Protein Identification by Mass Spectrometry

ISSUES TO BE CONSIDERED

Michael A. Baldwin‡

During the past two decades, mass spectrometry has become established as the primary method for protein identification from complex mixtures of biological origin. This is largely attributable to the fortunate coincidence of instrumental advances that allow routine analysis of minute amounts (typically femtomoles) of involatile, polar compounds such as peptides in complex mixtures, with the rapid growth in genomic databases that are amenable to searching with mass spectrometry (MS) data. Like many other developing fields in science, the creation of techniques and software tools and the initial generation and interpretation of data have been the domain of experts, people who are cognizant not only of the benefits of the methods but also of their actual and potential weaknesses. Now, as mass spectrometric techniques and proteomic tools become increasingly available and accessible, a much broader range of researchers is applying the same methodology, often with substantially less understanding of the major limitations that critically affect the reliability and significance of the results. Ideally, the MS community should establish criteria for mass spectrometric identification of proteins that should be employed by researchers. As this remains a rapidly developing field with many different experimental approaches and different ways of searching and interpreting the data, it is difficult to promulgate hard and fast rules. Nevertheless, Molecular & Cellular Proteomics is attempting to develop standards of acceptability for proteomics papers based on emerging knowledge as well as on principles of biological MS established over the last 20 years by the MS community. Authors of proteomics papers employing MS must make themselves fully aware of the key issues that are driving development of these guidelines. Hence, the paper that follows attempts to highlight the strengths and weaknesses of the methods in current use. It is particularly important to realize that for any protein match returned from a database search, there is a non-zero probability that it will be wrong. Many times, the quality of the data is such that the probability of a false positive can be disregarded, but in some cases the identifications returned by the search engines are very likely incorrect. Therefore, it is unacceptable to simply list all the hits that come back from any search engine and then discuss their biological significance as though they were categorically correct.

PEPTIDE ANALYSIS

Almost without exception, protein identification is based on the analysis of peptides generated by proteolytic digestion. The most widely used enzyme is trypsin, which hydrolyzes the protein on the C-terminal side of lysine and arginine, unless the subsequent amino acid in the sequence is a proline. This is advantageous as every peptide other than the protein C terminus has at least two sites for efficient protonation, the N-terminal amino group and the C-terminal basic residue, so peptides are readily ionized and detected as positive ions. However, for a variety of reasons, it is normal for only a subset of the potential tryptic peptides from any protein to be detected, particularly when the peptides are ionized directly from unseparated mixtures in which there may be competition for the available protons. There are also experimental limitations for the detection of peptides that are either very small or very large, a factor that is not controllable with a single enzyme as this is dependent on the distribution of lysine and arginine residues within a protein. Protein sequence coverage can be improved by carrying out a parallel digestion with a second protease of different specificity such as chymotrypsin, then combining the two digests for a common analysis. In practice, whether a large or small fraction of the peptides generated from any protein is detected depends on many variables: the amount of that protein present in the original sample; the efficiency of any protein extraction, digestion, and peptide extraction; the presence of other proteins and other impurities; and the sensitivity and performance characteristics of the mass spectrometer and its mode of ionization, mass separation, and ion detection.

Mass spectrometers employed in proteomic analysis use either matrix-assisted laser desorption/ionization (MALDI) or electrospray ionization (ESI) but they vary widely in their operation and performance characteristics. Early database searching was based on low-resolution linear MALDI-time-of-flight (MALDI-TOF) giving a mass accuracy of perhaps ±2 Da.

From the Mass Spectrometry Facility, Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-0446

Received, November 6, 2003

Published, MCP Papers in Press, November 6, 2003, DOI 10.1074/mcp.R300012-MCP200

1 The abbreviations used are: MS, mass spectrometry; MALDI, matrix-assisted laser desorption/ionization; ESI, electrospray ionization; TOF, time-of-flight; LC, liquid chromatography; MS/MS, tandem MS; QqTOF, quadrupole mass selector and quadrupole collision cell with orthogonal acceleration TOF; CID, collision-induced dissociation; HPLC, high-pressure LC; ICAT, isotope-coded affinity tag.

