High Throughput Mass Spectrometry-Based Proteomics: Principle and Application

Brett S. Phinney

Proteomics as it is most commonly referred to is the study of the total complement of proteins in a closed system, the protein complement of the genome. Proteomics is a new kind of systems biology. Instead of studying one or more components of a system at a time, proteomics strives to study the whole system at once. Proteomics strives to do for proteins, what microarrays do for nucleic acids. Proteomics has only become possible in the last 10 years or so because of several advances, both in technology and in information. On the technology side are powerful mass spectrometers capable of generating sequence information from thousands of peptides in a single day. On the information side is the rapid genome sequencing projects currently finished or in the process of finishing. Powerful computers can then compare the data generated by the mass spectrometers with DNA or protein sequences in public or private databases to rapidly identify thousands of proteins in a single day.

Proteomics is still in its infancy and suffers from the common failures of any new technology. The instrumentation is still relatively difficult to use, is changing rapidly and is very expensive (a mass spectrometer can be $500,000 or more). Software is often incomplete or full of bugs and most of the technical information needed for doing mass spectrometry based proteomics in not written down and must be re-invented by each particular laboratory. Even if you are able to get everything working properly, even with the best mass spectrometer and fastest computers we can still only sample 30% of a proteome in one experiment. Having said this, proteomics is able to do experiments in a few days or months that were not even conceivable a few years ago. As the result it offers immense promise for almost every area of the life sciences. From studying relative abundance changes between cell states to deep protein mining, to studying the members of protein complexes, the potential of mass spectrometry based proteomics can only increase with time.

B. S. Phinney
Room 3B Biochemistry
Michigan State University
E. Lansing, MI 48824

phinney1@msu.edu

Proceedings of the 55th Reciprocal Meat Conference (2002)

But why bother with proteomics at all if gene expression can tell us what is going on? Gene expression does not usually correlate well to the amount proteins do to such things as mRNA stability, translation efficiency and protein turnover. The proteome is dynamic and can rapidly change over time, as opposed to the genome, which is more or less static. The proteome is also different between different cells or even compartments within those cells. Proteomics can study the proteins in a particular cell type or even a certain cellular compartment at a certain time, something gene expression cannot always do.

In this manuscript I will attempt to describe several types of experiments that you can do using proteomic methodologies, how you typically do those experiments and where the future is heading in terms of sample through-put, instrumentation and sensitivity.

Proteomics experiments, as we practice it, centers around protein identification by mass spectrometry. This involves several important steps. First one must obtain the protein sample. This sample can be quite complex consisting of several thousand proteins. These complex samples can come from such places as a cell lysate, or sub-cellular fractionation, a purified envelope or organelle. Protein samples can then either be separated by SDS-PAGE gel (1d or 2d) (figure 1) or analyzed without prior separation (see below). Usually these proteins are digested with trypsin, to digest them into peptides (Shevchenko et al., 1996). These peptides are what are analyzed by mass spectrometry. Mass accuracy tends to be much better when dealing with peptides than with whole proteins, and statistically if you can sequence a five or six amino stretch of a peptide, you can uniquely identify the protein it was derived from. In addition, post-translation modifications and sequence errors in the public databases tend to decrease the utility of identifying a protein from its intact molecular mass.

Once these proteins are digested into thousands of peptides, they must be crudely separated and ionized. Usually these peptides are separated by reversed phase liquid chromatography before they are ionized and analyzed by a mass spectrometer (LC/MS). If the sample is very complex, it may contain hundreds of thousands of tryptic peptides. This is still much too complex a sample to be separated by standard LC/MS. Current mass spectrometers can sequence several co-eluting peptides (usually up to eight). But with that

