PROTEOMICS IN DRUG TARGET DISCOVERY –

High-throughput meets high-efficiency

Biochemistry is enjoying a renaissance under the guise of 'proteomics' due to the availability of sequenced genomes, advances in mass spectrometry and associated data analysis. Its application to drug discovery and development has obvious benefits to the multiple aspects of drug development but this is only beginning to be realised. This article explores the emerging technologies of non-gel based methods for protein quantitation and identification together with the trend towards focused studies rather than global approaches to protein expression level analyses.

By Dr Terence E. Ryan and Dr Scott D. Patterson

The completion of assemblies of the human genome, as well as those of the mouse and many pathogens, has allowed the science of drug discovery and development to move into the proteomic era[1]. In the past, potential drug targets were discovered by non-systematic methods: proteins were associated with a specific disease process either through phenotype dissection, by analogy with similar diseases, examining the differential expression of easily identified proteins in diseased versus normal tissue, or more frequently, by serendipity. With the new-found ability to predict all the possible protein coding regions in experimental systems, it is now possible to expand analyses beyond the most-abundant and best-characterised proteins of the cell, and to discover novel drug targets. When harnessed to recent advances in the industrialisation and automation of genetic and biochemical laboratories, drug target discovery enters a new realm where large numbers of samples from human tissues or experimental animal models can be subjected to in-depth examination. These advances include high-throughput DNA sequencing, transcript level analysis (using one of a number of techniques), or protein-based analysis (the subject of this article). In this fashion, previously undiscovered targets for small molecule therapeutics, antibody or other protein-based interventions, and even cell-based therapies may be identified.

What flavour of proteomics?

Like its antecedent genomics, the neologism 'proteomics' has become variable in its definition. The 'proteome' was originally proposed as a term to describe the protein complement of the genome. As the expressed portion of the proteome differs from cell type to cell type, in development and in response to environmental cues, the term 'proteome' has always been more fluid than that of 'genome'. Recently, proteomics (like genomics) has become a catchall, describing a variety of studies designed to systematically analyse proteins found in cells. For the purposes of this article, we restrict ourselves to describing recent advances in the technologies used to identify and compare the expression of proteins in diseased versus normal cells (target discovery), and exclude discussion of other 'proteomic' analyses, such as pathway identification and protein structure determination. Advances in protein chip technology have recently been summarised and are beyond the scope of this article[2]. Differential protein expression efforts are important components of the drug target discovery efforts