2 S. A. Carr, personal communication.
Issues in Protein Identification by Mass Spectrometry

(1). However, this is no longer acceptable because good mass accuracy increases the reliability of database searching by limiting the possible compositions of peptides for any given mass (2). A delayed extraction MALDI-TOF instrument with reflectron should give better than 50 ppm mass accuracy, and significantly better (~10 ppm) may be achieved with careful internal calibration. ESI has been the standard ionization method for liquid chromatography (LC)-MS and LC-tandem MS (MS/MS), although separated fractions can be deposited onto a MALDI target for either on-line or off-line analysis (3–5). ESI is usually employed for single and triple quadrupoles and quadrupole ion traps that typically give modest resolution. The combination of a quadrupole mass selector and quadrupole collision cell with orthogonal acceleration TOF (QqTOF) gives high resolution (~10,000) and perhaps 5 ppm mass accuracy if well calibrated (6–9). Fourier transform MS gives the ultimate performance with mass accuracy of perhaps 1 ppm, but such instruments are more expensive, more technically demanding, and until now their deployment has been mostly limited to large MS facilities.

PROTEIN IDENTIFICATION BY DATABASE SEARCHING

There are at least three approaches to protein identification based on peptide analysis. The first developed is generally referred to as peptide mass mapping or mass fingerprinting (1, 10–15). This relies upon a comparison of the experimentally determined MS peak mass values with the predicted molecular mass values of the peptides generated by a theoretical digestion of each protein in a database. Early observations with relatively small databases suggested as few as three or four peptide matches could be sufficient to identify the correct hit, even with relatively low-precision mass measurements from linear MALDI-TOF instruments (1). But recently, genomic databases have grown rapidly, e.g. as of June 14, 2003, the NCBInr database contained 1,446,218 entries, which was 50% more than 1 year earlier. Consequently, the criteria for protein identification have become more stringent and more accurate mass measurement is essential. It is also necessary to match a larger number of peptides and to cover a larger percentage of the protein sequence. The second approach uses collision-induced dissociation (CID) spectra of individual peptides from MS/MS. Prior to the development of database searching, peptide MS/MS spectra from high performance multi-sector instruments with fast atom bombardment ionization were used for de novo peptide sequencing (16). In the mid 1990s, such MS/MS spectra, generated by MALDI or ESI, were matched against sequence tags predicted for all proteins in a database, i.e. (short) series of fragment ions that could be attributed to coherent sequences of amino acids corresponding to subsets of the predicted peptide (17–20), as implemented in programs such as Protein Prospector’s MS-Tag. In their latest form, such searches are based on comparisons between the experimentally observed fragment ions and all predicted fragments for all hypothetical peptides of the appropriate molecular mass, based on known fragmentation rules. Each peptide match can be linked to a protein match, and it is possible that even a single peptide will identify a protein correctly, although identical sequences might be duplicated in closely related proteins, therefore matching multiple peptide sequences provides greater statistical confidence. Clearly, the greater the number of peptides being matched to any one protein and the greater the sequence coverage, the greater the probability of a correct identification. Error tolerant searching will allow a peptide to be identified when it differs from a database peptide by perhaps a single amino acid, and techniques have been developed to identify remote sequence homologies, although these are time-consuming and computationally intensive (21, 22). Finally, if no matches are found because the protein is not present in the database, de novo peptide sequencing continues to be a valuable method, based on known rules for peptide fragmentation (23). This is straightforward with good quality spectra and is valuable when a significant proportion of peptides diverge from those predicted, due to errors in databases, discrepancies between genomic sequence and processed or postranslationally modified proteins, species differences, and nonspecific enzyme cleavages. An early example of the benefits of extensive sequencing was the determination of the primary structure of Gal β-1,3 (4)GlcNAc α-2,3-sialyltransferase (24).

