
Lecture 11: Proteomics

Microbiology 343

David Wishart, University of Alberta, Edmonton, AB

Introduction

1. Proteomics Defined

High throughput or "global" technologies are redefining the way biology and chemistry are practiced. No longer are scientists studying a single gene or single enzyme in isolation for years at a time. Rather they are now looking to characterize hundreds of biomolecules at a time, as rapidly and efficiently as possible. In other words, science has gone "global". To understand why this change has come about, consider these facts: 1) high throughput DNA sequencers now make it possible to sequence entire genomes in a few weeks; 2) gene chips and DNA micro-arrays make it possible to study expressional changes of 20,000 or more genes in less than eight hours; 3) yeast two-hybrid methods allow hundreds of protein-protein interactions to be sorted out in just a few short weeks; 4) 2D PAGE systems allow up to 3000 proteins to be monitored in a single six hour run; 5) microplate readers allow up to 200,000 biological screening assays to be performed in less than 24 hours. These remarkable technological advances, combined with our own thirst for knowledge, have helped spawn two of the "hottest" fields in all of science: genomics and proteomics.

Genomics can be defined as the application of high-throughput technologies towards the characterization of all the genes in a given genome. The genome, of course, is the total DNA content in a cell, tissue or organism. Humans have a genome that is composed of 3,000,000 bases and about 32,000 different genes. Proteomics, on the other hand, is the application of high-throughput technologies towards the characterization of all of the proteins in a given proteome. The proteome is the total protein content in a cell, tissue or organism. Because proteins can be differentially processed, cleaved, phosphorylated, amidated or glycosylated, it is thought that the human proteome is actually composed of more than 500,000 different proteins. In other words, the proteome is a larger, and in many respects a more complex system than the genome. Proteomics, like genomics is a field more aligned with exploratory research than with traditional hypothesis driven research. Typically in a proteomics project you have little idea of what to expect until you see the result. In this regard, proteomics has been pejoratively labeled as a glorified "fishing expedition". However, proteomics is also more holistic than traditional "reductionist" approaches to research. This is because the technologies associated with proteomics actually allow a comprehensive understanding of a complete system (say a cell, an organelle or a whole organism).

Proteomics is a relatively new term. It appears to have been coined by P. James in 1997, who wrote a review in *Quarterly Review in Biophysics* (James, 1997). The most common interpretation of proteomics is that is a sub-field of genomics concerned with

characterizing proteins via 2D gel electrophoresis and mass spectrometry. This is a far too narrow a view of the field. For the purposes of this workshop we have divided proteomics into three different areas:

- **Functional proteomics** -- the identification of protein functions, activities or interactions at a global or organism-wide scale;
- **Expressional proteomics** -- the analysis of global or organism-wide changes in protein expression;
- **Structural proteomics** -- the high throughput, or high volume expression and structure determination of proteins by X-ray, NMR or computer-based methods.

Proteomics is a term that has been adopted by many (but not all) members of the biological community. Indeed, there is still some confusion about what constitutes a proteomics experiment and what constitutes a genomics experiment. For instance functional genomics is a term often used to describe the application of molecular biological techniques (yeast-two hybrids, SAGE, deletion analysis) to look at protein functions or protein-protein interactions. Structural genomics is a term used by many investigators in the US to describe the large scale study of protein structures. While

there is probably no "right" or "wrong", it is more consistent to use the term proteomics whenever the experiment involves the manipulation or handling of proteins or gives direct answers about proteins (their structure, function, amino acid sequence, or activity). If you are still confused by this ambiguous terminology, just remember that structural proteomics is essentially the same as structural genomics and functional genomics is almost the same as functional proteomics.

2. Functional Proteomics

As we indicated earlier, functional proteomics is concerned with the identification and classification of the functions, activities and interactions of all the proteins in a given proteome. It could be said that the other two branches of proteomics (expressional and structural) are also aimed at determining global protein function, although through slightly different means. Because the functional analysis of proteins is an inherently difficult task (even for a single isolated protein), functional proteomics makes use of a wide variety of specially developed experimental and computational methods to unambiguously ascertain protein function. For instance, functional proteomics employs a number of high throughput molecular biological techniques, such as yeast two-hybrid analysis, systematic gene disruption and immuno-precipitation to help with the identification of protein complexes and protein-protein interactions. By identifying

specific protein complexes or pairwise protein interactions, it is often possible to infer a protein's function or at least part of its function.

