

Proteomics

Proteomics is a buzz word, following on genomics, the study of the entire genome of a species, i.e. all the genes. Proteomics then is the study of all the expressed proteins of a species, or a developmental stage, or a tissue, or even an organelle, or assembly of proteins for a specific purpose, like the “spliceosome” assembly which cuts introns out of mRNA and splices the edited RNA back together. It describes a major change in the study of proteins in biological systems. In the era I grew up in, one identified a *function*, an enzymatic reaction or other function in a biological system, and attempted to identify, purify and characterize the protein responsible for the function. In the proteomic era, in contrast, one characterizes, typically by sequence, large numbers of proteins in a system, and then attempts to figure out what they do, often what other proteins they interact with. In principle, this is an admirable use of modern technical capabilities which enable characterization of large numbers of proteins. My concern is that protein function is often determined only by inference, by sequence similarity to another protein in another species, whose function may or may not be well established. This can lead to construction of a house of cards, in which X is likened to Y which was likened to Z, until a whole concept of function has been erected which rests on very dubious foundations. There have been many cases where a gene was cloned and assumed to encode a specific protein, only to be proven wrong later; in the meantime other proteins may have been described as similar to it, and the error may never be entirely corrected. Also, the interesting aspect of function may be exactly what you don't get from sequence similarity. For instance, identification of a protein as probably a serine protease specific for cleavage at lysine residues does not tell you where it acts in the organism's total functional metabolism – is it a blood clotting protease or a sperm maturation protease? The real function may have been recently established, and obscured by sequence similarity to a class of proteins whose function it does not share.

The essence of proteomics is techniques which allow looking at many proteins at once. The original such technique is two-dimensional electrophoresis, 2D electrophoresis for short, which has been around for nearly 30 years. Proteins in a crude extract are first separated by isoelectric focusing in a tube gel. The gel is extruded and laid along the top of an SDS slab gel, probably after soaking in SDS and heating. Proteins are then electrophoresed out of the tube gel and down the SDS gel. Finally the gel is stained, with Coomassie Blue or silver stain. This technique, depending on two independent properties of proteins, can resolve several thousand protein species, but it does not identify them. Originally this technique was used to compare different tissues of an individual, or the same tissue under different conditions, perhaps with and without some specific induction such as heat shock, to identify protein spots which changed in intensity between the tissues or conditions. However, gels with this many spots could not be compared by the human eye alone. Software and hardware had to be developed to record and compare the homologous spots, taking into account that two gels of two samples would never run exactly the same. But with such techniques proteins expressed differently between two tissues or conditions could be identified. Then the spot could be cut out, transferred to a polyvinylidene difluoride membrane, and at least some sequence determined by gas-phase Edman degradation. Sometimes the spot could be degraded proteolytically in the gel before transfer, and the sequence of a number of peptides determined.

One development of this technique has been the use of differential fluorescence labels. One extract is labeled with one fluorescent label, say fluorescing blue; an extract of a different tissue or condition is labeled with a different label, say fluorescing yellow. The two extracts are then mixed and electrophoresed on one gel. If there is the same amount of a protein in both

extracts, the corresponding plot will appear green under UV light, as both labels fluoresce; but a protein differently expressed in the two extracts will appear blue or yellow, depending on which it is most expressed in. This avoids the problem of comparing different gels.

The same concepts are now applied using other means of separation. Typically, proteins are separated by high performance liquid chromatography. They are then either directly analyzed by mass spectrometry, or digested with a sequence-specific protease such as trypsin and the resulting peptides analyzed by mass spectrometry. This is a technique, originally limited to reasonably small molecules, in which charged ions move down a tube and are deflected, I think by a cross electrical field, to land along a detector. Mass can be determined to a fraction of a dalton of molecular weight (or rather molecular mass).

Mass spectrometry of whole proteins typically uses matrix-assisted laser desorption ionization (MALDI). The protein sample is solidified in an acidic matrix – which I think means drying on some target plate in presence of a solid acid, one which absorbs a particular wave length of UV light. When the spot of dried matrix and protein is zapped with UV light, the matrix rapidly volatilizes, and the protein along with it, into a gas phase where it acquires charge. It is then propelled through the mass spectrometer by an electric field. How fast and far it flies through the mass spectrometer tube depends on its mass to charge ratio. The molecular mass of the protein can then be compared with molecular mass calculated for proteins encoded by open reading frames of the organism's genome, if the genome sequence is known. Assuming that the mass spectrometer also quantitates the amount of the protein, you have a way to identify and quantitate many expressed proteins of an organism.