The use of MS/MS and CID is becoming the accepted standard for protein identification and is steadily replacing peptide mass fingerprinting, although the quality of tandem data varies considerably with instrument type. Sequence ion information from MALDI-TOF instruments using post-source decay (25) is painstaking to collect and the accuracy of mass assignments is relatively low, therefore it is little used now, particularly as MALDI has become available on higher performance QqTOF (26–30) and TOF/TOF (31–33) instruments. Low-resolution three-dimensional ion traps have been very popular and are well-suited to high-throughput LC-MS/MS, but cannot distinguish different charge states when operating in full scan mode. This is a limitation as ions are separated not by mass but by mass/charge (m/z), and ESI typically gives multiply charged ions. Library searching based on such low-resolution ion trap data typically tests multiple possibilities based on the assumption that the precursor ions can have one, two, or three protons attached. For some search engines, all fragments are assumed to be singly charged, thereby increasing the probability that any peptide match will be incorrect. However, the popularity of ion traps has presented a challenge for software development and several search engines now assume that fragment ions from multiply charged precursors might also be multiply charged. Precursor ion charge states may also be deduced from an analysis of fragment ion spectra (34). With higher resolution instruments, charge states are readily determined unambiguously from the peak spacing, e.g. with QqTOF instruments, and Fourier
transform MS has identified charge states for protonated molecular ions and fragment ions of proteins of up to 45 kDa (35). It is also important to add that new ion trap designs, including linear traps, provide much higher resolution and, in combination with other mass analyzers, will form the basis of a new generation of powerful tandem instruments.

A largely uncontrolled variable factor in database searching is the choice of search engine, of which there are several in common use (36). A number of web sites offer free access to web-based database searching programs for peptide mass fingerprinting and the identification of sequence tags, all of which provide other tools as well, such as programs for calculating isotope patterns, predicting enzyme digestion patterns, and theoretical prediction of CID fragments. Some sites allow free downloading of the programs. An alternative to internet access is to purchase a license to have the programs resident in-house, and some mass spectrometer vendors supply software and software licenses with the purchase of an instrument. An enumeration of all the current search engines is a moving target, but well-known web-based examples include tools on the ExPASy proteomics server provided by the Swiss Institute of Bioinformatics (www.expasy.ch/tools), Mascot from Matrix Science (London, UK; www.matrixscience.com), and Protein Prospector provided by the University of California (San Francisco, CA; prospector.ucsf.edu). Systems supplied by instrument manufacturers include Sequest from Thermo-Finnigan (San Jose, CA) and Spectrum Mill from Agilent Technologies (Palo Alto, CA).

DIGESTION OF ISOLATED PROTEINS VERSUS DIGESTION OF PROTEIN MIXTURES

Protein identification by digestion and peptide mass fingerprinting is not effective for complex protein mixtures unless preceded by a separative step, most often a two-dimensional gel separation. Using large format gels, it is possible to separate and analyze hundreds or even thousands of protein spots from a single gel. This is manually repetitive and time-consuming but automated methods have been developed (37), and several companies offer robots for spot identification, spot cutting, digestion, and spotting of the peptide mixtures onto MALDI plates. In practice, gel separation is equally applicable to MS/MS, but an alternative approach is to digest a complex mixture and then to separate the peptides chromatographically before introduction to the mass spectrometer. Such mixtures may contain thousands of proteins and multi-dimensional high-pressure LC (HPLC) may be necessary, e.g. an initial separation by strong cation exchange chromatography may be followed by reversed-phase chromatography (38). Peptide mass fingerprinting is no longer possible with this approach as the association is lost between particular peptides and the proteins from which they were derived, therefore CID and peptide sequencing is essential. The latter method is more easily automated than two-dimensional gel analysis, consequently in many laboratories it is replacing the use of gels. It can also be combined with one-dimensional gel separation in which individual bands that may contain multiple proteins are cut out, digested, and the peptides are introduced to the mass spectrometer via an HPLC separation (39).