In addition to its use of molecular biology techniques, functional proteomics also employs biophysical methods such as mass spectrometry and micro-sequencing to assist in the rapid identification of proteins. By identifying a protein through its mass fingerprint or by determining its relationship to better studied proteins through comparative sequence analysis it is often possible to ascertain a protein's putative function. Mass spectrometry may also be used to identify protein modifications (phosphorylation, amidation, glycosylation, etc.) which can be used in the classification or identification of protein interaction pathways. High throughput screening using a variety of functional assays (enzyme assays, binding assays, ligand "chips") can also be used to identify possible functions, substrates or activities for novel proteins. This experimental screening approach can also be complemented with computational approaches where sequence motifs and sequence signatures can be used to identify putative functions, substrates or activities.

The end result of these functional assignment efforts is supposed to be a complete partitioning of a given proteome into different functional classes. A subfield of functional proteomics, called comparative proteomics, is often employed to assist with this classification and to assess the results. In comparative proteomics, the sequences, structures or functional classifications between different organisms (or cells) are

compared in a at a proteome-wide level using various computational techniques. This allows one, in a global way, to infer ancestral relationships, to identify key evolutionary events, to identify differences in metabolic functions, to predict 3D structures or to extract additional information about the functions or interactions of unclassified proteins. This kind of information can be archived in organism-specific databases to enhance their utility and functionality or used to create brand new databases. Some examples of the databases where this information is already being archived include the YPD (yeast proteome database), EcoCyc (the E coli database), ACEDB (the C. elegans database), Flybase (the drosophila database) and BIND (the biomolecular interaction database).

If one succeeds in identifying or classifying all (or most) of the proteins in an organism, it is potentially possible to use this information to predict its phenotype or to predict its response to given mutations or environmental perturbations. This is already being done with simple bacteria (*H. influenzae*) where Edwards and Palsson (1999) used mathematical modeling techniques to predict the metabolic and phenotypic consequences of gene disruptions to the *H. influenzae* genome. This work has laid the foundation to a very new and exciting field called "Systems Biology". It may not be too unreasonable to expect that someday soon, mathematical or computational techniques will be used to model the effects of gene disruptions in higher organisms or in more sophisticated applications such as developmental biology

3. Expressional Proteomics

Expressional proteomics is fundamentally concerned the display, measurement and analysis of global changes in protein expression. In expressional proteomics, the primary tool for protein display (i.e. the 2D gel) has close parallels to the gene chips or DNA micro-arrays commonly used in genomics. Both are display tools used help compare and measure global expression levels of known (or unknown) biological molecules. Indeed, at least one company (called CIPHERGEN) actually produces a so-called "protein chip" that looks a lot like a gene chip. Unlike DNA or RNA, proteins have complex three dimensional structures and unpredictable physical properties that precludes the use of straight-forward techniques such as complementary base-pairing to facilitate their extraction, amplification, separation or display. This makes the presentation, measurement and analysis of proteins on a global level something of an unsolved problem. While 2D gels are still the preferred tool in expressional proteomics, other techniques are starting to be developed, including protein chips (chips made of arrayed antibodies specific to certain proteins), ligand chips (chips made of arrayed small molecule ligands or peptides that are specific to certain proteins), tandem capillary electrophoresis (where two or more CE columns with different separation matrices are paired together) and tandem HPLC (where two ore more HPLC columns with different separation matrices are paired together). All four techniques allow a two (or more) dimensional display of hundreds to thousands of different proteins at one time. They

also allow for direct quantitation and (indirect) identification with a level of sensitivity that often exceeds that obtained with 2D gels.

Another technique that facilitates the display of proteins in expressional proteomics is called immuno-precipitation or the "protein pull-down" experiment. In this technique, if an antibody to at least one protein in a complex has been prepared, it is often possible to precipitate or "pull down" the entire protein complex using an antibody bound to a solid substrate. The use of antibodies to isolate sub-sets of proteins (say between 5-50 different proteins) allows one to employ far simpler and far more robust strategies (1D SDS PAGE, RP-HPLC) to separate and display proteins of interest. Other groups are pursuing protein "display" techniques that actually avoid gels and columns altogether. In particular, isotope tagging, a method pioneered by Rudy Aebersold (while he was at the University of Washington), allows one to work with unseparated peptide/protein mixtures and to not only identify but quantify proteins using mass spectrometry alone. This novel approach promises to greatly simplify protein identification and quantitation in expressional proteomics.

