abundant species.

Robert Hettich (Oak Ridge National Laboratory, ORNL) described a protein-analysis method developed in his laboratory that integrates top-down and bottom-up approaches. In the bottom-up procedure, cellular lysates are fractionated via anion exchange LC. A portion of each fraction is digested, and the resulting peptides are analyzed by LC quadrupole ion-trap MS. Another portion of the cell lysate fraction (still containing the intact proteins) is examined top-down by FTICR. This analysis provides a means of characterizing post-translational modifications, signal peptides, and gene start sites that are not detected with bottom-up techniques involving digestion steps. This dual system has been used to identify nearly 900 proteins from *Shewanella oneidensis*, representing every functional class, including 43 previously hypothetical species. The dual system was demonstrated with a specific protein that had been identified using trypic peptides in the bottom-up approach at 85% sequence coverage. In the top-down assay, no match was found to the predicted protein's molecular mass of 29,366.7 Da. Instead, a protein with a mass of 26,444.688 Da was identified. Additional analysis using both FTICR and computational tools confirmed the removal of a 28-residue signal protein from the original protein. In other examples, the presence of post-translationally modified proteins was established by this dual approach, further demonstrating its value for confirming preliminary protein annotation of the sequenced genome. Hettich also noted the need for robust computational and bioinformatics tools to help interpret and catalog data.

Lloyd Smith (University of Wisconsin) presented a concept that could lead to the analysis of a single protein molecule. Smith stated that although proteins ideally would be ionized with 100% efficiency, currently only a small fraction of the sample is ionized. He and Michael Westphall have developed an approach that employs a nanoelectrospray source based on a piezoelectric drop-on-demand device. This allows individual droplets to be generated in a manner that he described as a "railgun" to inject the ions along a defined trajectory. The piezoelectric device is quite robust in generating reproducible droplets. Smith noted that this device could be coupled with an LC. A cubic trap prior to the mass analyzer is envisioned as a means of controlling ion injection and facilitating desolvation of the electrosprayed droplets. This device is currently under construction in his laboratory. A new aerodynamic lens stack is also being designed to assist in transfer of the ions into the mass analyzer. Initial tests with the piezoelectric electrospray (ES) source have shown that a single droplet yielding 80 amole of insulin may be detected in an 80-pL droplet. Smith envisions that in future the protein expression profile of a single cell can be established using MS.

Jean Futrell (PNNL) briefly described technologies being developed in his laboratory based on FTICR, including surface-induced dissociation (SID) in conjunction with DREAMS, new designs for the ion funnel and FTICR cell, and database-searching algorithms that can be employed with SID spectra. He stressed that these developments are a result of fundamental MS research and provide direct benefit to proteomics studies.

Protein Complexes

Joe Loo (University of California, Los Angeles) overviewed current capabilities for analyzing protein complexes. He noted that MS can support structural biology programs because it can provide insight into binding affinities as well as information on the stoichiometry of complexes by measuring molecular mass. Matrix-assisted laser desorption ionization (called MALDI) TOF has been used to

examine protein-protein interactions directly from protein chips. The protein chips have discrete spots, each containing a particular immobilized protein that will selectively bind its interaction partners from solution. Protein complexes have been examined directly from solution by electrospray ionization (ESI), although Loo noted that this approach has some limitations. First, the ionization technique, particularly using nanospray, is slow and difficult to automate. Second, the ESI process has limited compatibility with buffers and other components present under physiological conditions that would preserve protein complexes. This problem arises because under high-electrolyte conditions, the spray process is severely impacted. A problem with solubility further complicates the direct analysis of protein complexes. Each complex is unique, and experience has shown that it is difficult to identify a single condition under which complexes can be analyzed. Also, weaker complexes may not be detected by direct electrospray introduction. Finally, Loo pointed out the disadvantage of having to analyze larger complexes on an instrument with a high mass-to-charge range.