many peptides, the mass spectrometer will only be sampling a small percentage of the peptides eluting from the reversed phase column. With such complex mixtures of peptides it is common to employ a technique called multi-dimensional liquid chromatography protein identification technology or MUDPIT (Washburn et al., 2001). The MUDPIT approach is starting to be considered as an alternative to two-dimensional electrophoresis. Two-dimensional electrophoresis has modest detection limits, limited dynamic range (2-3 orders of magnitude), has an inherent protein bias based on a limited pI range and is difficult to automate and is very labor intensive. The MUDPIT approach, on the other hand, has a much larger dynamic range (10^5), has a limited protein bias and can be highly automated. This technique involves first separating the peptide mixture by strong cation exchange and then taking the fractions and subjecting them to a reversed phase separation (figure 2). This technique can either be done automatically online with the mass spectrometer (Washburn et al., 2001) or can be done off line and the fractions from the SCX column manually collected and put in the autosampler of a LC/MS system. Both techniques have their respective advantages and disadvantages. The offline method allows you to load a rather large amount of protein on the strong cation exchange column, up to 4-5 mg for a 2mm SCX column (Schlatzer et al., 2002). This gives you the possibility of identifying low abundance proteins that may only be present at 1-2 copies per cell. The online method allows greater automation and limited human interaction, but suffers from a lower sample loading amount usually 5-50 ug of total protein (Hancock et al., 2002). There are also other techniques of decreasing sample complexity such as purifying only the peptides containing cysteines (Gygi et al., 1999) or methionines tryptophans or histidines (Hancock et al., 2002).

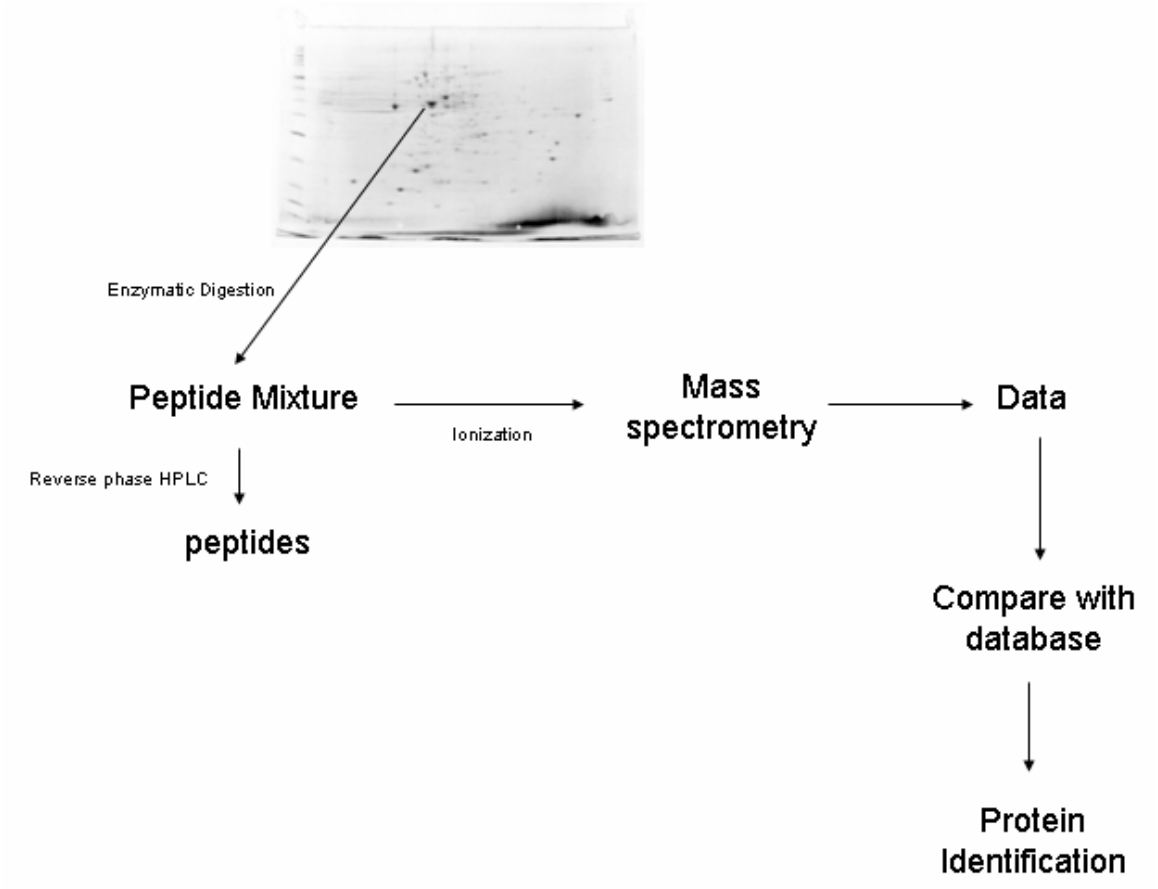

Figure 1. Representative diagram of a proteomics experiment, utilizing a conventional 2-d gel. Gel spots from the 2-d gel are excised and digested with trypsin. The extracted peptides are then analyzed by mass spectrometry and identified.

