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 Recently, labor-intensive, time-consuming gel electrophoresis has been replaced by liquid chromatography, greatly improving total analysis time.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.