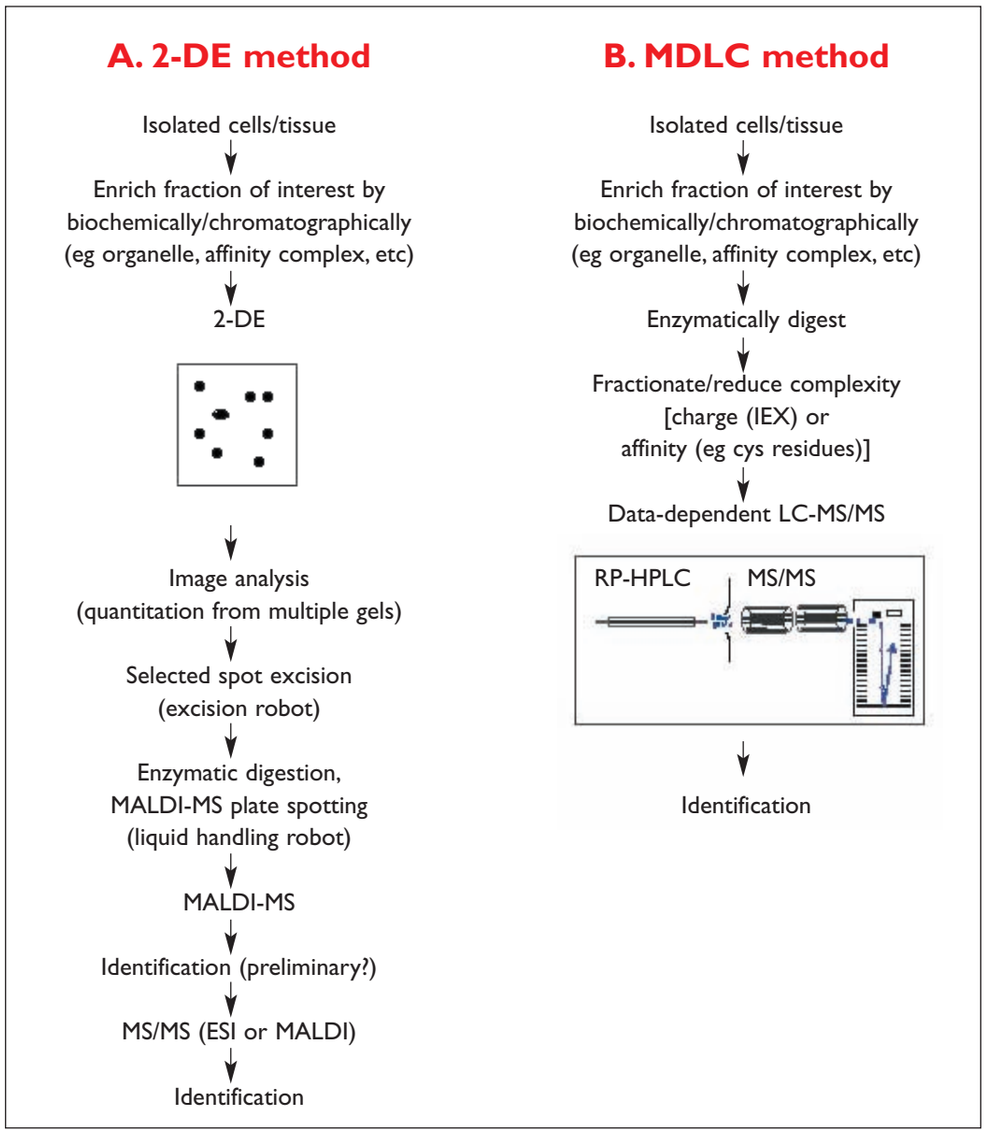

Figure 1
A comparison of the steps involved in 2-D gel and multi-dimensional chromatography-based proteomics approaches. **A** 2-D gel electrophoresis utilises a high-resolution separation at the protein level prior to image analysis (with multiple samples comparative images allow relative abundance measurements). Proteins of interest are enzymatically digested to yield peptides for identification. **B** Multi-dimensional chromatography can utilise specific/limited fractionation prior to enzymatic digestion to yield peptides. All quantitation and identification is conducted at the level of peptides

of Amgen Inc (Thousand Oaks, CA), Beyond Genomics (Waltham, MA), Celera Genomics (Rockville, MD), Bristol-Myers Squibb (Princeton, NJ), Genentech (South San Francisco, CA), Geneva Proteomics (Geneva, Switzerland), Large Scale Biology (Gaithersburg, MD), MDS Proteomics (Toronto, Canada), Pfizer (Groton, CT), Hoffman-La Roche (Basel, Switzerland) and several other companies. Of those companies with industrial proteomic laboratories, a variety of technologies from 2-D gel electrophoresis to new emerging analytical methods are utilised. These new methodologies include techniques which provide high-resolution separation of proteins or peptides, peptide quantitation and identification via a new generation of mass spectrometers, protein 'tagging' chemistries to facilitate comparisons between samples, and new methods for isolating cells and subcellular fractions for analysis. We will describe the evolution of protein-based methodologies and technologies and how the strategies are being employed to integrate them into a cohesive target identification strategy.

An amicable separation: two-dimensional or multi-dimensional?