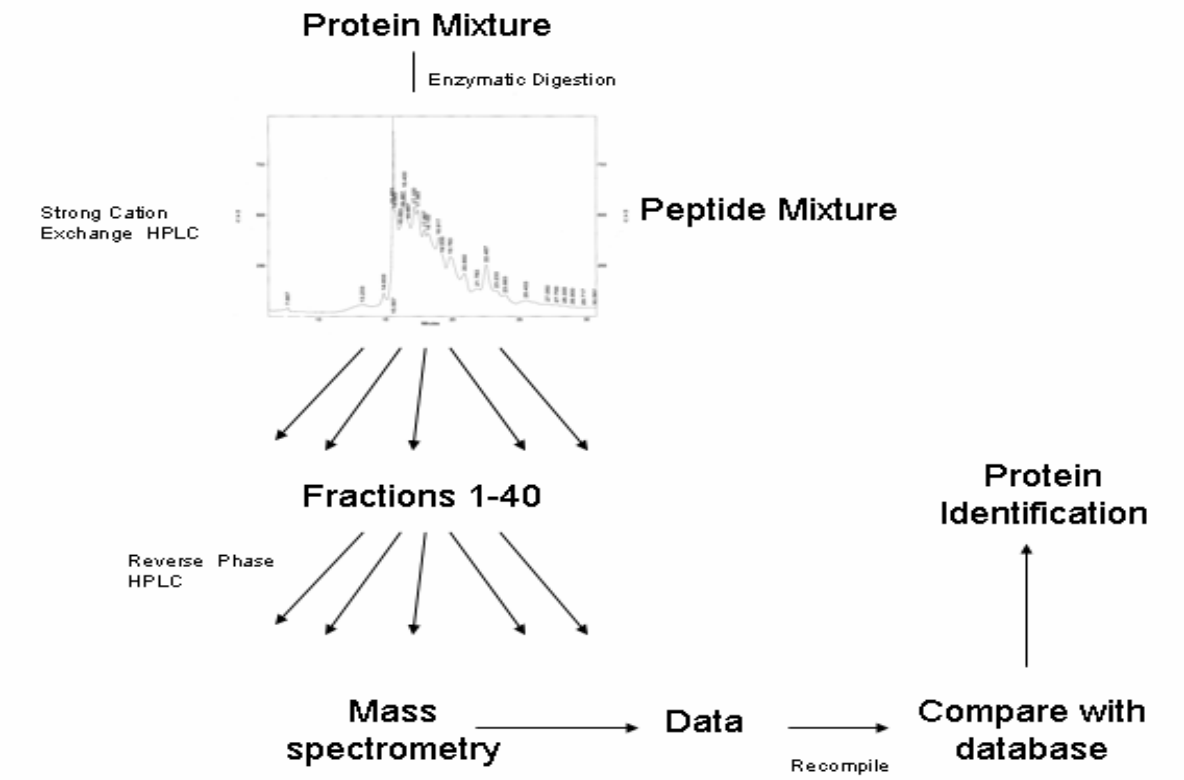

Figure 2. Representative proteomic MUDPIT experiment. A complex mixture of proteins is digested with trypsin and then the tryptic peptides are separated by strong cation exchange chromatography (SCX). Each individual fractions from the SCX separation is subjected to reverse phase separation coupled with tandem mass spectrometry (LC/MS/MS)

Once the complexity of a sample has been decreased, it must be ionized and fed into a mass spectrometer. This is done in one of two ways, either by electrospray ionization (ESI) or matrix assisted laser desorption ionization (MALDI). Electrospray ionization has the advantage that the LC system can be directly coupled to the mass spectrometer allowing integration between the two instruments. It also allows easy sample clean-up using a small reverse phase trap which can be automated by automatic column switching to remove salts and other impurities from the peptides. It suffers from its difficulty in set up and use and high cost of consumables (reversed phase columns). In our lab we use 75 um x 15 cm self-packed reverse phase columns with a flow rate of 200 nl/min into our mass spectrometers. MALDI on the other hand is relatively easy to use and is very fast, some instruments have a 200 Hz laser capable of ionizing a sample at an astonishing rate. Also, your sample is not completely destroyed by the ionization process allowing you to go back and analyze your sample again. The down side is that it is much harder to connect an LC system to a MALDI interface and work around steps must be employed such as using specialized equipment to mix the eluant from a LC column with the MALDI matrix and then lay the LC "track" on a MALDI plate and then manually placing the plate in the mass spectrometer.

