Data analysis—the Achilles heel of proteomics

Scott D. Patterson

During the past few years there has been a resurgence of research using parallel protein-based analysis, now commonly referred to as proteomics. However, our ability to generate data now outstrips our ability to analyze it. This occurs even though proteomics is inherently a substrate-limited science and proteins exist over a wide concentration range in biological samples. Therefore, it is not surprising that the entire proteome of any species has yet to be observed. In this article, I address some of the primary issues currently facing researchers in this field, with an emphasis on the computational aspects affecting progress, including the accuracy of matches from mass spectrometric data to sequence databases and the integration of the results of proteomics experiments to yield biological meaning.

Parallel protein-based analysis first came to the fore during the mid-1970s with the introduction of two-dimensional gel electrophoresis, which for the first time allowed a staggering number of different protein species to be revealed in a single experiment and permitted the comparison of expression patterns between samples. At that time, the method suffered from irreproducibility, the limited ability to identify the proteins that displayed differential expression, and the considerable computational challenges of large-scale experiments resulting from the state of image-analysis software and the cost of computing and data storage.

During the past decade, the field has garnered renewed interest, with advances in mass spectrometry allowing the analysis of ever-diminishing quantities of proteins, and the refinement of identification software providing the means to match these data with translations predicted from nucleotide sequence data (originating from the genome projects)[1]. Software has improved, and the cost of computing and data storage has continued to decline.

Proteomics data sets have the potential to provide biological information unattainable through parallel transcript—that is gene expression—analysis. There are numerous examples in which transcript amounts at steady state, or even after induction, do not correlate with the amount of protein present in the system because of well-known, but currently mostly unpredictable, effects of post-transcriptional and post-translational regulation[1]. Therefore, analysis at the level of both the transcript and the protein is critical to understanding biological systems. Gene expression analysis can, in principle though rarely in practice, determine the quantity of every transcript in a cell because of the wide dynamic range of some of the detection systems.

Table 1. Proteomics experiments require handling of diverse data sets

Stage of process	Type of data
Preparation for analysis	Project information Sample information Separation, fractionation
Sample processing	Quantitative analysis (LC-MS/MS, 2-DE MS/MS) Identification, MS/MS data analysis
Data analysis	Data capture and validation Data management and integration
Monitoring	All processes require quality assurance and quality control

"...our ability to generate data now outstrips our ability to analyze it"

Large-scale protein analysis is substantially more difficult for several reasons: because proteins are fragile, considerable losses occur at each step of enrichment; proteins can exist in multiple isoforms (post-translational modifications are variable and dynamic); there is no protein equivalent of PCR, preventing specific amplification of proteins in a sample (i.e., proteomics is substrate-limited); the analytical instrumentation available can analyze only ~3–4 orders of magnitude of protein concentrations (cellular proteins cover ~6 orders and serum proteins 10 orders of magnitude[1]); finally, proteomics experiments are inherently more complex and as such have a lower throughput than gene expression experiments.

Even putting these experimental issues aside, proteomics raises particular data analysis problems because of the nature of the two main technologies used in studying proteins: first, systems based on two-dimensional gel electrophoresis (i.e., quantification through image analysis, proteolytic digestion of the isolated proteins of interest, and identification by mass spectrometry) and non-gel based systems (i.e., quantification and identification by mass spectrometry of mixtures of proteolytically digested proteins)[1].

Two-dimensional gel software continues to be improved, but after 20 years of development it still requires some manual intervention. Mass spectrometers measure the mass-to-charge ratio of charged molecules, and the most common approach to protein identification is the proteolytic digestion of proteins into peptides. If the digest is of a purified protein, then those masses may be sufficient to identify the protein. Identification occurs with greater confidence if the structure of individual peptides is revealed through fragmentation, generating tandem mass spectrometry (MS/MS) data. Peptides derived from complex protein mixture digests can be separated chromatographically and fed into a mass spectrometer online. The development of instrument control software that allows mass spectrometers to select and fragment ions automatically produces large data sets for analysis.

