Proteomics
from a drug discovery perspective

The genomic sequencing efforts spawned the field of proteomics by providing the basic blueprint for protein sequences, enabling proteomics to become the contemporary technology for studying the underlying mechanisms of disease. This review will address the technologies related to proteomic research and the application of such research.

By Dr Ashok R. Dongre and Dr Stanley A. Hefta

The research and development of drugs is a complex and challenging process requiring the integration of many disciplines and technologies. The past century saw a tremendous growth in pharmaceutical products that were mainly driven by chemistry, pharmacology, and clinical sciences. While these sciences remain essential to the drug discovery and development process, advances in the molecular, cellular and biochemical sciences have enabled the characterisation of the molecular basis for disease in a fashion not possible by prior art. The advent of genomic sciences in the mid- to late-1990s, and the recent completion of the human genome, is shifting the drug discovery paradigm. This 'genomic'-based approach enables researchers to rapidly correlate changes in gene expression or structure with disease, observe differences in gene structure between multiple species or organisms, and relate genetic variation to disease susceptibility and drug efficacy. These capabilities further enhance the traditional approaches to drug development by providing a mechanism to examine large ('global') segments of the genome in a high-throughput manner.

While these capabilities of genomics are impressive, they are not sufficient to fully comprehend the disease process or to identify protein targets for which to develop drugs against. This is because the gene expression levels may not directly correlate with protein expression levels, the physiologically active protein may be in a complex with other proteins, or the activity may be regulated by post translational modifications or processing events. Whereas the genome is a fairly consistent cellular element, the proteome (the expressed genome) is constantly changing. Protein function can be regulated by synthesis, degradation, trafficking, modifications, or the association with accessory molecules, to mention a few examples. Hence, a complete understanding of a disease process requires a comprehensive analysis of changes in all proteins in the proteome.

Proteomics is the systematic study of the proteome, a field made possible by extensive genomic sequence data generated by the Human Genome Initiative. This data provides the blue-print for identifying proteins when only partial amino acid sequences are available. Consequently, investigators can identify individual proteins in complex mixtures without prior purification or complete sequence elucidation. In contrast to previous biochemical research that tended to focus on individual proteins, proteomics attempts to characterise the changes in protein expression levels or structure throughout the entire proteome. In effect, proteomics attempts to view the molecular basis (the individual proteins involved) for cellular functions in the absence of functional assays. Such capabilities could significantly reduce the time required for drug development by directly, and with high-throughput, characterising the molecules that most drugs are developed against: proteins. This review will address the technologies related to proteomic research and the applications of such research.

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.

The first step in most proteomic analyses involves separating a mixture of proteins by two-dimensional gel electrophoresis, and quantifying the individuals by staining and imaging technologies. Two-dimensional gel electrophoresis technology was developed more than 25 years ago and improved in

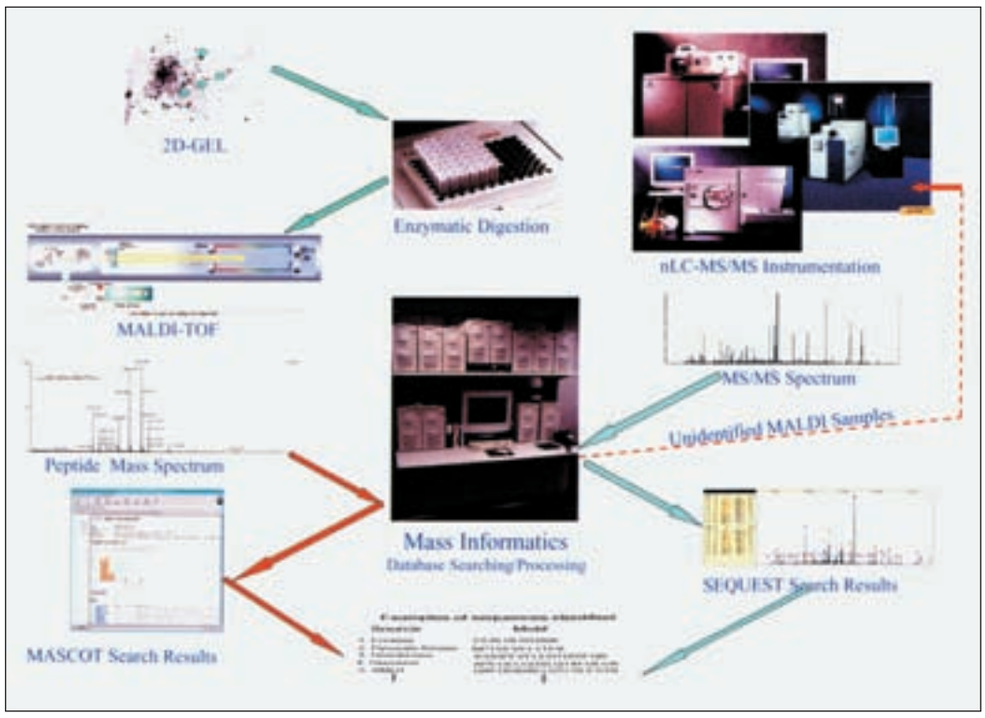

Figure 1
The 'layered approach' to protein identification

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 immobilised 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 immobilised ampholytes. Currently, there are several companies offering the hardware and precast materials for running 2D gels.