Once the peptides have been ionized the mass over charge (m/z) of a peptide ion is measured and the peptide ion is usually isolated from all the other peptide ions that are being simultaneously ionized. This isolated peptide ion is then usually collided with an inert gas such as argon or helium with fragments the peptide ion into a complex set of ion series (Figure 3). The m/z of the fragment ions are then recorded and associated with its precursor ion. The mass of the precursor ion and its associated fragmentation ions are what are used for identifying the protein from a database.

Identifying a protein involves searching the non- interpreted product ion spectra using several commercially available programs such as SEQUEST (Yates et al., 1996) and Mascot (Perkins et al., 1999). Thousands of spectra can be matched daily using powerful computer clusters. Of course if your protein of interest is not in an amino acid or nucleic acid database, the software will not be able to identify it. Often a homologous protein from a different species is sufficient to identify your protein if the organism you are working on does not have a large set of known proteins or genes.

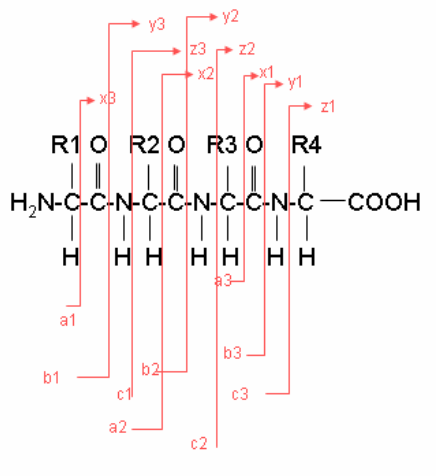

Figure 3. Low energy peptide fragments typically seen in tandem mass spectrometry experiments.

At Michigan State University we can acquire approximately 20,000 MS/MS spectra daily using our LCQ Deca Ion tap mass spectrometer (Figure 4). We operate the instrument in a top 3 or 4 double play type of experiment where the top 3 or 4 peptide ions that are co-eluting are fragmented and recorded (figure 5). We can on a routine basis identify hundreds of proteins a day using this method. Ongoing projects include looking at the proteome of the chloroplast, mitochondria, tomato and various other bacteria and organisms.

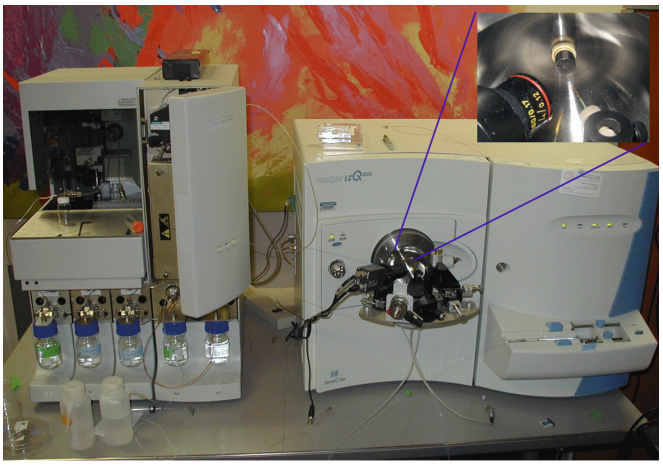

Figure 4. A Thermo-Finnigan Ion trap mass spectrometer coupled with a waters CapLC. The zoomed image is the tip of a 75 um column in front of the inlet of the mass spectrometer.