Loo described how ESI MS may be used to study the gas-phase stability of protein complexes. He also noted that the flow rate of nanospray may affect the resulting mass spectra. Reports of online protein complex analysis have been accomplished using capillary isoelectric focusing (CIEF) and by size exclusion chromatography (SEC), although separation by SEC was not as good as with CIEF. In general, he reported a success rate of about 75% for direct analysis of protein complexes (ones that remain soluble), although keeping them in solution is often difficult and requires optimization of conditions from one complex to another.

Greg Hurst (ORNL) presented a concept for scaling up the analysis of protein complexes by MS. His approach involves the use of “two-handed” cross-linkers with an affinity label that could be used to fish linked peptides produced from cross-linked protein complexes out of a much more complicated mixture, thus simplifying analysis. The two linker arms could be made different lengths to “measure” distances between residues and provide insight into the interaction interfaces between partners in the complex. This would require developing cross-linkers with high yields, and a combinatorial approach is being used to assess cross-linking under a range of reaction conditions. In addition, the method would require computational tools to assist in interpreting the resulting mass spectra. For complexes with low binding affinities, this approach may be the best to pursue, noted Yates.

David Chen (Los Alamos National Laboratory) provided insight into the analysis of protein complexes, particularly DNA-repair complexes. He noted that preparation of the complexes is critical and that any method must ensure that identified complexes are authentic and not artifacts.

Mark Biggin (Lawrence Berkeley National Laboratory, LBNL) added comments about analysis of protein complexes. He stated that methods for isolating protein complexes are not optimized in a way that preserves all-important complexes and prevents isolation of artifactual, contaminating polypeptides that are biologically irrelevant members of a complex. In fact, there may be no perfect solution to these problems, which will continually bedevil high-throughput assays. Biggin pointed out that any high-throughput analysis of protein complexes must be tightly integrated to genome-wide analysis of other data, including protein-expression profiling, imaging, and gene-network studies. These data can provide supporting evidence for which putative protein-protein interactions are “real.” This will require that various types of data be generated in a single organism, and bioinformatics will be needed to build an integrated database and conduct network modeling.

Sample Preparation

In giving a general overview of sample-preparation issues for protein-complex analysis, Katheryn Resing (University of Colorado) pointed out the importance of developing robust protocols. The purification of protein complexes is a limiting factor for the “working” biologist. Currently, complexes

are targeted one at a time and identified via binding to an oligonucleotide sequence. Complexes typically are purified using affinity purification, usually with antibodies to one or more components of the complex, along with low nonspecific binding resins (e.g., magnetic beads); high-affinity antibodies are required. Recently, expression vectors to make "tagged" proteins have been used to purify complexes with targeted proteins. Resing has done some preliminary work with DNA aptamers generated by systematic evolution of ligands by exponential enrichment (SELEX) and these have affinities and specificities comparable to good antibodies. [Note: In the SELEX system (McGown et al., *Anal. Chem*. **67**, 663A–668A, 1995), a number of DNA oligonucleodies are passed over a column with an immobilized protein, and those that bind to the protein over several cycles are retained. They can then be sequenced, and large numbers can be produced by synthesis or PCR for use as affinity agents.]

Resing recommended that when pulling down proteins, antibodies are not as effective as SELEX. If antibodies are required, she uses magnetic beads. Once the protein complexes are isolated, they typically are analyzed either by separation on gels followed by in-gel digests and MS or by complex digestion followed by peptide-mixture analysis with LC MS. She noted that the top-down, bottom-up, and AMT methods use high resolution to take a "snapshot" of all proteins.

