KEYWORDS: mass spectrometry, proteomics, genomics

# Proteomics in the Post-Genomics Era: Weighing in the Data

**Ben Montpetit**

Biochemistry and Molecular Biology, University of British Columbia
Submitted December 2002

## The Birth of Proteomics

The post-genomics era – a time when genomes are being sequenced and released with fewer flourishes in each instance. Part of the reason for this is that the human genome has been completed, but a more significant reason is that there has been a realization that the genome provides a basis for evaluating what is possible, but does not provide the critical information about what "is" in a given organism, organ, tissue or cell. Genome sequences do not always provide a direct link to biological activity, since it is the complex interwoven pathways governed by proteins that are responsible for an organism's phenotype. For this reason the genomic era will be partly responsible for the birth of the proteomics era.

Proteomics can be defined as the large-scale study of proteins (ultimately the whole proteome) usually by biochemical methods. Following this definition, the start of proteomics can be traced back to the 1970's with the establishment of 2-D electrophoresis[1], during which time 2-D gels were used to map all genes being expressed in specific tissues and the resulting changes when that system was perturbed. This process led to the construction of large databases of 2-D gels, but the lack of a simple, fast, and accurate method of identifying these proteins retarded the development of this field. However, the adaptation of mass spectrometry (MS) for use on proteins removed these obstacles and provided a tool to bring proteomics in to the forefront. Coupled with the introduction of MS was the ever-increasing power of computers and the development of software to provide an ability to store, manipulate, display large amounts of data in a biologically significant way, and to automate the process to some degree.

## Why proteomics, when we have the genome?

As mentioned above the genome does not provide us with a direct link to biological activity because it is proteins, not genes, which are the active components in cells. For this reason and the ones to follow, proteomics will become very important in dissecting complex biological systems by doing the things that genomics cannot do at this time. For example, the arrangement of genes in the genome provides no information about the protein complexes that a gene product may be a part of, or it's level of expression, or for most cases it's cellular localization. Furthermore, an open reading frame does not always mean a protein will be produced. Gene prediction itself is a difficult task. Recently, it was concluded that the error rate for prediction was at least 8% in the annotation of the 340 genes from *Mycoplasma genitalium*[2]. When this is extrapolated to the human genome which contains ~60,000 genes this could mean an error involving upwards of ~6500 genes. Protein-protein interactions and post-translational modifications are another level of detail that is not addressed by genomics, but are very important in determining cell function. For example, recent work indicates that >83% of proteins exist in at least one complex (i.e. several proteins working together) and only ~17% of proteins function as single entities[3]. When all this is considered it can be seen that the genome is insufficient on its own. The genome provides a large database from which ideas and hypotheses can be drawn, but it is going to be the use of proteomics in conjunction with genomics that will provide a powerful tool for studying complex biological problems. One of the main foundations of proteomics in this quest for information will be MS.