The first systematic attempts to perform 'global' protein expression analysis made use of technologies that are surprisingly 'old' relative to those used

in genomics – protein electrophoresis and mass spectrometry (**Figure 1A**). Polyacrylamide gel electrophoresis has been widely utilised to provide high-resolution separation of total cellular proteins, particularly in its two-dimensional (2-D) incarnation. In this method, pioneered independently by O'Farrell[3], Scheele[4] and Klose[5], cellular proteins are first separated in an acrylamide gel matrix along a pH gradient, allowing each protein to migrate in the gel to its isoelectric point. This initial resolution is followed by a second separation (perpendicular to the first) in a denaturing polyacrylamide gel according to molecular weight, resulting in a two-dimensional display of proteins in a gel matrix that can be visualised and crudely quantified by any of several protein stains, or by radioisotope detection. Following separation and quantitation, the proteins of interest must be isolated from the acrylamide gel matrix before they can be subjected to further analysis. This isolation can be accomplished by punching out 'spots' from the gel and eluting the contained protein, or the separated proteins can be transferred via an electric current to a supporting membrane, where the original 2-D separation is preserved in solid-state form. Isolated proteins from the gel or membrane are then proteolysed with trypsin to yield peptide fragments ranging from ~10 to 20 amino acid residues in size. The mass of the resulting peptides can be accurately determined by mass spectrometry (MS), yielding a mass-to-charge (m/z) ratio for each peptide (for further description see below). A fingerprint of peptide m/z ratios can be characteristic of a protein, allowing its identification by comparison with calculated m/z ratios derived from each potential protein sequence in the database being searched (a method known as peptide mass fingerprinting). However, gas phase fragmentation of peptide ions in the mass spectrometer, known as MS/MS (or MS^2), allows conclusive identification of a specific peptide through matching of the experimentally derived fragment ion masses with those calculated for all peptides of the same size in the database (a method known as uninterpreted MS/MS matching)[6]. This is an especially powerful approach, as it allows matching of peptides to small regions of translated genomic sequence without the need to assemble full-length transcripts, and is of particular value for identification of proteins from complex genomes, where many peptides could be present that differ in mass by amounts too small to be reliably distinguished by conventional mass spectrometry.

While 2-D gels can provide high-resolution separation of intact proteins, significant limitations apply. First, proteins at extremes of isoelectric point and molecular weight are usually poorly resolved or not observed. Second, membrane proteins tend to be under-represented in this type of separation, risking the exclusion from analysis of an entire class of proteins thought to include targets of high therapeutic importance. Third, high-resolution polyacrylamide gel matrices are relatively thin, severely limiting the amount of sample that can be applied to any one gel, which has a net effect of eliminating low-abundance proteins from analysis unless protein 'spots' from many successive gel analyses can be pooled. Finally, 2-D gels are difficult to prepare and run in a reproducible fashion, limiting their adaptability to high-throughput, factory-type laboratories. However, 2-D gels as a separation technology have advantages to recommend their use, such as a long body of experience with this methodology where problems have been well identified and understood, as well as the ease of visualisation of post-translational modifications, where slight changes in pI or molecular weight can be associated among related proteins. In addition, information on the relative mass of the intact protein is retained. Indeed, the chief advantage of 2-D gel separation is that single spots are thought to represent single proteins (although this is not always true), simplifying the computational requirements for mass spectrometry data. This approach of coupling 2-D electrophoresis for protein separation and quantitation to mass spectrometry for identification is a hallmark of the approaches popularised by Large Scale Biology (Gaithersburg, MD), Oxford GlycoSciences (Abingdon, UK), Pfizer (Groton, CT) and Hoffman-La Roche (Basel, Switzerland).

Recently, a new approach known as 'complex mixture analysis' (**Figure 1B**) has been introduced, which takes advantage of multi-dimensional chromatography, the new generation mass spectrometers, high-speed computing resources and genomic assemblies[7-10]. Instead of separating individual proteins using 2-D polyacrylamide gels, complex mixture analysis can start with protein pools partially fractionated by multi-dimensional liquid chromatography, a technique that accomplishes serial protein separations over a variety of chromatographic matrices[11]. The protein pools subjected to this type of separation can represent any easily obtained protein mixture, such as affinity-purified proteins, subcellular organelles, hydrophobically partitioned proteins, proteins of a particular size range, or even total cellular proteins from simple organisms such as bacteria or yeast[11]. The complex mixtures are subjected to