Another suggestion was to use a cross-linking agent to "fix" the complexes and then identify the cross-linked peptide (described by Hurst). Resing noted that this approach would have the advantage of easy adaptability to analysis of samples under many experimental conditions. Three analytical strategies were proposed for identifying the cross-linked proteins: (1) computationally analyze the peptide masses and physical chemical properties in uncross-linked vs cross-linked samples, looking for new peptide masses; (2) use a cross-linking agent containing biotin (or other affinity tag) to pull out the cross-linked peptides; or (3) use a cleavable linker, resolve the complexes on a 1D SDS-polyacrylamide gel electrophoresis (PAGE) gel, then cut out and place the gel on top of a second gel, cleaving the linkers before a second SDS-PAGE, where components of complexes would run together vertically. Several variations on these strategies are possible. Resing outlined a number of things on her "wish list," including better complex-identification and -affinity methods; techniques for weakly bound complexes; more useful ways of addressing stoichiometry, heterogeneity, and the analysis of intact proteins; improved reagents (especially cell-permeable ones) for cross-linking; and better computational procedures.

David Goodlett (Institute for Systems Biology) also addressed issues in sample preparation. The strategy employed in his laboratory includes tryptic digests of proteins from a mixture separated by microbore LC and analyzed by tandem MS (triple quadrupole instrument). He employs what he calls "gas-phase fractionation," which analyzes the samples many times over different mass-to-charge (mz) ranges to obtain increased coverage, essentially eliminating most intense peaks to get improved dynamic range. An important part of the approach is a multiprocessor database search to identify the proteins. Samples are prepared using either in vivo or in vitro stable ICAT, which provide quantification capabilities and improved identification by introducing constraints in the database search. The proteins are initially digested with trypsin, and fractions are generated. The fractions are separated into ICAT-labeled peptides (containing Cys) and avidin affinity flow-through peptides (not containing Cys); these peptides are subsequently analyzed by MS/MS. This approach has been used to identify differences in the Jurkat cell lipid raft proteins before and after stimulation. Goodlett commented that this generates a "ton of data" and pointed to the need for more programs to cull out the data. After a discussion on centralized and decentralized approaches to software development, attendees agreed that a decentralized approach would be better.

Roland Annan (GlaxoSmithKline) led a general discussion on preparation techniques. He noted that this type of work needs to be done in "factories" and that these factories will have multiple mass spectrometers to attack the problem in parallel, with big computers all dedicated to interpretation of spectra.

He stated that with methods such as the use of single and dual tags as affinity-capture "handles" now in place for the front end, the next steps are scale-up and automation. He also noted that sample preparation should be kept close to the MS laboratory because variations in preparation may impact results in repeat experiments. The general consensus was that MS analysis will be relatively easy and fast. The rate-limiting step is not in collecting MS data but in preparing the sample and then analyzing the thousands of resulting spectra. Sample preparation and data analysis will be much harder, but these two latter capabilities will be the key to success.

Computational Biology

Ed Uberbacher (ORNL) led a discussion on computational tools that will be needed in tandem with MS for the characterization of protein complexes. Three areas identified were data analysis, data management and sharing, and characterization of protein complexes. Data analysis must be able to address such issues as post-translational modifications, quantification, reproducibility, and comparison and integration of data with those from other techniques. Cross-linking strategies hold considerable promise for MS studies of protein complexes, but new algorithms will be needed to identify constituents in database searches and pick out transient complexes. Data management must include ways to share data and databases with the scientific community, although some data will be archived locally. New approaches to represent data will be required. The needs with respect to characterization of protein complexes are huge. For example, how can data give a picture of the complexes? How do we get from single observations to a description of complex dynamics? Using cross-linking to provide constraint for additional information about the complex is very important. Several in the audience noted that critical computational tools must be developed in close collaboration with the MS and biology experts, not in a vacuum. Computational experts should be immersed in the area and sitting with MS and biology experts. Computational tools will be vitally important in integrating MS and other data and then linking that data to biological function. We need to think about what we are going to capture before we start—these are the same discussions held 15 years ago when DOE's Human Genome Program started.