Regardless of how one chooses to separate and display proteins, ultimately these proteins must be identified and their levels of expression quantified. This is where bioinformatics comes in. Whether it is software for normalizing, overlaying or landmarking 2D gels, or software for extracting protein ID's from 2D gel data banks, or software for comparing mass spectrometric fingerprints to known sequence data --

bioinformatics applications play an absolutely essential role in protein quantitation and identification. Expressional proteomics also depends on bioinformatics software to extract and assess significant changes in protein expression. These software applications are generally less well developed than those used in gene-chip analysis. Nevertheless, it is only through comparisons of expression levels of previously identified proteins that one can determine the influence that a drug, a toxin, a pathogen, a mutation or a change in environment might have on the proteins in a proteome. In this regard, it can be expected that expressional proteomics will likely be used not only for basic research but for many applications ranging from medical diagnosis to therapeutic drug monitoring to crop and livestock management.

3. Structural Proteomics

Structural proteomics is perhaps the most computationally intensive of all three areas in proteomics. This is because it is a nexus for both experimental structural biologists (who work with computers some of the time) and computational structural biologists (who work with computers all of the time). Structural proteomics is largely being driven by the Protein Structure Initiative (PSI) a multimillion dollar international effort to solve as many protein structures as quickly and efficiently as possible. This 10 year project which has got underway about 4 years ago (Sali, 1998) has several objectives and a variety of underlying motives. At one level PSI (and therefore, structural proteomics) is concerned

with developing computational tools to extract function from structure. At a second level it is concerned with developing computational tools to facilitate X-ray and NMR structure generation. At a third level, it is concerned with developing improved tools to help with protein structure prediction, protein classification and protein modeling. However, the ultimate aim of both structural proteomics and PSI is to "solve the protein folding problem".

The premise behind structural proteomics is that despite the near infinite variety of protein sequences, the number of protein folds is quite limited (perhaps less than 5,000). To date approximately 1000 of those folds have been determined or identified. It has been argued that if the right subset of proteins could be identified and if these proteins could have their structures solved by X-ray or NMR spectroscopy, then all 5,000 template folds could be known and that all remaining protein sequences could have their structures (or approximate structures) determined through comparative modeling (Sali, 1998; Skolnick et al. 1999). In other words, the protein folding problem could be solved by brute force. Based on some reasonably optimistic estimates it has been argued that the subset of proteins (~5,000) needed to solve the protein folding problem could be mostly generated using conventional technologies by 2010 at a cost of \$300 million.

Not to be outdone, a group of computational structural biologists announced late in 1999 that they would be developing the worlds fastest and most sophisticated

computer to specifically tackle the protein folding problem. The targeted completion date for this super computer (called Blue Gene) is 2007. Given that both experimental and computational experimental biologists are in a bit of a race to see who can solve the protein folding problem first, it is little wonder that members of both groups are hedging their bets by trying to more involved with each other. That way if and when a winner is finally declared, everyone (experimentalists and theoreticians alike) can share in the acclaim. This increased level of collaboration (perhaps motivated more out of fear than need) is actually having a very positive effect in the whole field of structural biology and structural proteomics.

In particular, computational approaches (I'll use the term bioinformatics from now on) are being actively developed to assist experimentalists in many areas including the identification of appropriate crystallization conditions, the prediction of which proteins will likely crystallize, the identification of which proteins will be soluble and stable for NMR, and the prediction of which proteins will express well and which will not. In addition to this work, structural bioinformaticians are developing tools (such as threading and template recognition methods) to help predict whether new protein sequences will have similar folds to previously known proteins. This work is helping to focus the efforts of X-ray and NMR spectroscopists towards solving the structures of proteins with truly novel folds. Bioinformaticians are also helping crystallographers and NMR spectroscopists with many of the "book-keeping" aspects of the PSI project. For instance, novel

bioinformatics applications are being developed to coordinate protein target selection, to facilitate fold classification (CATH, SCOP, HSSP) and to improve reporting and reduce redundancy in structure generation efforts.