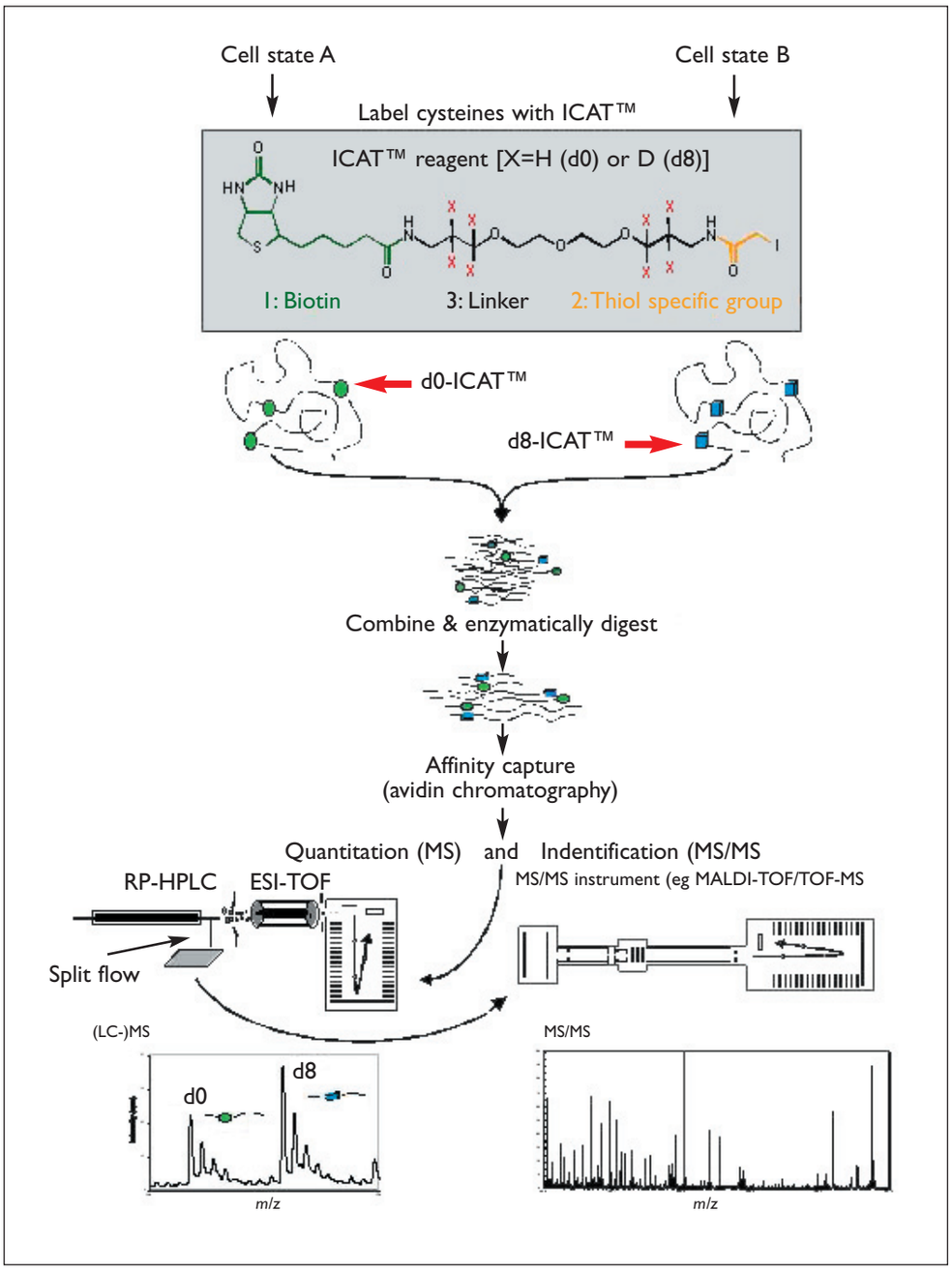

Figure 2
The Isotope Coded Affinity Tag (ICAT™) method. Cysteine residues on proteins are labelled with either the isotopically light (d0) or isotopically heavy (d8) ICAT™ reagent at which time the samples are combined prior to fractionation and proteolysis. The peptides appear as pairs in the LC-MS profile and can be identified by MS/MS either in a separate MS analysis (such as the MALDI-TOF-TOF-MS, shown) following LC-MS separation (ESI-TOF shown) and fraction collection, or through the use of a mass spectrometer capable of both MS analysis and ion selection for MS/MS under instrument control (NB: not all peptides of interest will be fragmented in a single run with this approach and quantitation will be compromised to some degree)