Bill Cannon (PNNL) described computational approaches for deciphering cellular networks that included peptide state, protein state, state of the cell, and cellular networks. Ying Xu (ORNL) described downstream computational requirements, including using data-constrained (from MS data with cross-linking, for example) protein threading to help predict structure and docking.

Summary

Workshop participants agreed that to meet GTL goals, a two-pronged, parallel approach will be needed initially:

(1) Use present technologies to categorize protein complexes (a "brute force" procedure).
(2) Develop new technologies, especially sample preparation, automation, and computational tools, to move these investigations into the robust, high-throughput mode required for GTL program success.

Branscomb noted that getting off the ground, even with an aggressive but perhaps sloppy approach, is probably the best way to start the process. The data coming out would enable the biology community to communicate about how to meet the community's needs. He suggested that perhaps a goal of

getting 50 to 60% of the complexes from a single organism might be a good start, a draft product that would help demonstrate feasibility.

Attendees generally agreed that the time and effort needed to obtain the MS data will not be the critical issue with respect to high-throughput analysis of proteins and protein complexes. The real challenges will be the steps involved in preparing samples and interpreting data. A number of improvements in MS technologies, however, were identified to increase throughput and information content, especially in sensitivity and dynamics, to allow visualization of low-abundance complexes. Additional technologies mentioned were enhancements of sample introduction, ionization, and detection capabilities. Robust and automatable sample-preparation techniques that maintain protein-complex integrity are critically important to this endeavor's success. Areas identified as needing attention in the near term include better ways of isolating the complexes (including "pulling down" and cross-linking), devising standard methods for cell preparation to improve run-to-run reproducibility, and procedures for quantification.

In response to the question of what computational tools are needed, the answer was literally "a lot!" A key point, however, is that computational tools can be used to reduce the demands on analytical technology and to increase both analysis throughput and information content. This obviously is an area where careful planning is needed, with close collaboration between the experimental and computational communities. Although new computational methods will be required in many areas for Goal 1, specific needs for the near term are (1) more efficient algorithms for database searching and statistical data analysis, (2) data-visualization tools specifically designed to characterize nodes in networks (e.g., interpretation of data from complex components represented by a set of multiple cross-links among many protein members) and algorithms designed to identify significant changes in these networks upon experimental cell manipulation, and (3) a format that is available to other researchers.

Finally, participants noted that high-throughput analysis of protein complexes should be tightly integrated to genome-wide analyses of other types of data such as protein-expression profiling and gene network studies that can provide supporting evidence for which putative protein-protein interactions are "real." This will require a large bioinformatics effort to build an integrated database and conduct network modeling. Furthermore, complementary data from such analytical techniques as microscopy, NMR, neutron scattering, and others can provide additional insight into the interactions of protein complexes present in microbial cells.

The role of MS in Goal 1 extends far beyond just identifying the full complement of proteins and protein complexes in a cell. MS can be used to monitor changes in proteins and protein complexes within the cell in response to a specific stimulus, thus providing insight into stoichiometry and heterogeneity, post-translational modifications, and other alterations that will yield information on the role of specific protein interactions in cellular biochemical networks and pathways. Additional information may be obtained by identifying the complexes' potential interaction interfaces (i.e., the chemistry between two interacting partners) and by combining MS and other experimental data with computational methods (structure prediction, surface disorder, and docking).

Michelle Buchanan, Oak Ridge National Laboratory, Workshop Organizer

Appendix 1

DOE-OBER GTL Mass Spectrometry Workshop
Marriott Wardman Park Hotel, McKinley Room
Washington, D.C.
December 10–11, 2001