AUTOMATION OF DATA COLLECTION AND ANALYSIS

In so-called data-dependent experiments, with HPLC coupled to a tandem mass spectrometer, repetitive recording of MS spectra is interleaved with the selection and analysis of peaks for CID analysis (40). The amount of data generated by such experiments running on a continuous basis is overwhelming for manual interpretation, consequently automation is becoming essential. Effective automated transfer of mass spectrometric data to the informatics programs is dependent on the reliable performance and accuracy of so called “peak-picking” algorithms. Such programs are often proprietary to the instrument manufacturers, although some mathematical approaches have been described (41). Ideally, they convert the instrument-specific and technique-specific raw mass spectra into generic lists of monoisotopic mass and intensity, usually after some spectral processing to enhance the peak envelopes (ion current profiles) while reducing electronic and random noise. The m/z value for each peak is usually defined by its centroid (the center of mass of the peak), determined above a certain fraction of the peak height such as 50%, a variable that is selected to avoid the inclusion of noise at the baseline and to allow some failure to fully resolve adjacent peaks. MALDI-generated ions are singly charged but ESI-generated ions usually carry multiple charges, therefore ESI peaks are separated by fractional m/z values. A challenge is to correctly identify the first peak in each isotopic cluster. For well-resolved, isolated peptide ions of molecular mass less than ~1500 Da with good signal-to-noise ratios, this is relatively straightforward as the first peak in each cluster, i.e. the monoisotopic peak (42), is the most abundant. Above 2000 Da this is no longer true, and the correct identification of the monoisotopic peak becomes progressively more difficult with increasing mass, particularly if it falls below a threshold selected to discriminate between actual peaks and background noise. This problem is compounded by the overlapping isotopic clusters likely to occur in the spectra of peptide mixtures.

SETTING VARIABLE PARAMETERS AND RANKING HITS

Once the experimental data have been recorded, subjected to smoothing, possible baseline subtraction, the peaks extracted, and possibly filtered according to criteria that exclude nonpeptide peaks, the data must then be searched against an appropriate database, as selected by the researcher. All search engines allow for the setting of a number of variables such as protein molecular mass range, pl range, mass tolerance for peaks in a normal spectrum, mass tolerance for CID precursor ions, mass tolerance for fragment ions, number of
Issues in Protein Identification by Mass Spectrometry

charges, number of peptides required for a match, and possible modifications to certain residues such as alkylation of cysteine or oxidation of methionine. The program then returns a series of hits, ranked according to one of a number of possible criteria. Unfortunately, there are presently no universal standards for scoring the output from these programs, therefore deciding/determining what is a significant match is not straightforward. The information provided by the search engines is variable, e.g. the scoring within Protein Prospector has been based on MOWSE (Molecular Weight Search), which allows for the nonuniform distribution of peptide sizes that result from proteolysis (14). This is not particularly informative, but some researchers have specified a MOWSE score of at least 100 for a hit from peptide mass mapping to be significant. Note however that no MOWSE threshold can be defined above which a match is definitely correct and below which it probably is not. At the University of California, San Francisco we are developing a new scoring system for Protein Prospector based on the frequency of observing certain ion types in MS/MS spectra, backbone ions such as y and b scoring higher than side-chain ions such as d, v, and w. Other algorithms that employ probability-based scoring systems including Sequest (18), Mascot (43), and Sonar (44) are being developed to provide more reliable significance thresholds, but again putative matches close to the threshold may well be incorrect.

A number of groups have addressed these problems by developing statistical tools that improve the scoring for existing search engines, particularly Sequest (45–47), while others have developed independent tools (48–51). Ideally, the probabilities assigned will be independent of the mass spectrometer used and the search engine used. Some approaches focus on improved peptide identification, and they generally employ a training set to derive a knowledge base of peptide fragmentation rules. An explicit objective of most of these methods is to make high-throughput experiments more reliable and to minimize the need for time-consuming, visual inspection of data and its subjective, manual interpretation. Such an approach should facilitate the analysis of entire datasets from experiments involving multiple LC-MS/MS runs with perhaps tens of thousands of MS/MS spectra. One method described uses an expectation maximization algorithm to assign a true probability value to each peptide identification. An analysis of the whole dataset identifies “sibling” peptides originating from a single protein. A higher number of sibling peptides observed for a given protein increases the probability of identifying that protein correctly. Thus based on statistical principles, protein identifications based on single peptides carry less weight than those based on multiple peptides (51). Consideration has also been given to the case in which a particular peptide sequence is part of more than one protein, as in many enzyme families.