proteolysis with trypsin, the resulting peptides separated via liquid chromatography, and the eluate of the chromatograph fed directly to a mass spectrometer. The mass spectrometer measures the mass (MS mode) and relative abundance of the peptides, and their selection for MS/MS analysis allows their identification. This method of analysis is designed to identify as many components in a sample as possible, and is commonly known as 'profiling'.[10,12]

The utility of the complex mixture analysis approach can be extended to compare protein quantities found in normal and diseased tissues utilising the recently developed technique known as Isotope-Coded Affinity Tag (ICAT™)[13]. In this method, proteins from two different cell sources

are modified on cysteine residues by one of two forms of the ICAT™ reagent. As shown in **Figure 2**, the ICAT™ reagent consists of three parts (1) an affinity tag (biotin), (2) a cysteine reactive group (thiol specific group), and (3) a linker region. The two forms of the ICAT™ reagent are identical except for the linker region, in which there are either eight hydrogen atoms or eight deuterium atoms. As a consequence, cysteine-containing peptides modified with the heavier form of the ICAT™ reagent will display an apparent mass difference of eight atomic mass units relative to the same peptide labelled with the lighter form. This difference is easily distinguished in the mass spectrometer, and direct comparison of the ion signal strengths of the two labelled forms of the peptide correlate directly to the ratio of the expression level of the particular peptide in question (a surrogate for the level of the protein from which it was derived). The presence of a biotin group on the ICAT™ reagent allows only the modified peptides to be enriched through avidin chromatography and subjected to analysis, providing a level of complexity reduction that increases the efficiency and sensitivity of this analytical method. Therefore, ICAT™ provides the ability to quantitate relative abundance levels of peptides/proteins in pairs of samples derived from multi-dimensional chromatography experiments[14].

In addition to the ability to use the ICAT™ technique to compare protein expression levels, several other advantages inherent to complex mixture analysis suggest that this technique will rival 2-D gel-based proteomic analyses in both industry and academia. First, rather than separating at the level of proteins, the high-resolution separation step in complex mixture analysis is conducted at the level of peptides, allowing quantitation and identification to be performed on the same molecular species (the peptide). The proteolysis of proteins into peptides at an early stage eliminates some of the biases inherent in gel-based systems, as tryptic peptides have a much narrower distribution of pI, size, solubility and hydrophobicity than intact proteins, meaning that a more complete representation of proteins can be obtained if peptides are used for high-resolution separation (albeit with loss of information on the relative mass of the intact protein as determined by gel electrophoresis). Secondary fractionation procedures can also be conducted on peptides to reduce the complexity of a sample in a targeted way, further improving the sensitivity of the method (eg, through enrichment of peptides carrying specific post-translational modifications or reactive moieties on amino acid side chains[15]). In addition, liquid chromatography columns can be easily scaled to separate large quantities of a sample in a single run, a feature that is difficult to accomplish with 2-D gels, and which has limited the ability of gel systems to sample low-abundance proteins. This more complete representation comes at a cost, however: such a 'shotgun' approach puts a premium on computing algorithms and hardware needed to identify individual peptides.