Alternatively, mass spectrometry uses the electrospray ionization method, often in two stages (MS/MS). Proteins, or peptides from proteolytic degradation with a sequence-specific protease, are separated by hplc, and sprayed through an orifice to create droplets from which the solvent evaporates, leaving charged protein in the gas phase, where it can be separated by mass/charge ratio. Particular peptides can be directed into a second mass spectrometer, where they are broken down at the peptide bonds, creating a ladder of peptide fragments whose mass can be measured. From this the sequence of the peptide can be deduced. In the best case, the whole sequence of a protein can be deduced from fragments analyzed by mass spectrometry. Failing that, with large fragments of protein sequence one can search the protein sequence data base to find a protein with these same sequences.

A mass spectrometry-based method for determining differential expression involves isotope-coded affinity tags (ICAT). The reagent has a reactive group, a spacer and a biotin group; two versions differ by whether the spacer has 8 hydrogen atoms or eight deuterium atoms. The reactive group reacts with cysteines. One sample is reacted with the hydrogen version, the other with the deuterium version, and the mixture of them digested with a protease. The labeled peptides can be enriched by binding the biotin tag to streptavidin, resulting in a simpler peptide mixture. This could then be analyzed by mass spectrometry. Quantitation of expression could be done by comparing the amounts of fragments 8 mass units apart, representing the same peptide from the two initial samples. The authors followed the differential expression of more than 1400 proteins in yeast by this technique.

A second major technique in proteomics is the use of microarrays. This involves the attachment of known reagents, for instance monoclonal antibodies, at defined positions on a solid support, often a chemically derivatized glass slide. An extract, often with the proteins labeled with fluorescent or radioactive tags, is flowed onto the array slide, allowed to bind, and unbound protein washed off. The amount of protein bound to each reagent can then be measured by fluorescence or radioactivity, or by attaching a secondary antibody which is labeled with an

enzyme or fluorescent label. Thus the amounts of known antigens can be measured – though not always linearly, only 30% of antibodies in a study bound antigen in linear proportion to the amount present.

The reagents in a microarray of course are not limited to antibodies; they can be antigens which bind their antibodies, DNA or RNA sequences which bind transcription factors, polysaccharides, allergens, small synthetic organic molecules, etc. The key point is that you know what you put on the slide and where you put it, so that you know what reaction occurred with. I don't know that microarrays can be combined directly with identification techniques such as mass spectrometry, but they don't have to be. If you know that a certain reagent, say one of a large number of peptides created by combinatorial chemistry, binds an appreciable amount of protein, one can use the reagent in a sort of affinity chromatography module to bind a larger amount of the protein, then desorb the protein and subject it to mass spectrometric analysis for identification. Conversely, you could immobilize on a slide a number of proteins you are interested in, then apply some mixture – either of proteins from a tissue or cell extract, or of small molecules, perhaps from combinatorial synthesis – and see what binds to what protein.

Protein post-translational modifications can be investigated as a class. For instance, you can use Fe^{+++} , immobilized by chelation, to bind proteins that have phosphate groups on them. The enriched phosphoproteins can then be digested with trypsin, and the resulting fragments identified by mass spectrometry. Or, especially for rarer modifications, you could use an antibody to the particular modifying group to precipitate proteins containing it.

Another major topic is protein-protein interaction. We believe that many proteins in the cell interact with other proteins, in loose ways which do not allow copurification. Many stronger interactions have been identified by affinity purification. One labeled member of a complex binds to an affinity material, for instance an antibody to the so-called FLAG epitope, and proteins bound to it are resolved by electrophoresis and identified by mass spectrometric analysis.