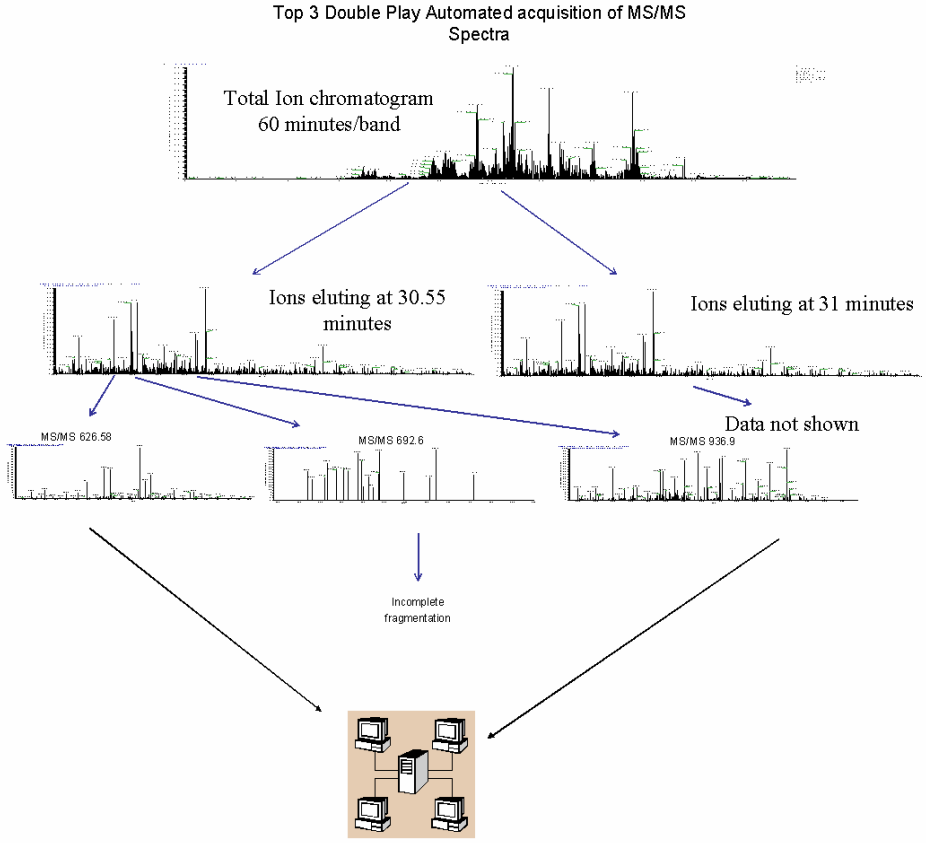

Figure 5. Top 3 double play experiment. The top three co-eluting peptide ions at any one time are isolated and fragmented. The non-interpreted product ion spectra are automatically searched against the non-redundant database. MS/MS ion matches are pooled and protein identification hits are scored using a probability based Mowse Score.

Another type of proteomics experiment we offer is looking at the relative quantitative differences in proteins between samples. This can be done by labeling the proteins in each sample with a stable isotope tag. Such tags can be 18O, 15N (Washburn et al., 2002) or a similar tag. We use the ICAT Tag (isotope-coded affinity tag) (Gygi et al., 1999) produced by Applied Bio-Systems. A tag is considered necessary by many because different peptides tend to have differing ionization efficiencies, although a few papers have quantitated without the use of stable isotope tags (Bondarenko, 2002). The ICAT tag selectively binds to cysteine containing peptides and has an attached biotin label and either 0 or 8 deuterium atoms attached. This reagent is added to the samples; the samples are then mixed and processed together. This allows sample handling errors between experiments to be minimized. You can then compare the intensity of the D0 and D8 peptide ions and derive a relative abundance increase or decrease, usually within 10-20%. This technique also has the added benefit that the complexity of the sample is reduced to only cysteine containing peptides by using an avidin column to selectively remove the tagged cysteine containing peptides. Figure 6 shows an ICAT labeled tryptic peptide ion pair from two samples of ovalbumin. One sample contained 2 ug and the other contained 4 ug of total protein. The samples were labeled with the ICAT reagent, digested with trypsin and processed according to the manufactures instructions. As expected the ICAT labeled pairs have a 1:2 intensity ratio. This type of experiment can be done on a sample containing hundreds or thousands of proteins. This allows an investigator to get a snap shot of what proteins are increasing or decreasing in abundance, at any given time.