Mass spectrometry – the next generation

Proteomics has been enabled by the advent of the genome sequencing efforts and the development of mass spectrometers capable of ionisation of peptides and proteins, as well as by computer algorithms able to use MS data to identify the gene coding for the peptide, and in some cases to quantitate its expression level. As a background to the technology, the simplest description of a mass spectrometer is an instrument that measures charged species under vacuum, and that consists of an ionisation source and a mass analyser/detector. Ionisation of peptide/proteins is accomplished using either a matrix-assisted laser desorption ionisation (MALDI) or an electrospray ionisation (ESI) source. MALDI ionisation is triggered by a laser fired at the sample co-crystallised with a low molecular weight organic matrix, and ESI ionisation by spraying a solution through a charged needle (eg, the effluent of a HPLC column). The ions generated are then measured in a mass analyser, of which the ion-trap (IT) and time-of-flight (TOF) are the most commonly used for either ionisation source today, the TOF analyser providing the best mass accuracy and resolution. Advances in both mass accuracy and instrument control have been essential for the integration of these instruments into proteomics. Data-dependent instrument control allows the mass spectrometers to run a number of complex operations without operator intervention, the most common of which is the ability of the instrument to select observed peptide ions for MS/MS analysis (ie, automatic generation of spectra for identification). Therefore, the data acquisition rate is determined by the speed (duty cycle) of the instrument. In this regard the TOF-based mass spectrometers provide the fastest data acquisition rates, and a new class of MALDI-based instrument (a true TOF-TOF-MS) referred to as the Voyager 4700 Proteomics Analyser (Applied Biosystems, Inc, Framingham, MA) promises to be the fastest yet with a 200Hz laser (an order of magnitude faster than other commercial instruments), allowing more spectra to be collected per unit time[16]. MALDI

References

1 Subramanian, G et al (2001). Implications of the human genome for understanding human biology and medicine. JAMA 286 (18), 2296-2307.
2 Kiplinger, JP (2001). Protein Arrays: New technologies for the proteomics era. Drug Discovery World 2 (3), 40-46.
3 O'Farrell, PH (1975). High resolution two-dimensional gel electrophoresis of proteins. J. Biol. Chem. 250, 4007-4021.
4 Scheele, GA (1975). Two-dimensional gel analysis of soluble proteins. Characterization of guinea pig exocrine pancreatic proteins. J. Biol. Chem. 250, 5375-5385.
5 Klose, J (1975). Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues: A novel approach to testing for induced point mutations in mammals. Humangenetik 26, 231-243.
6 Patterson, SD (2000). Proteomics: the industrialization of protein chemistry. Curr Opin Biotech 11, 413-418.
7 McCormack, AL et al (1997). Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. Anal. Chem. 69 (4), 767-776.
8 Link, AJ et al (1999). Direct analysis of protein complexes using mass spectrometry. Nat. Biotechnol. 17 (7), 676-682.
9 Mintz, PJ et al (1999). Purification and biochemical characterization of interchromatin granule clusters. EMBO J. 18 (15), 4308-4320.
10 Patterson, SD et al (2000). Mass spectrometric identification of proteins released from mitochondria undergoing permeability transition. Cell Death Diff. 7 (2), 137-144.
11 Washburn, MP et al (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19 (3), 242-247.