While structural bioinformaticians are doing much to help experimentalists, experimental structural biologists are also been actively helping structural bioinformaticians. Through the CASP (computer aided structure prediction) competitions, experimentalists have been providing structural data in a blind test fashion to structural bioinformaticians for the better part of 10 years. This allows protein folders or would-be protein folders to test their methods on homology modeling, threading and *ab initio* structure prediction and to be independently evaluated. The competitive nature of this work has lead to a steady improvements in quality of the predictions and the robustness of the prediction algorithms. Indeed, homology modeling is now so accurate and so routine that many X-ray crystallographers now use molecular replacement to help rapidly generate 3D structures of homologous proteins. It is likely that NMR spectroscopists will eventually make use of this technique in the near future.

A continuing challenge to structural proteomics is the need to extract function from structure. In the future it is likely that structural biologists will rely more and more on structural bioinformaticians to assist them with aspects of functional identification and that structural bioinformaticians will use the growing body of structures and sequence

data to develop more sophisticated approaches for predicting or identifying 3D functional motifs and active sites.

4. Proteomics: What's it good for?

The promise of proteomics is quite tantalizing. Because it is a high throughput "holistic" science, proteomics offers the possibility of learning a great deal about complex systems in very short order. This can be seen in the area of functional proteomics with the now classic publication by Uetz et al. (2000) describing a comprehensive analysis of all detectable protein-protein interactions in yeast. Had conventional methods been used it might have taken decades to get the same result. Similarly impressive results have been obtained in the analysis of ribosomal components in yeast (Link et al., 1999) where more than 80 proteins, some of which were novel, were identified as being part of the complex using liquid chromatography and mass spectrometry. Functional proteomics methods have also allowed the rapid identification of the major proteins in the yeast spindle-pole body (Wigge et al., 1998) and in both the human and yeast spliceosome complex (Gottschalk et al., 1998).

In the area of structural proteomics, it is likely the efforts to determine the structure of as many unique proteins as possible, as quickly as possible will actually enable the protein folding problem to be solved by brute force sometime in the next 7 to 10 years (Sali, 1998). Again, if traditional methods of structure determination were to be

used or if homology modeling methods were not available, it would likely take many decades to reach the same goal.

In the area of expressional proteomics, it is likely that medical diagnostic tests will increasingly use proteomics display or analysis methods to detect disease markers or identify prognostic indicators. This is already being done for cancer detection (Jungblut et al., 1998) and it will likely be done in the analysis of bacterial or viral pathogens (Link et al., 1997). Similar kinds of monitoring are being done in agriculture, aquaculture and forestry.

So whether it is in basic molecular biology, structural biology, medicine, pharmaceutical research or agriculture, it is likely that proteomics will play an increasingly important role in Canadian science. This is already evident in the very active proteomics efforts going on in Ontario (\$75 million), in Alberta (\$5 million), in the National Research Council (\$2 million) and in projects to be supported by Genome Canada (\$300 million over 3 years) and PENCE (\$3 million). Because Canada largely missed out on many of the major genomic projects of the 1990's (genome sequencing, technology development, database placement), federal and provincial governments and their funding bodies have decided to focus on providing substantial support to disciplines arising from the "post-genomic" era. Proteomics is, by definition, a key part of that post-genomic era.

References

Edwards JS, Palsson BO. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J Biol Chem.* **274**:17410-17416 (1999).

Gottschalk A, Link AJ, Hays LG, Carmack EB, Yates JR 3rd. A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins. *RNA* **4**:374-393 (1998).

James P. Protein identification in the post-genome era: the rapid rise of proteomics. *Q Rev Biophys.* **30**:279-331 (1997).

Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR 3rd. Direct analysis of protein complexes using mass spectrometry. *Nature Biotech.* **17**:676-682 (1999).

Link AJ, Hays LG, Carmack EB, Yates JR 3rd. Identifying the major components of Haemophilus influenzae type-strain NCTC 8143. *Electrophoresis* **18**:1314-1334 (1997).

Sali, A. 100,000 protein structures for the biologist, *Nature Struct. Biol.* **5**:1029-1032 (1998).

Skolnick J, Fetrow J and Kolinski, A. Structural genomics and its importance for gene function analysis *Nature Biotech.* **18**:283-287 (2000).

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae* *Nature* **403**:623-627 (2000).

Wigge, PA, Jensen ON, Holmes S, Soues S, Mann M, Kilmartin JV. Analysis of the *Saccharomyces spindel* pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry *J. Cell. Biol.* **141**:967-977 (2000).

Yates JR Mass spectrometry: from genomics to proteomics, *Trends Genet.* **16**:5-8 (2000).