Data output from these platforms thus presents two primary problems for proteomics research. First, the vast majority of MS/MS spectra are not of peptides but instrument noise or minor contaminants and therefore have no value, so that their analysis wastes enormous amounts of computing time. The development (and continuing refinement) of algorithms for matching MS/MS spectra to sequence has been an important breakthrough for pro-

Scott D. Patterson is chief scientific officer, Farmal Biomedicines, LLC, Pasadena, CA 91106 (scottp@farmalbiomed.com).

Proteomics caveats

Aside from data analysis, the accuracy of quantitative data is an issue in proteomics, whether it be based on two-dimensional gels or other means, but this needs to be considered in light of the experimental aim, In some instances, knowing that expression varies by 20% could be critical; in others, 200% changes may represent the minimum change considered of interest. In addition, proteomics measurements are essentially all relative, because global measures of absolute abundance are currently not feasible, although they are being discussed[1]. As mentioned in the main text, proteomics is substrate-limited, and, together with the somewhat limited dynamic range of protein concentrations interrogated, only a portion of the proteome can be examined at one time. This is an important caveat that was lost on many investigators over the past few years in the hype of what proteomics could produce. *SP*

teomics. However, these produce scores that indicate *only* the significance of a match; they do not provide an absolute measure. As a consequence, results are often plagued by false positives, with manual review often the only means of gaining confidence in a finding (if identification is critical, the peptide can be synthesized and then fragmented to determine whether the same MS/MS spectrum is produced).

A second problem is that analysis of search results by inexperienced operators can lead to acceptance of false matches and lists of identified proteins that are inaccurate. This is exacerbated by the large number of spectra being sent for search and evaluation. In addition, the lack of commercially available tools for the matching of peptide MS/MS spectra, which can link matched spectra, means that every spectrum, whether earlier identified or not, has to be searched and the results evaluated. Because of these constraints, experiments are often analyzed individually rather than as part of a larger program that allows storage into a relational database. This is probably the biggest problem facing researchers new to this field.

The wide adoption of proteomics approaches into biological research will require several developments to combat data overload and ensure data quality. First, tools must be readily available to deselect MS/MS spectra from search routines that are unlikely to yield a match because of poor quality. Second, search algorithms require further refinement to diminish the false positives and false negatives (merely setting scores high to diminish false positives is counter to the aim of the experiment); this problem is beginning to be addressed through the development of true probability-based scores[2,3] that are akin to the assignment of quality scores to each base in DNA sequencing. Third, spectral matching algorithms for peptide MS/MS spectra need to be made commercially available. And fourth, a database of truly nonredundant transcripts of the organism under study is required, together with an extensive relational database that can acquire data from the diverse range of instruments involved in each stage of the proteomics experiments (see Table 1)[4].

With the caveat that only a portion of the proteome is being analyzed at any one point in time (see "Proteomics Caveats"), the most obvious first step in data analysis is to determine whether differential protein expression affects any known pathways. If so, have all members of that pathway been identified in the analysis and have they been similarly affected? Can new pathways be discerned? How well are the data correlated with other data sets, such as gene expression analysis or even metabolite data? An excellent example of an early integration of such data sets was demonstrated in yeast[5], but application to higher organisms is certainly proving more complex.

Using the genome as the key to linking sets of gene-related information is one means of beginning to tie together biologically relevant information sets[6], but clearly new tools and visualization methods are required. And even when data are generated and analyzed under carefully controlled conditions, distilling biological significance from these enormous data sets will remain a challenge, which will require the help of new types of software (a challenge shared by gene expression analysis).

Advances in proteomics offer much to biological discovery. However, the realization of the full potential of technological advances will require concurrent intensive efforts on the computational front.

1. Patterson, S.D. & Aebersold, R.H. *Nat. Genet.* **33** (suppl.), 311–323 (2003).
2. Keller, A. *et al. Omics* **6**, 207–212 (2002).
3. MacCoss, M.J., Wu, C.C., Yates, J.R. *Anal. Chem.* **74**, 5593–5599 (2002).
4. Ryan, T.E. & Patterson, S.D. *Trends Biotechnol.* **20** (Suppl. Proteomics), S45–S51 (2002).
5. Ideker, T. *et al. Science* **292**, 929–934 (2001).
6. Cherry, J.M. *et al. Nucleic Acids Res.* **26**, 73–79 (1998).