Continued from page 52

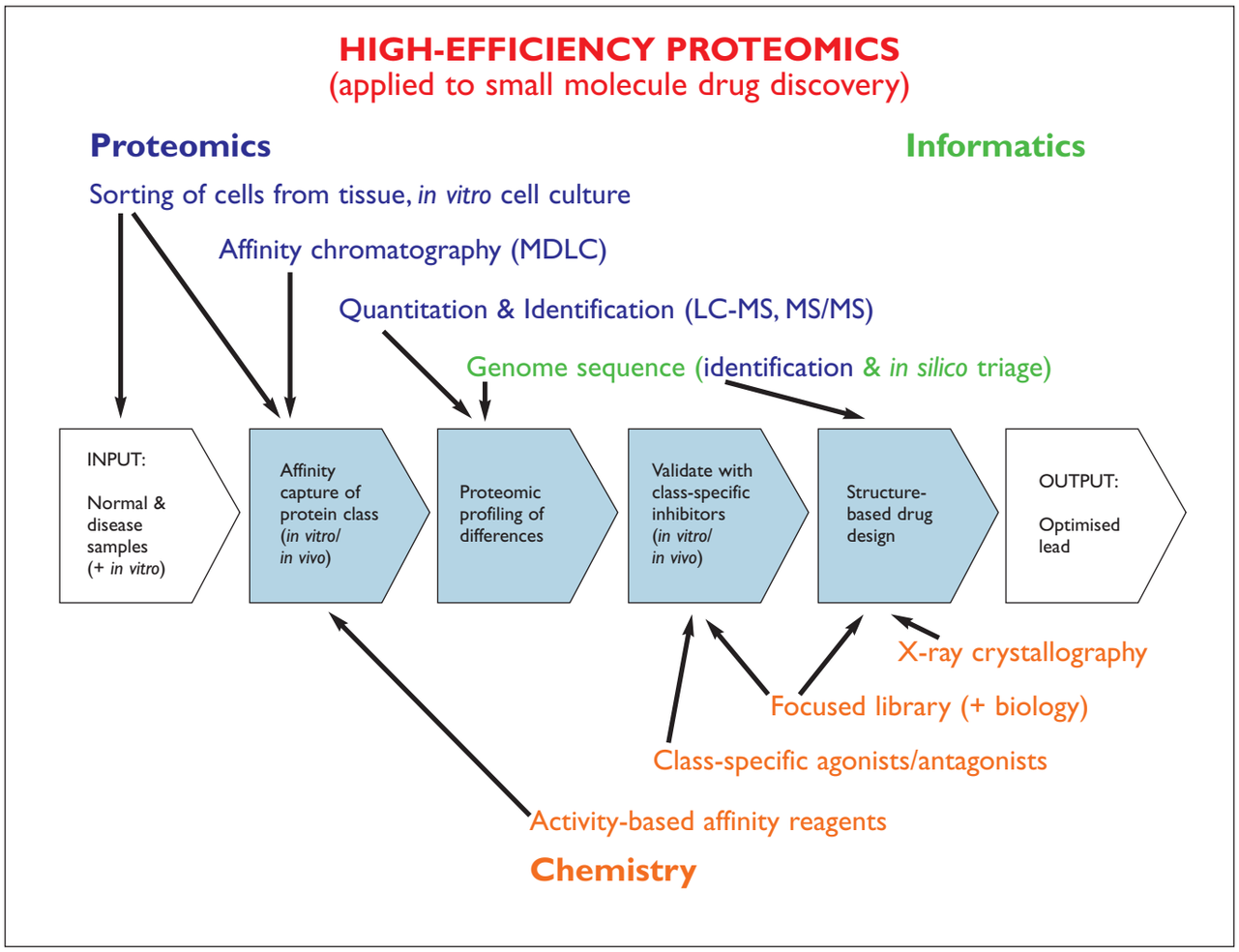

Figure 3

High-efficiency proteomics. Through the combination of targeted capture and high-throughput chromatography-based proteomics, discovery can be linked to screening and development capabilities with the aim of reducing the attrition rate in drug discovery and development. The schematic represents discovery through to an optimised lead compound. The capabilities required for such an approach are listed and grouped under the general areas of Proteomics (incorporating Cell Biology, Protein Chemistry and Mass Spectrometry), Informatics (incorporating associated software for quantitative analysis, identification, structural databases, etc) and Chemistry (incorporating medicinal and combinatorial chemistry, X-ray crystallography and *in vitro*/*in vivo* biology and pharmacokinetics)

TOF-TOF instruments are being commercialised by Applied Biosystems, Inc and Bruker (Bremen, Germany), and are beginning to be operated by Celera Genomics (Rockville, MD), Oxford GlycoSciences (Abingdon, UK), and Geneva Proteomics (Geneva, Switzerland).

