Ashok R. Dongre
Gregory Opiteck
Wesley L. Cosand
Stanley A. Hefta
Department of Applied Genomics,
Bristol-Myers Squibb Pharmaceutical Research Institute,
Princeton, NJ 08543

Proteomics in the Post-Genome Age*

Abstract: *The genome sequencing effort has helped spawn the burgeoning field of proteomics. This review article examines state-of-the-art proteomics methods that are helping change the discovery paradigm in a variety of biological disciplines and, in particular, protein biochemistry. The review discusses both classical and novel methods to perform high-throughput qualitative and quantitative "global" as well as targeted proteome analysis of complex biological systems. From a drug discovery standpoint, the synergy between genomics and proteomics will help elucidate disease mechanisms, identify novel drug targets, and identify surrogate biomarkers that could be used to conduct clinical trials.* © 2001 John Wiley & Sons, Inc. Biopolymers (Pept Sci) 60: 206–211, 2001

Keywords: *proteomics; proteome; drug discovery; multidimensional chromatography; liquid chromatography–mass spectrometry; protein analysis; gel electrophoresis; post-genomics; genome*

INTRODUCTION

The advent of Expressed Sequence Tag (EST) sequencing and the Human Genome Initiative changed the face of protein chemistry forever. With databases of expressed and predicted genes, the experimental challenge of identifying a protein has changed to recognizing a sequence that already resides in a database. Instead of sequencing a protein in its entirety, for many purposes it is only necessary to recognize fragments of the protein to confirm its presence. Since this recognition can be done by mass spectrometry, it is possible to work with very complex mixtures that the analytical technology can deconvolute.[1–4]

This new experimental technology has led protein biochemists to change how we think about biological problems. Instead of spending months to characterize a given protein, we aspire to characterize all of the proteins expressed by the genome under a given set of biological conditions. That is to say, we would like to

* This contribution is dedicated to the team of Bruce and Libby Merrifield on the occasion of his 80th birthday.

Biopolymers (Peptide Science), Vol. 60, 206–211 (2001)

Correspondence to: Wesley L. Cosand; email: wesley.cosand@bms.com

Biopolymers (Peptide Science), Vol. 60, 206–211 (2001)
© 2001 John Wiley & Sons, Inc.

characterize the proteome that exists under a set of biological conditions. In this paper, we explain some of the recent advances in the discipline that convince us that this is not unjustifiable hubris, but rather a rational although ambitious vision of the near future of protein characterization.

Why is this a useful endeavor? It is now well accepted that a single genome can give rise to qualitatively and quantitatively different proteomes under different biological conditions. DNA chips have proven to be a powerful tool to profile the amounts of mRNA resulting from each gene.[5–8] Why is this not a sufficient characterization of the biological state of a system? One answer to this question is that there is only very loose coupling between the level of mRNA and the level of expressed protein resulting from that mRNA.[9] In addition, there are other important mechanisms for the regulation of a biochemical pathway that can be investigated only by characterizing the proteins involved. These include mechanisms such as post-translational proteolytic processing and post-translational modifications such as phosphorylation and acylation. At an even higher level, the activity of a given molecule may be dependent on which of multiple partners are concurrently expressed in a cell and are available for the formation of multiprotein complexes. All of these regulatory mechanisms are quite amenable to investigation with the tools of proteomics.[10–12]

What changes in science and technology have made the "global" characterization of the proteome a reasonable goal? The most obvious one is the acquisition of the sequence of the genome, although the formidable job of identifying the 1.5% of the genome that codes for protein is widely underappreciated. Second, the development of algorithms for searching sequence databases with uninterpreted mass spectral data has made possible the automation of protein identification. The identification of proteins (protein fragments) in real time as they are still eluting from a liquid chromatography–mass spectroscopy (LC/MS) interface makes it possible to use this information to control the instrument. Although this remains a computationally intensive task, the availability of inexpensive computing power has made this capability affordable by more than just the best funded laboratories.

Finally, advances in mass spectrometry resulting in higher sensitivity and greater mass resolution and progress in micro column chromatography all have had very significant roles in laying the foundation for this new discipline.