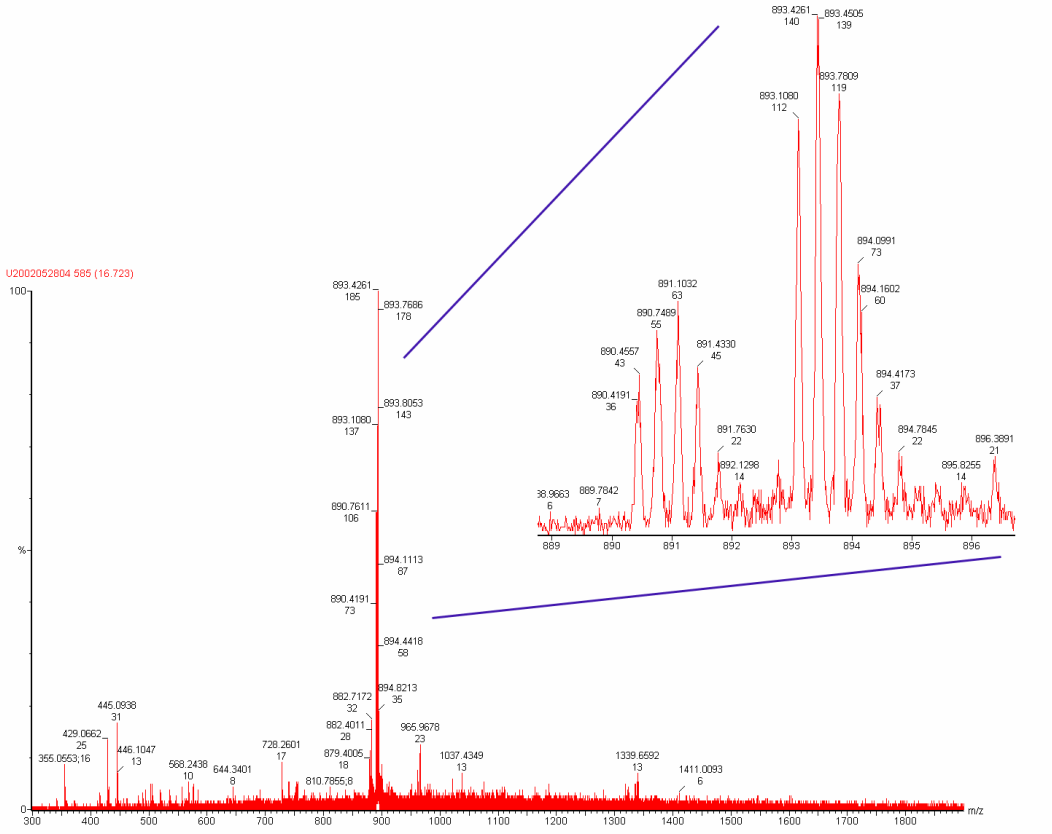

Figure 6. An ICAT pair of tryptic peptides derived from a digest of two samples of ovalbumin. One sample contained 2 ug of protein while the other contained 4 ug. The relative abundance ratio of both samples can be ascertained by the peptide ion intensity.

High-throughput mass spectrometry is a definition that is constantly changing. At one time obtaining one or two MS/MS spectra in a single day must have been considered high-throughput. Now with advances in instrumentation such as the new ABI Tof-Tof one can hope of generating ten thousand MS/MS spectra in a single day. The question then becomes how to make sense and store that massive amount of data.

References

Bondarenko, D.C.a.P.V. 2002. Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry. Journal of Proteome Research. ASAP Article.

Daniela Schlatzer, J.A., Kevin Blackburn, William Burkhart, Roderick Davis, Alex Iving, Jungping Jing, Sue Kadwell, Joanna Krise, Mary Moyer, Arthur Moseley, Kieran Todd, James Ward & Patrick Warren. 2002. LC/LC/MS/MS Analysis of a Complex Proteome and Protein Classification Using Gene Ontology. In ASMS. Orlando: ASMS 2002 poster.

Gygi, S.P., et al. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature Biotechnology 17:994-999.

Hancock, J.T.S.D.A.K.L.O.C.T.M.K.C.S.-L.W.P.S.W.S. 2002. A Comparison of Shotgun Sequencing and 2-D Electrophoresis for the Identification of Proteins in Complex Mixtures: *Escherichia col*i Cell Lysate. in ASMS. Orlando: ASMS 2002.

Perkins, D.N., et al. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551-3567.

Shevchenko, A., et al. 1996. Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels. Analytical Chemistry 68:850-858.

Washburn, M.P., D. Wolters, and J.R. Yates. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature Biotechnology 19:242-247.

Washburn, M.P., et al. 2002. Analysis of quantitative proteomic data generated via multidimensional protein identification technology. Analytical Chemistry 74:1650-1657.

Yates, J.R., et al. 1996. Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. Journal of the American Society for Mass Spectrometry 7:1089-1098.