Visualisation of the protein 'spots' requires staining of the proteins by one of a variety of methods. Classically, coomassie and silver staining was used. However, several fluorescent dyes have recently been introduced, which provide good quantification and linear dynamic range. Silver staining remains the most sensitive non-radioactive method capable of staining proteins in the low nanogram (<10ng) level; however, quantitation by silver is problematic. The fluorescent stains are nearly as sensitive (~10 ng), but offer a greater range of linear response than silver (three orders of magnitude versus 2-3 fold). Coomassie offers the least sensitivity of these stains, requiring 100ng of material to visualise.

Analysis of the staining patterns, and quantification of the relative amounts present is the next step in this study. Several specialised 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 which specialise 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 analyse (image and quantify) 2D gels in a high-throughput mode. Additionally, several other biotech start-ups that specialise in proteomics are setting up proprietary platforms to analyse 2D gels. However, such a high-throughput 2D-gel analysis package is not commercially available at present.

Because of the high costs of labour and reagents required for producing highly reproducible 2D gels, a number of laboratories are attempting to substitute gels with a multi-dimensional chromatography

system. These systems are varied in design, but promise a direct interface between the sample separation steps (chromatography) and analysis steps (mass spectrometry). Because of this direct interface, it is believed that such systems will limit sample loss, leading to increased sensitivity, and provide for higher-throughput. Although these technologies show promise, the development work is still largely being pursued in academia, and has not been fully validated for routine applications.

Following 2D gel electrophoresis and image analysis, the selected proteins are 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 ionisation techniques: matrix-assisted laser desorption ionisation (MALDI) and electrospray ionisation (ESI). These methods made it possible to ionise and thereby detect large biological molecules using a mass spectrometer. Furthermore, ESI technique made it possible to interface chromatographic systems to mass spectrometers. The mid- to late-1990s saw improvements in sensitivity, automation and throughput of these mass spectrometry-based protein identification techniques. These improvements are credited to development of nanoscale chromatography devices and of software for rapidly searching experimental mass spectrometry data against protein and genomic databases. 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 analysing proteins: MALDI-MS for peptide mass fingerprinting, and liquid chromatography-electrospray ionisation tandem mass spectrometry (LC-ESI-MS/MS) for peptide sequence analysis. In most 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.

One protocol involves using a 'layered approach' to protein identification (See **Figure 1**). After protein spots are excised from a 2D gel they are enzymatically fragmented with proteases. The resulting protein fragments (peptides) are extracted from the gel matrix and analysed by MALDI-MS peptide mass fingerprinting. 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 two to three hours. 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 ~80% 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 yield a protein identification, or that yield an ambiguous protein identification, are then further analysed by LC-ESI-MS/MS (the layer approach). 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 post-translational modification, including phosphorylation, methylation and acetylation. The technique, however, is considerably slower than the MALDI-MS peptide mass fingerprinting, and as such is not as 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 co-ordinated efforts to develop software for performing this function. This software, which will be available in 4Q00, will support 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 Bio-Rad or Micromass. Other commercial software packages are also being developed by a variety of companies.

An important and essential component of this proteo-informatics platform is mass informatics. Mass informatics is a compilation of various software packages including database search programs to analyse 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 categorised into peptide mass fingerprinting

programs which use MALDI-MS data, and peptide sequencing programs which use LC-ESI-MS/MS data. Two software packages namely SEQUEST and MASCOT have the capability to handle vast amounts of data from a high-throughput proteomics effort and are both commercially available. SEQUEST, marketed by Finnigan Corporation, is a specialised 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. However, the rate of database searching using MASCOT is limited because the current version of software does not permit batch processing. SEQUEST, on the other hand, is scaleable and provides high-throughput search capabilities using a version, PVM-SEQUEST (parallel virtual machine-SEQUEST), which can run on a LINUX Beowulf Computer Cluster. This arrangement allows one to process an enormous volume of mass spectral files. On a typical day the four tandem mass spectrometers in our proteomics laboratory can acquire and generate 20,000 MS/MS data files, for example. Using PVM-SEQUEST each MS/MS data file can be searched against all known protein or EST sequences in 1-1.5 seconds. This processing speed, when interfaced in real time with LC-MS/MS data acquisition, will provide the ability to sequence peptides every 2-5 seconds and thus enable the identification of proteins every few minutes.