TECHNOLOGY PLATFORMS FOR PROTEOMICS

Proteomics technologies can be divided into two broad categories: technologies associated with sample separation and quantification, and technologies associated with sample identification.[13–15]

The first step in most proteomic analyses involves separating a mixture of proteins by two-dimensional (2D) gel electrophoresis, and quantifying the individual components by staining and imaging technologies. Two-dimensional gel electrophoresis technology was developed over 25 years ago and improved in the late 1970s and 1980s to achieve high resolution, reproducible protein separation. Complex protein mixtures are applied to tube gels and separated in the first dimension based on their isoelectric points. The gel is then placed on top of a polyacrylamide slab gel and the proteins are electrophoresed into the gel and resolved according to their molecular weight. As a result, the two dimensions of migration are according to isoelectric point, or the net charge of the protein, and the molecular weight of the protein. Advances have been made in this basic technology by substituting the first dimension tube gel with polyacrylamide gel "strips" containing immobilized ampholytes to form a pH gradient from one end of the strip to the other. These strips are more durable than the tube gels, easier to handle, and produce more reproducible results because of the immobilized ampholytes. Currently, there are several companies offering the hardware and precast materials for running 2D gels.

Visualization of the protein "spots" requires staining of the proteins by one of a variety of methods. Classically, Coomassie and silver staining were used. However, several fluorescent dyes have recently been introduced, which provide good quantification and linear dynamic range. Silver staining remains the most sensitive nonradioactive method capable of staining proteins in the low nanogram ($<$10 ng) level; however, quantification by silver is problematic. The fluorescent stains are nearly as sensitive ($\sim$10 ng), but offer a greater range of linear response than silver. Coomassie offers the least sensitivity of these stains, requiring $>$100 ng of material to visualize.

Analysis of the staining patterns, and quantification of the relative amounts present, is the next step in this study. Several specialized software packages are available to collect images, detect spots, perform quantitative analysis, compare multiple gels, and even generate composite 2D gel images from several individual gels. Some of the more comprehensive software packages on the market are PDQUEST from BioRad and BioImage from Genomic Solutions. Biotechnology companies that specialize in proteomics, such as Oxford GlycoSciences (OGS) and Large Scale Proteomics (LSP), have custom designed hardware and software capabilities that allow them to run, stain, and analyze (image and quantify) 2D gels in a high-throughput mode. Additionally, several other biotech startups that specialize in proteomics are setting up proprietary platforms to analyze 2D gels. However, such a high-throughput 2D-gel analysis package is not commercially available at present.

Because of the high costs of labor and reagents required for producing highly reproducible 2D gels, a number of laboratories are attempting to replace gels with a multidimensional chromatography system.[2,4] Such an approach gives up the use of staining and densitometry to quantify proteins. These systems are varied in design, but they share a direct interface between the sample separation steps (chromatography) and analysis steps (mass spectrometry). Because of this direct interface, such systems limit sample loss, leading to several hundredfold increased sensitivity. Another factor driving such research is that this technology is much more amenable to automation than 2D gels. These technologies are the subject of research in academic analytical laboratories and within industrial settings they are beginning to be applied to important biological problems.

Regardless of the mode of separation, the selected proteins must be processed for identification. Classically this identification was performed by transferring the protein spot to a membrane followed by N-terminal sequencing. However, this technique is slow, tedious, and less sensitive than the currently available state-of-the-art mass spectrometric (MS) techniques. The late 1980s and early 1990s saw an explosion of application of mass spectrometric techniques to biological samples. This proliferation in utility is mainly credited to the invention of two ionization techniques: matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI). These methods made it possible to ionize and thereby detect large biological molecules using a mass spectrometer with reasonable throughput. Furthermore, ESI techniques made it possible to interface chromatographic systems to mass spectrometers. The middle to late 1990s saw improvements in sensitivity, automation, and throughput of these mass spectrometry based protein identification techniques. These improvements are partially due to development of nanoscale chromatography devices. However, a factor that is at least as important was the development of software that made possible the automated searching of protein and DNA sequence databases by raw experimental mass spectrometry data. The sum total of these advances was protein identification in a matter of seconds, rather than days or months as required by N-terminal sequencing.