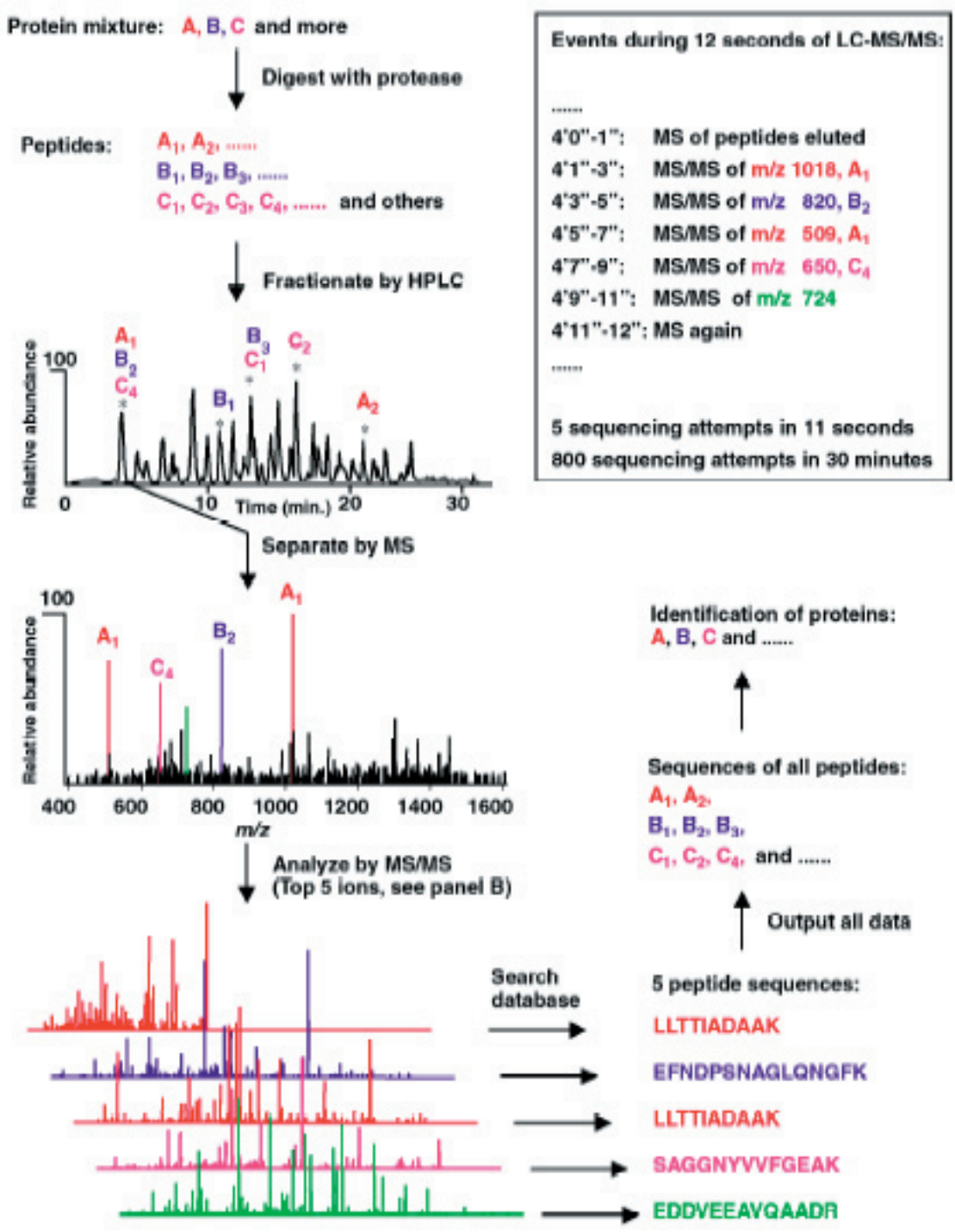


**Figure 1** Schematic representation of a protein identification procedure using tandem MS/MS and database searching (from reference 4)

## The Basics of Mass Spectrometry

MS is a process of weighing molecules. It has been noted in the literature that MS could be working at the smallest scale in the world because of the things it weighs. MS is a technique in which a molecule (ranging from small chemical compounds to large supermolecular protein complexes) is fragmented into small charged fragments that are then analyzed based on their mass to weight ratio, providing the "parts" from which a large molecule is built. When applied to