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 such as OGS and LSP have invested huge resources to set up proprietary data management and archival schema. The first commercial package that shows promise will be WorksBase from Bio-Rad. WorksBase will use Oracle relational database architecture to efficiently perform both data management and data archival tasks.

Impact of proteomics in drug discovery

The genome provides the sequence of the gene and, by extension, the sequence of the protein; proteomics provides information as to the function of the protein. Genomics and proteomics are thus complementary technologies. In a commercial setting these technologies benefit from being highly integrated with the biological, disease-related research being pursued in therapeutic areas. This integration allows for a concerted development of experimental strategies, a seamless flow of experimental data, and the more orderly transition of programs from one organisational structure to the next.

The application of proteomics crosses several areas of drug discovery from exploratory research to full drug development. Work in the exploratory phase research is directed at identifying 'drugable' protein targets and pathways. This work is largely directed at identifying proteins involved in signal transduction events in cells (kinases and phosphates) or at identifying binding partners in protein complexes. A proteomics approach to these research areas is uniquely suited because proteomics technologies are capable of relating seemingly disparate members of a genomics database and monitoring structural alterations. For example, the identification of signalling proteins usually involves identification of phosphorylated proteins (post translational modifications) after cellular stimulation. Conversely, protein-protein interactions can not be fully approximated with genomic data, and thus require the direct analysis of the proteins comprising the complex.

Traditional drug development strategies involve the systematic study of disease leading to the isolation of disease-related proteins that can be used for screening chemical libraries to identify compounds that modulate the activity of the target protein. In many instances, however, compounds are selected that modulate cellular functions without prior knowledge of the drug target. Unfortunately, lack of knowledge about the drug target may limit the opportunity to build a blockbuster franchise through 'life cycle management' activities such as second-generation improvements to the drug chemotype, and other proprietary advancements. To address this, proteomics studies can be directed at identifying the mechanism of action of drugs, and at differentiating the effects of various drug chemotypes on global changes to the transcribed proteome. Such applications rely on the availability of synthetic chemistry support to produce appropriate compounds for the studies. Radioactively labelled compounds with reactive moieties for binding covalently to target proteins are generally required. The key to these types of studies is having compounds with sufficient binding affinities and specific activities to label targeted proteins in trace amounts. Failure to obtain such compounds generally results in the labelling of non-specific, higher abundant proteins. As such, the application of proteomics technologies to mechanism of action studies requires considerable planning and may not be appropriate if the proper reagents are unavailable.

Another area of opportunity for using proteomics involves the identification of surrogate biomarkers to monitor disease progression or drug treatment. Biomarkers are used in drug discovery for multiple purposes. A biomarker profile can enable decision making during preclinical development, or provide

an efficacy or safety measure in a clinical setting. Biomarkers are typically developed as part of the early phase research; however, for several diseases, such as Alzheimer's or atherosclerosis, the development of such markers is not as straight forward as it would be for conditions such as hypertension, where a simple non-invasive check of blood pressure provides the output required. For these diseases, proteomics efforts are directed at identifying protein markers in bodily fluids that would be indicative of disease progression/regression or drug efficacy. As described for target identification, preclinical biomarker discovery is facilitated by a close interaction between the biologist studying the disease and the technologist in the proteomics group. Reagent sets, such as a knock-out mouse line or appropriate cell line, can significantly enhance and enable the proteomics approach. It is important to realise that proteomics is a mechanism for identifying and characterising proteins. Complementary approaches, such as bioinformatics, transcriptional profiling, and biochemical analysis are required to transition preclinical biomarkers into the clinical setting.

Conclusions and future prospects

The genomic sequencing efforts spawned the field of proteomics by providing the basic blueprint for protein sequences. What before required extensive efforts to purify and sequence proteins is now accomplished quickly and with relative ease. Future advancements directed at automating, multiplexing, and interfacing the technologies will advance the capability of performing proteomic analyses in a much more inclusive and encompassing manner than is currently possible. Proteomics is the contemporary technology for studying the underlying mechanisms of disease. DDW

Ashok R. Dongre PhD *is a Research Investigator (Proteomics) in the Department of Applied Genomics at Bristol-Myers Squibb PRI in Princeton NJ 08543. E-mail: dongrea@bms.com*

Stanley A. Hefta PhD *is an Associate Director of Proteomics in the Department of Applied Genomics at Bristol-Myers Squibb Princeton NJ 08543. E-mail: heftas@bms.com*