Monday, December 10

8:00 a.m.	Coffee and Sweet Rolls
8:30 a.m.	Welcome Marvin Frazier, DOE OBER
8:40 a.m.	Workshop Goals Michelle Buchanan, Oak Ridge National Laboratory
8:50 a.m.	Genomes to Life: Overview and Discussion Ray Gesteland, University of Utah
9:05 a.m.	Task 1: Overview and Discussion Barbara Wold, California Institute of Technology (Note: Elbert Branscomb substituted for Dr. Wold – she was detained by a family medical emergency and could not attend)
9:45 a.m..	MS and Proteomics: Overview of Status John Yates, Scripps Research Institute
10:15 a.m.	Break
10:30 a.m.	Proteomics: General Discussion of Strategies and Breakthroughs Needed Patrick Griffin, Merck Research Laboratories Dick Smith, PNNL, and Robert Hettich, ORNL (Participants Encouraged to Bring Viewgraphs to Talk about Other Approaches)
11:30 a.m.	Group Lunch
12:30 p.m.	General Discussion on Proteomics (cont.) Lloyd Smith, Univ. of Wisconsin
1:30 p.m.	Protein Complexes: Overview of Status Joe Loo, University of California, Los Angeles
2:15 p.m.	Break
2:30 p.m.	General Discussion: Protein Complex Strategies, Breakthroughs Needed Katheryn Resing, University of Colorado Greg Hurst, ORNL (Participants Encouraged to Bring Viewgraphs to Talk about Other Approaches)
5:00 p.m.	Adjourn for Evening

Tuesday, December 11

8:00 a.m.	Coffee and Sweet Rolls
8:30 a.m.	Sample Preparation: Issues and Needs Dave Goodlett, Institute for Systems Biology
9:00 a.m.	General Discussion: Sample Preparation Roland Annan, SmithKline Beecham Pharmaceuticals
9:30 a.m.	General Discussion: Data Analysis: Approaches and Needs Ed Uberbacher, ORNL
10:30 a.m.	Break
10:45 a.m.	Summary of Needs to Accomplish Goal 1 of GTL Program Michelle Buchanan, ORNL
11:30 a.m.	End of Workshop

Attendees
DOE Genomes to Life Mass Spectrometry Workshop
December 10-11, 2001
Marriott Wardman Park Hotel, Washington, D.C.

Dr. Carl W. Anderson
Biology Department
Brookhaven National Laboratory
Upton, New York 11973

Dr. Roland S. Annan
SmithKline Beecham Pharmaceutical
Post Office Box 1539, UW-2940
King of Prussia, Pennsylvania 19406

Dr. W. Henry Benner
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Mailstop 70A-3363
Berkeley, California 94720

Dr. Mark Biggin
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Mailstop 84-171
Berkeley, California 94720

Dr. Elbert W. Branscomb
Lawrence Livermore National Laboratory
7000 East Avenue
Post Office Box 808
Livermore, California 94551-0808

Dr. Michelle V. Buchanan
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, Tennessee 37831-6129

Ms. Brenda W. Campbell
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, Tennessee 37831-6129

Dr. William R. Cannon
Pacific Northwest National Laboratory
W. R. Wiley Environmental Molecular
Sciences Laboratory
Post Office Box 999, K1-83
Richland, Washington 99352

Dr. David Chen
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Mailstop 74-157
Berkeley, California 94720

Dr. Dan Drell
OBER (SC-72)
U.S. Department of Energy
19901 Germantown Rd..
Germantown, Maryland 20874-1290

Dr. Charles Edmonds
National Institute of General Medical Sciences
National Institutes of Health
45 Center Drive MSC 6200
Bethesda, Maryland 20892-6200

Dr. Brendlyn Faison
OBER (SC-74)
U.S. Department of Energy
19901 Germantown Rd..
Germantown, Maryland 20874-1290

Dr. Marvin Frazier
OBER (SC-72)
U.S. Department of Energy
19901 Germantown Rd..
Germantown, Maryland 20874-1290

Dr. Jim K. Fredrickson
Pacific Northwest National Laboratory
Post Office Box 999, MS P7-50
Richland, Washington 99352