High-efficiency vs high-throughput

While the modern proteomics laboratory can be engineered and scaled to provide a high-throughput of biological specimens, the number of 'druggable' proteins encoded by the human genome probably represents a tiny fraction of its potential coding capacity. Proteomic analyses of the sum of human proteins, either on a whole-body basis or that of an individual cell, are likely to be inefficient approaches for identifying valid drug targets. In preference to this 'atlas' approach, we propose that the maximum efficiency of the proteomic laboratory can be reached by preselecting the classes of proteins that are considered 'druggable' by a pharmaceutical development company. For example, a

Continued from page 49

12 Spahr, CS et al (2001). Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry I. Profiling an unfractionated tryptic digest. Proteomics 1 (1), 93-107.
13 Gygi, SP et al (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. 17 (10), 994-999.
14 Han, DK et al (2001). Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotechnol 19 (10), 946-951.
15 Spahr, CS et al (2000). Simplification of complex peptide mixtures for proteomic analysis: Reversible biotinylation of cysteinyl peptides. Electrophoresis 21, 1635-1650.
16 Medzihradszky, KF et al (2000). The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. Anal Chem 72 (3), 552-558.

company skilled in development of monoclonal antibody therapeutics would most efficiently utilise proteomic profiling of potential targets found on the plasma membrane, rather than a profile that included diverse and abundant proteins from other cellular compartments. Other companies, which have expertise in chemical inhibitors of a particular enzyme class, could use immobilised inhibitors to provide an affinity enrichment of proteins of that class. This latter approach also lends itself to rapid target validation, as chemophores of a similar specificity to those used for capture can be utilised for *in vitro* or *in vivo* experiments to evaluate whether they induce the desired phenotypic response. If so, such chemical 'knock outs' provide a starting point for structure-based drug design or directed small molecule library screening.

Targeted complexity reduction removes proteins which are irrelevant to drug discovery from analysis, freeing bandwidth for focused analyses in greater depth or over a broader range of specimens. Thus, application of this approach at a large scale (high-throughput) results in what we refer to as 'high-efficiency' (targeted analyses at scale). This 'high efficiency' approach, when coupled with the in-depth analyses possible with complex mixture analysis and modern instrumentation, maximises the opportunity to discover drug targets that can withstand a rigorous process of target validation, resulting in fewer false leads to divert development efforts (**Figure 3**).

The future of proteomics

At this early stage in the development of proteomic expression profiling and quantitation, it would be foolhardy to predict the arc of methods and instrumentation development. However, broad trends are clear, and point to areas certain to be active over the next several years. First, a new generation of mass spectrometry instrumentation will considerably speed the acquisition and processing of peptide mass measurements. In addition, new techniques for data-dependent MS-MS analysis will make new instruments increasingly useful in factory-scale proteomic laboratories, and dramatically lower the cost of peptide sequence identification. Second, multi-dimensional liquid chromatography is likely to replace 2-D gel technology as the high-resolution separation method of choice. Indeed, proponents of 2-D gel separations such as Geneva Proteomics (Geneva, Switzerland), Oxford GlycoSciences (Abingdon, UK) and Bristol-Myers Squibb (Princeton, NJ) have recently begun to describe their use of liquid chromatography in presentations at recent press conferences and scientific meetings. Third, new protein identification algorithms will be developed to facilitate use of complete genomic sequence databases, as will supercomputing methods to deal with the large data collections resulting from massive application of parallel proteomic analyses in factory-type settings. Finally, high-efficiency approaches will increasingly focus on reducing cellular protein complexity to the various classes of druggable target proteins, streamlining target identification by dramatically improving the statistical confidence in proteomic experiments that identify differentially-expressed proteins. DDW

Dr Scott D. Patterson *has been at Celera Genomics for the past year and holds the position of Vice-President, Proteomics. Prior to that he was at Amgen, Inc in California where he established its proteomics programme as well as conducting research in apoptosis, work that he had begun when he was on the faculty of Cold Spring Harbor Laboratory in New York. He received his doctorate from the University of Queensland, Australia where he also worked.*

Dr Terence Ryan *is Director of Cell Biology in Celera's Proteomics Department. Prior to that he was Director of Protein Sciences and Cell Technology at Regeneron Pharmaceuticals and directed new virus discovery at Pandex Laboratories, part of Baxter International. Dr Ryan holds a doctorate from Rutgers University and was a fellow in Molecular Virology at the University of Wisconsin.*