There are two distinct biological mass spectrometry platforms for analyzing proteins: MALDI-TOF-MS; and liquid chromatography-electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS). In many proteomics laboratories one platform is preferred. However, we believe that a comprehensive proteomics analysis is advantageous, as the data obtained from each platform can either provide mutual validation or can be complementary, as in those cases where data from both platforms is required to make an unambiguous protein identification.

Our protocol involves using a "layered approach" to protein identification. After protein spots are excised from a 2D gel, they are enzymatically fragmented with proteases. The resulting peptides are extracted from the gel matrix and analyzed by MALDI-TOF-MS peptide mass fingerprinting. In this type of analysis, the masses of the cleaved peptides are measured and then the protein is identified by comparing these empirical masses against a database of the masses of all theoretical proteolytic fragments. This technique is both sensitive (low femtomole levels) and affords high throughput. In the current format, 96 proteins can be identified in a matter of 2–3 h. Advances continue to be made in this technology that will permit the analysis of 384 protein digests in approximately the same amount of time in the near future. This type of approach is generally successful for identifying 40–50% of the protein samples without additional analysis. Some proteins, however, are not identifiable by this method for any number of reasons. The protein spots that fail to be identified, or that yield an ambiguous protein identification, are then further analyzed by LC-ESI-MS/MS. This technique produces data that can be interpreted to obtain the direct peptide sequence and thereby lead to unambiguous protein identifications. Tandem mass spectrometry is also employed to identify sites of posttranslational modification, including phosphorylation, methylation, and acetylation. The technique, however, is considerably slower than the MALDI-MS peptide mass fingerprinting and, unless automated, is not amenable to high-throughput applications.

A critical component of high-throughput proteomics is the informatics infrastructure necessary for tracking samples through the various processing and handling steps involved in such analysis, and the bioinformatic tools needed for higher level analyses of the results obtained. Software has been developed for specific steps in the process, but few programs are available for compiling the information generated from multiple instruments throughout the process. Recently, BioRad and Micromass have coordinated efforts to develop software for performing this function. This software supports sample tracking, and provide tools for the comparison and integration of data generated in multiple fields, such as genomics, proteomics, biochemistry, and pharmacology. Furthermore, the open architecture feature gives the end user the ability to modify and/or include other proprietary packages that are not supported by either BioRad or Micromass. Other commercial software packages are also being developed by a variety of companies.

An important and essential component of this protein informatics platform is mass informatics. Mass informatics is a compilation of various software packages including database search programs to analyze mass spectral data, data management programs, and archival tools. The database search programs take mass spectral data and correlate it with genomic and protein databases. These programs are broadly categorized into peptide mass fingerprinting programs, which use MALDI-MS data, and peptide sequencing programs, which use LC-ESI-MS/MS data. SEQUEST™[16] and MASCOT™[17] are two such software packages that have the capability to handle large amounts of data from a high-throughput proteomics effort and are both commercially available. SEQUEST, marketed by Finnigan Corporation, is a specialized peptide sequencing program that uses peptide MS/MS data. MASCOT, sold by Matrix Science (UK), is able to handle both peptide MALDI-MS and MS/MS data. SEQUEST is scaleable and is available from Thermo-Finnigan in a version, PVM-SEQUEST (parallel virtual machine-SEQUEST), which provides truly high-throughput search capabilities. This PVM version can run on a LINUX Beowulf computer cluster, an arrangement that can process an enormous volume of mass spectral files. For example, on a typical day the four tandem mass spectrometers in our proteomics laboratory can acquire and generate 20,000 MS/MS data files. Using PVM-SEQUEST each MS/MS data file can be searched against all known protein or EST sequences in 1–1.5 s.

The next component in mass informatics is managing and archiving mass spectral data. Currently, there is no commercial software package available to efficiently perform these tasks. Proteomics companies like OGS and LSP have invested huge resources to set up proprietary data management and archival schema. The first commercial package that meets these needs may be WorksBase from Bio-Rad. WorksBase uses an Oracle relational database architecture to efficiently perform both data management and data archival tasks.

IMPACT OF PROTEOMICS IN DRUG DISCOVERY

Our experience is with a department of proteomics that operates as an integral part of a genomics division within a pharmaceutical company. We view this setting as a marked advantage because of the collaborative opportunities to bring a wide choice of complementary technologies to bear on biological problems. Methodology development occurs as a result of needing better experimental approaches to these biochemical questions. It might be instructive to look at several systems where this technology has been applied in our laboratories.