Dr. Jean Futrell
Pacific Northwest National Laboratory
W. R. Wiley Environmental Molecular
Sciences Laboratory
Post Office Box 999 (K8-84)
Richland, Washington 99352

Dr. Raymond F. Gesteland
Department of Human Genetics
University of Utah
15 N. 2030 E, Room 7410
Salt Lake City, Utah 84112-5330

Dr. David Goodlett, Director
Institute for Systems Biology
4225 Roosevelt Way, NE
Seattle, Washington 98105

Dr. Patrick R. Griffin
Merck Research Laboratories
P.O. Box 2000, RY 121-146
Rahway, New Jersey 07065

Dr. Robert L. Hettich
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, Tennessee 37831-6365

Dr. Roland Hirsch
OBER (SC-73)
U.S. Department of Energy
19901 Germantown Rd..
Germantown, Maryland 20874-1290

Dr. Greg B. Hurst
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, Tennessee 37831-6365

Dr. Arthur Katz
OBER (SC-72)
U.S. Department of Energy
19901 Germantown Rd..
Germantown, Maryland 20874-1290

Dr. Michael L. Knotek
Consultant
10127 N. Bighorn Butte Dr.
Oro Valley, Arizona 85737

Dr. Dean Cole
OBER (SC-73)
U.S. Department of Energy
19901 Germantown Rd..
Germantown, Maryland 20874-1290

Dr. John C. Houghton
OBER (SC-74)
U.S. Department of Energy
19901 Germantown Rd..
Germantown, Maryland 20874-1290

Dr. Peter Kirchner
OBER (SC-73)
U.S. Department of Energy
19901 Germantown Rd..
Germantown, Maryland 20874-1290

Dr. David Koppenaal
Pacific Northwest National Laboratory
W. R. Wiley Environmental Molecular
Sciences Laboratory
Post Office Box 999 (K8-98)
Richland, Washington 99352

Dr. Joseph Loo
University of California-Los Angeles
402 Paul Boyer Hall
Post Office Box 951569
Los Angeles, California 90095-1569

Dr. Noelle Metting
OBER (SC-72)
U.S. Department of Energy
19901 Germantown Rd..
Germantown, Maryland 20874-1290

Dr. Douglas Ray
Pacific Northwest National Laboratory
P. O. Box 999 / MS K9-90
Richland, Washington 99352

Dr. Katheryn Resing
University of Colorado
Department of Chemistry and Biochemistry
Post Office Box 215
Boulder, Colorado 80309

Dr. Lloyd Smith
University of Wisconsin
Department of Chemistry
3335A Chemistry Bldg.
1101 University Avenue
Madison, Wisconsin 53706

Dr. Richard D. Smith
Pacific Northwest National Laboratory
W. R. Wiley Environmental Molecular
Sciences Laboratory
Post Office Box 999 (K8-98)
Richland, Washington 99352

Dr. Prem C. Srivastava
OBER (SC-32)
U.S. Department of Energy
19901 Germantown Rd.
Germantown, Maryland 20874-1290

Dr. Marvin Stodolsky
OBER (SC-72)
U.S. Department of Energy
19901 Germantown Rd.
Germantown, Maryland 20874-1290

Dr. David Thomassen
OBER (SC-72)
U.S. Department of Energy
19901 Germantown Rd.
Germantown, Maryland 20874-1290

Dr. Edward C. Uberbacher
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, Tennessee 37831-6480

Dr. Gary J. Van Berkel
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, Tennessee 37831-6365

Dr. Mike Viola
OBER (SC-73)
U.S. Department of Energy
19901 Germantown Rd.
Germantown, Maryland 20874-1290

Dr. Michael Westphall
Department of Chemistry
University of Wisconsin
1101 University Ave.
Madison, Wisconsin 53706

Dr. John Yates, III
Scripps Research Institute
Department of Cell Biology, SR11
10550 North Torrey Pines Road
La Jolla, California 92037

Dr. Ying Xu
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, TN 37831