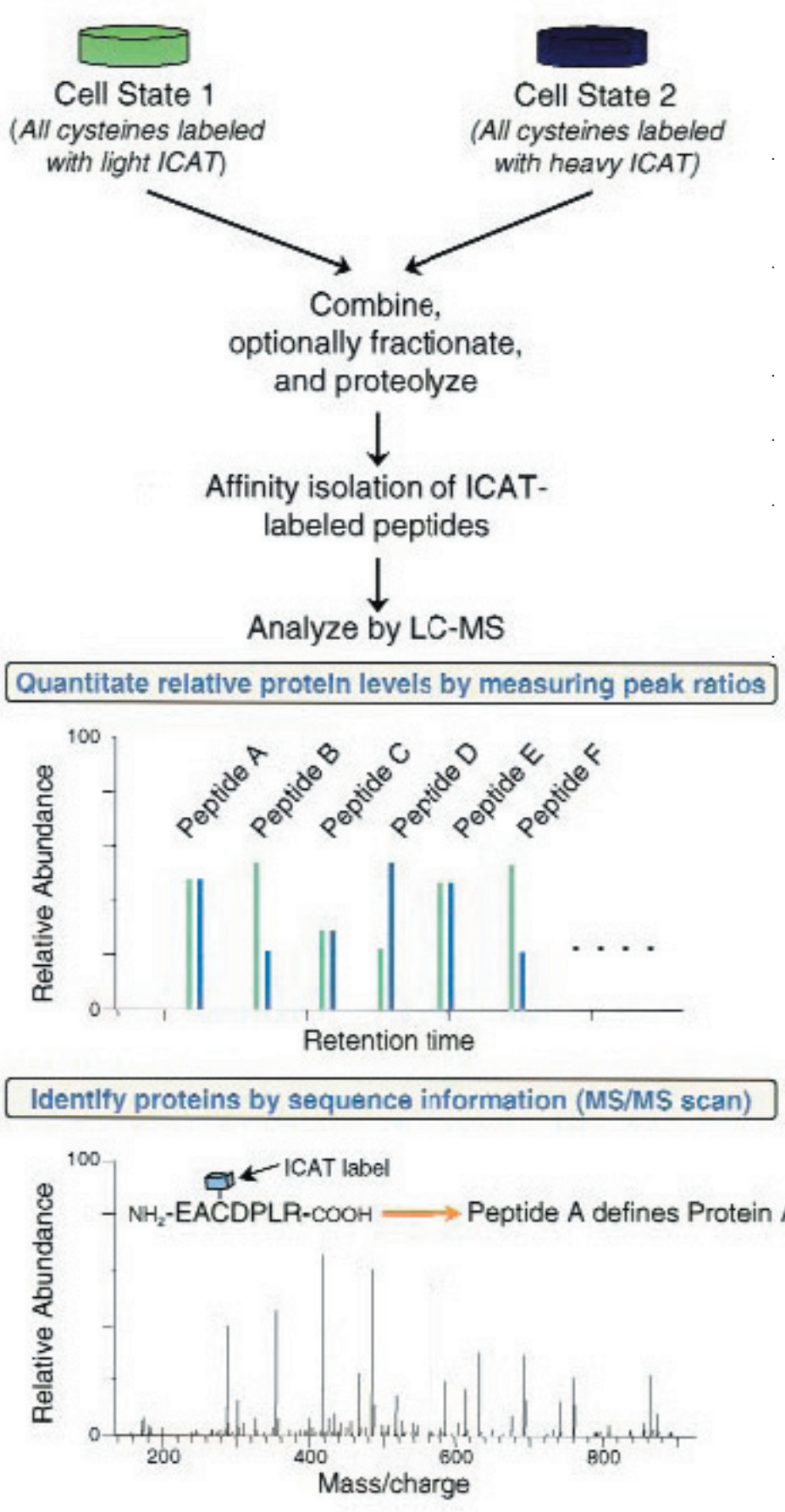


**Figure 2** Schematic representation of the general ICAT labeling scheme.

- proteins from two different cellular states are labeled with a isotopically heavy (deuterium in place of hydrogen) or light isotope coded affinity tag (ICAT)
- the tag consists of a reactive group used to covalently bind proteins through cysteines, a light or heavy linker, and a biotin group used for purification of labeled proteins
- protein from each state are pooled, proteolyzed, and purified over a biotin affinity column
- this peptide pool containing the differentially tagged proteins in the analyzed by MS
- since the two samples were taken from cells expressing the same proteome, with exception of any proteins that are differentially regulated, proteins from each sample should co-elute with the difference being the light and heavy tag
- proteins showing differences in relative abundance are selected for further analysis by tandem MS/MS to obtain sequence information and ultimately identifying the protein.

proteins, one can use a protease so that the protein is fragmented in a particular manner that is reproducible, which then generates a specific "fingerprint" for every protein. Coupled with the ability of computers to store and search databases that include previous identified spectra it is possible to identify isolated proteins by using the spectra of these known catalogued proteins. Recent developments have taken this one step farther by combing two MS machines in a row (tandem MS/MS), which allows fragments to be analyzed, specific fragments to be selected from the primary pool and further fragmented in the second mass spectrometer to produce fragments consisting of individual amino acids (i.e. a protein sequence) (see Figure 1). These amino acid sequences can then be searched against ORF, cDNA, EST, and genome databases allowing the identification of novel proteins never catalogued before. This is a very powerful tool and is currently being used in numerous fields to identify novel interacting proteins within known complexes[5].

## ICAT: A unique approach

As is the case with most techniques, MS is continuously being readapted to produce information that is more detailed. Recently, Gygi et al. published a paper[6] that uses MS not only to identify the presence of specific proteins, but also to quantify the changes in protein levels in two biological states. This is schematically depicted in Figure 2.

ICAT was used by Gygi et al. to monitor the changes in gene expression caused by changing the carbon source of growing yeast. They were able to identify proteins that were previously known to respond to these different carbon sources. Moreover, they were able to identify proteins that responded to the change, but were previously not known to do so. The identification of numerous proteins that were previously unknown to cooperate in a certain pathway demonstrates the powerful ability of MS to produce large amounts of data that can rapidly be followed up and verified by other biochemical or molecular methods.

## Proteomics, proteomics, proteomics

Proteomics – this is likely to be a buzzword for the next few years taking the place of genomics in the minds of many. There are good reasons to be excited by this expanding field, which includes the ability to study the active components of cells directly. Key to this success is the fast, accurate, scalable and automatable MS, which is able to look at problems involving post-translational modifications and protein-protein interactions, areas inaccessible by genomics. However, one must stress that the field of genomics is far from being left behind and it is only because of the information developed by the large scale genomic projects that had allowed proteomics to come so far so fast. Furthermore, advances in genomics over the next few years will likely push proteomics even farther through the generation of data and hypotheses that can only be addressed by proteomic means.

## References


1. O'Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. 250, **4007–4021 (1975).**
2. Brenner, S. E. Errors in genome annotation. *Trends Genet.* **15,** 132–133 (1999).
3. Gavin, A.C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147(2002)
4. Peng, J. Gygi, SP. Proteomics: the move to mixtures *J Mass Spectrom.* **10**,1083-91. (2001)
5. Sanders SL, Jennings J, Canutescu A, Link AJ, Weil PA. Proteomics of the eukaryotic transcription machinery: identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry. *Mol Cell Biol.* **Jul;22(13)**:4723-38. (2002)
6. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R.Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol.* **Oct;17(10)**:994-9 (1999)