One of the first areas where proteomics made a significant contribution to our work was in the area of microbial biology. We have had bacterial genomes for several years, and these systems have the marked advantage of a smaller number of proteins and the accumulated knowledge of many years of thorough characterization of their biochemical pathways. Bacterial cell death is not a good indicator of antimicrobial activity in modern high-throughput compound screening. There was a need to find a technology that would yield a fluorescent signal when a given biochemical pathway was inhibited. For example, the gene RpoS is a central regulator in "reconfiguring" bacterial metabolism, resulting in increased survival of the bacteria in the presence of antibiotics. If it were possible to find bacterial genes that were overexpressed whenever the RpoS pathway was inhibited, the promotor of such a gene could be used to drive the expression of a fluorescent protein, providing a technology to screen for RpoS inhibitors. To discover such genes, we undertook to grow a strain that possessed a temperature-sensitive mutation in the RpoS gene in parallel with the control strain. When the proteins of these preparations were analyzed on 2D gels, it was evident that a set of genes were expressed at markedly higher levels in the strain lacking Rpos activity, and a set of genes were expressed at lower levels. A number of similar experiments helped confirm the reproducibility and specificity of these expression changes. The proteins of these differentially expressed spots were excised from the gels, digested with trypsin, and analyzed by LC-MS/MS and identified using SEQUEST. The result was the identification of more than ten genes from which the microbiologists could choose genes for the preparation of reporter gene constructs. Our antimicrobial scientists viewed this contribution as a major advance for whole cell screening for antimicrobial compounds.

Recently scientists at most pharmaceutical companies have been investigating signal transduction pathways as a route to developing more specific drugs. Most of these pathways involve protein phosphorylation, and so these investigators have sought out proteomics scientists as essential collaborators in this effort. One can get an overview of the changes that occur in phosphorylated proteins with Western blots of proteins from ligand-stimulated cells using antiphosphotyrosine antisera. These blots can be compared to parallel gels stained with fluorescent dyes from which the spots can be excised for identification by mass spectrometry. Another approach we have found useful is to fragment the protein preparation with trypsin, use metal ion affinity resins to capture phosphorylated peptides, elute them with a metal ion solution, and then characterize the peptides by MS. However, most recently we have approached this not by separating out the phosphorylated peptides, but by concentrating on improving the chromatographic resolution of our LC/MS/MS technology, collecting the data in an automated fashion, and writing software scripts that select only MS/MS data where the characteristic fragmentation demonstrating the loss of a phosphate group is detected. Our biology colleagues are finding these data indispensable.

Much of the excitement surrounding proteomics has been generated by the ability to characterize protein–protein interactions. Some biotechnology companies are undertaking the goal of using technologies based on yeast 2 hybrid studies to determine the global network of protein–protein interactions. Our goals have been more modest. We have used immunoprecipitation of protein complexes to obtain material for LC-MS/MS characterization. One recent project was designed to characterize the protein cofactors of nuclear hormone receptors. When a steroid binds to a nuclear hormone receptor (NHR), the receptor dimerizes and then the dimer binds to the Hormone Response Element motif on DNA. Cofactors then bind to this complex, initiating transcription. Our colleagues would like to understand the role tissue specificity plays in determining which cofactors are involved. An important factor in designing experiments to answer this question is the low concentration of these cofactors in cells. An approach that relied on running 2D gels and then characterizing the spots on the gels would be quite challenging. Recently our collaborators gave us an immunoprecipitate from tissue cultured cells, and rather than running gels, we chose to digest the entire preparation with trypsin, load the sample on our LC-MS/MS apparatus, and then run a 2 h reverse phase gradient. We used the Finnigan/Micromass Mass Dependent Data Acquisition software to automate the capture of the data, which was then analyzed by SEQUEST. The software reported the presence in the sample of every cofactor of this nuclear hormone receptor that had been reported in the literature, but the report also included two other proteins that were known to be nuclear proteins but otherwise were poorly annotated. When these data were reported to our collaborator, we were informed that our data confirmed studies conducted over the past 18 months in their laboratory that had demonstrated the roles of these two proteins as NHR cofactors.