The loose interactions are often identified by the so-called yeast two-hybrid procedure. This typically uses a transcriptional activating protein such as the GAL4 protein in yeast. One part of this binds to the promoter sequence of the galactose operon, another part to RNA polymerase, to activate transcription of the operon, which can be assayed with a substrate for β -galactosidase. The DNA sequences for the two parts of the GAL4 protein are separated. One is fused to the gene for the protein whose interactions you want to investigate, termed the bait protein; the other is fused to genes for as many proteins as possible, the prey proteins. Both are introduced into yeast. If the bait protein interacts with a prey protein, the two halves of the GAL4 protein are able to get together and activate the GAL operon, β -galactosidase is made, and you have a positive, probably a positive colony on a plate or in a microtiter dish. You then reisolate the prey plasmid and determine what DNA sequence it contains and thus what protein it coded for. Obviously I'm skipping over the molecular biology of making these constructs. This method has drawbacks: it has a lot of false positives, but is limited to nuclear interactions, i.e. if the interaction doesn't happen in the nucleus it can't activate transcription. If the prey is a membrane protein it may not be able to get together with the DNA to initiate transcription. Other two-hybrid techniques have been developed to avoid some of these problems. These techniques can be carried out in a high-throughput system, essentially robotically, so lots of interactions can be identified and then further characterized.

Another big topic is functional microarrays of proteins for analysis of biochemical activities. This requires that a very large number of proteins be individually expressed, purified, and immobilized on a glass plate or in 'nanowells' less than 1 mm in diameter in a plastic plate.

This is not as daunting as it sounds. cDNAs for as many transcribed genes as possible from, say, yeast are put into plasmids with a sequence for glutathione-S-transferase, so that all proteins are expressed fused to this protein. Each colony must then be transferred to a well of a microtiter plate – a plastic plate with typically 96 wells – and grown there. Then glass beads are put in and the plate vortexed to break open the yeast cells. The contents of the wells were filtered – they must be sucked out of the wells and transferred to a 96-well filter – and the filtrate mixed with glutathione-Sepharose so that the specific proteins adsorb to this and other yeast proteins can be washed away. Then the adsorbed proteins are eluted, presumably with free glutathione, and transferred to new wells or glass surfaces where they can be immobilized and assayed for some biochemical activity. For instance, they were assayed for binding the regulatory protein calmodulin, using biotinylated calmodulin. The washed slide was then incubated with dye-labeled streptavidin, which binds to biotin residues. Where the dye was, streptavidin was, therefore calmodulin had bound to the yeast protein. Thirty-three new binding partners of calmodulin were identified in this way. They also tested the expressed proteins by electrophoresis and western blotting – using antibody to the glutathione-S-transferase - and showed that 80% of the expressed proteins were as long as predicted from the cDNA sequence.

Obviously this approach is best for fairly general types of activity which many proteins may be expected to have, such as protein kinase activity. One can for instance immobilize various substrates in wells, apply a different kinase, plus [³²P]-ATP, to each well, and determine by presence of radiolabel after washing which kinases could phosphorylate that particular substrate.

Once you have a slide or ‘chip’ with many proteins immobilized on it, you can use it to look for many activities, until it wears out. The only limit is devising an assay which is measurable on chip, and whose product can then be washed away for further use of the chip. One can also probe the proteins for post-translational modifications, using lectins for glycosylation or antibodies for other types of modification.

Arrays of peptides and of carbohydrates can similarly be constructed and used to look at, for instance, sequence specificity of kinases - which peptides get phosphorylated? Therefore, what in the substrate sequence is important? Similarly, small molecules, especially from a combinatorial chemistry library, can be immobilized on chips and used to probe for proteins which bind to them, using a fluorescent labeled protein mixture. One hopes that some small molecule will not only bind to but inhibit the function of some cellular protein, offering a way to affect cellular metabolism. The structure of the small molecule must be known through some way of tagging how it was synthesized, but this is a topic in discussion of combinatorial synthesis (typically the assembly of molecules with similar architecture but varying sequence, as for instance different peptides, but not limited to standard peptide bonds; any attachment which can be repeated is useful).

General reference: Zhu, H., Bilgin, M., and Snyder, M. *Ann. Rev. Biochem.* **72**:783-812 (2003).

ICAT labels: Gygi, S.P., et al., *Nature Biotechnol.* **17**:994-9 (1999)

Kinase specificity: Zhu, H., Klemic, J.F., Chang, S., Bertone, P., Casamayor, A. et al., *Nature Genetics* **26**:283-9 (2000)

Proteome chip, calmodulin binding: Zhu, H., Bilgin, M., Bangham, H., Hall, D., Casamayor, A., et al., *Science* **293**:2101-5 (2001)