FUTURE PROSPECTS FOR PROTEOMICS

Experiments like these have convinced us of the power of methods that avoid 2D gels as a separation step. Consequently, we are investing in methodological research in multidimensional micro chromatography to resolve the very complex mixtures that result when a protein preparation from a subcellular fractionation is treated with a proteolytic enzyme. These efforts are beginning to come to fruition and we are convinced that micro chromatography represents technology that will replace gels for many applications.

The role of software development to enable proteomics by mass spectrometry cannot be overemphasized. An entirely automated data analysis pipeline represents the “price of admission” to this experimental discipline today and, while such a pipeline is not commercially available, there are vendors working on such an integrated work environment. Beyond this achievement of a pipeline, we believe that there are significant advances in the analytical power of the technology that will result from current efforts in software development.

What does the future hold for the application of this technology to answering biological questions? Within the pharmaceutical industry, there is the hope that proteomics will be able to provide a solution to the need for surrogate biomarkers that could be used to conduct clinical research in areas of largely unmet medical need. Clinical trials in diseases such as Alzheimer’s or atherosclerosis are challenging to conduct because it is difficult to detect a change in the disease state. If it were possible to find proteins in the blood whose presence or concentration was an indication of an improvement of the disease state, one might be able to conduct clinical research in these diseases far more effectively. Clinical research is so expensive that any technology that ameliorates the huge risks of clinical development results in greatly improved business value. So it is not surprising that in pharmaceutical proteomics laboratories more effort may be directed toward this goal than to all others combined. But beyond the legitimate business drivers that justify this research, an advance that shortened the time necessary to conduct trials in these diseases, and that enabled more such clinical trials to be conducted, would certainly be a major step toward our transcendent corporate mission "to extend and enhance human life."

REFERENCES

1. Haynes, P. A.; Yates, J. R., III. Yeast 2000, 17, 81–87.
2. Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, M. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. Nature Biotechnol 1999, 17, 676–682.
3. Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. Proc Natl Acad Sci USA 1996, 93, 14440–14445.
4. Gygi, S. P.; Rist, B.; Gerbe, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. Nature Biotechnology 1999, 17, 994–999.
5. DeRisi, J. L.; Iyer , V. R.; Brown, P. O. Science 1997, 278, 680–686.
6. Lashkari, D. A.; DeRisi, J. L.; McCusker, J. H.; Namath, A. F.; Gentile, C.; Hwang, S. Y.; Brown, P. O.; Davis, R. W. Proc Natl Acad Sci USA 1997, 94, 13057–13062.
7. Shalon, D.; Smith, S. J.; Brown, P. O. Genome Res 1996, 6, 639–645.
8. Zong, Q.; Schummer, M.; Hood, L.; Morris, D. Proc Natl Acad Sci USA 1999, 96, 10632–10636.
9. Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. Mol Cell Biol 1999, 19, 1720–1730.
10. Williams, K. L. Electrophoresis 1999, 20, 678–688.
11. Hochstrasser, D. F. In Proteome Research: New Frontiers in Functional Genomics; Wilkins, M. R., Williams, K. L., Appel, R. D., and Hochstrasser, D.F., Eds. Springer-Verlag: New York, 1997; pp 187–219.
12. Gooley, A. A.; Packer, N. H. In Proteome Research: New Frontiers in Functional Genomics; Wilkins, M. R., Williams, K. L., Appel, R. D., and Hochstrasser, D.F., Eds. Springer-Verlag: New York, 1997; pp 65–91.
13. Jensen, O. N.; Wilm, M.; Shevchenko, A.; Mann, M. Methods Mol Biol 1999, 112, 513–530.
14. Yates, J. R., III; Carmack, E.; Hays, L.; Link, A. J.; Eng, J. K. Methods Mol Biol 1999, 112, 553–569.
15. Jensen, O. N.; Wilm, M.; Shevchenko, A.; Mann, M. Methods Mol Bio 1999, 112, 571–588.
16. Eng, J. K.; McCormack, A. L.; Yates, J. R., III. J Am Soc Mass Spectrom 1994, 5, 976–989.
17. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Electrophoresis 1999, 18, 3551–3567.