Experience shows that for any analysis of a digest from a single protein, there will be mass spectrometric peaks observed that are not matched by the search engine. Some of these may be identified as nonpeptide peaks by virtue of fractional mass values (mass defects) that fall outside an allowable window for the elemental compositions of normal peptide ions. Others may be peptides from the targeted protein but resulting from nonspecific cleavages or carrying post-translational or chemical modifications. However, in very many cases, a sample cut from a single gel spot or a fraction corresponding to a single peak in an HPLC run will actually contain multiple proteins, and only a subset of peptides will match any given protein hit. If peptide mass fingerprinting is employed, the peptides that match to the top-scoring hit can be subtracted, then the amended peak list can be searched again for second and subsequent proteins. However, MS/MS is better able to differentiate between peptide and nonpeptide signals and can also identify modified peptides. Multiple proteins will be detected more successfully when MS measurements are coupled with HPLC separation, either on-line or off-line, in part because this allows a wider dynamic range and because it reduces suppression effects, i.e. the selective ionization of only a subset of the peptides present in a mixture.

THE IMPACT OF EXPERIMENTAL DESIGN AND METHODOLOGY ON THE QUALITY OF THE RESULTS

In considering the reliability of protein identifications by MS, a variety of sources of obtaining erroneous results must be considered. First, there is contamination by other proteins during the initial isolation and digestion of the sample, which is difficult to avoid for low-level samples, e.g. during gel electrophoresis, cutting spots from the gel, from sample tubes, pipettes, buffers, or extraneous matter dropping into the sample (such as hair, skin, material from clothing, laboratory dust). Then there is the efficient extraction of the peptides and preparation of the sample for MS while trying to minimize the presence of salts or detergents that may adversely affect ionization. The analysis technique chosen should be appropriate for the sample, e.g. a highly complex mixture should not be analyzed without a separative step, either prior to analysis or on-line as in LC-MS/MS. If the preceding steps are optimized, the mass spectral method is appropriate, the instrument is well maintained, and suitable calibration spectra are run, the data obtained should be as good as can be.

In submitting the data to a search engine, the researcher must first select a suitable database and must ensure that all the variables are set appropriately. If possible and particularly if the user is not very familiar with the relevant search engine, multiple searches should be submitted using different values for the same dataset, and ideally using different search engines. At the University of California, San Francisco, we frequently compare hits from both Mascot and Protein Prospector. Strong protein hits usually coincide with both search engines and may become stronger because of additional peptide identifications. Weak hits based on single weak peptides may also improve if the other search engine finds a
different peptide, which can happen as we consider some different ion types with Protein Prospector, or even if both searches identify the same peptide. In contrast, we are also less likely to believe weak hits that are found by only one search engine. For high-throughput analysis, it is not realistic to review every spectrum, but a selection of spectra should be scrutinized, particularly any that yield particularly important or unexpected findings. In these, the signal-to-noise ratio should be examined, the presence of appropriate isotope peaks should be confirmed, and the isotope spacing should be checked to establish that charge states and mass assignments are correct. However, this will only be meaningful if the raw data can be examined, rather than smoothed, de-isotoped, or otherwise manipulated versions of the spectrum.

The information derived from search results will depend on the nature of the sample, the techniques used, and the particular search engine, but there are some general principles to consider. For peptide mass fingerprinting, there should be adequate sequence coverage for the peptides identified. Thus the matching of four peptides representing 10% of the sequence does not constitute a reliable hit and should not be listed as a positive identification. In contrast, six peptides representing 20% of the sequence may be adequate for a tentative identification. At this level, there is no definitive rule so users should err on the side of caution, and bearing in mind the ever-increasing size of databases and the risk of ambiguous or incorrect results, it should be considered essential to obtain corroborating data from at least one CID spectrum. For protein identifications based on CID data, it is important to appreciate that the first step requires correct peptide identifications. Only if this is achieved can protein identifications have any relevance. Ideally, an examination of a peptide match should reveal that the fragment peaks matched tend to be the stronger ones within the CID spectrum, preferably distributed throughout the m/z range rather than all coming from the low-mass ions. Hopefully, there will be some chemical logic to the spectrum and the fragmentation, e.g. histidine represents another site for protonation and gives higher charge states in ESI, C-terminal arginine favors y rather than b ion formation, cleavage N-terminal to a proline is favored, etc. Then it must be appreciated that the protein is a gene product that may represent one of a number of forms. If it is truncated, this will be revealed only if the appropriate terminus is identified in a peptide, and then it is important to consider that this peptide could correspond to a nonspecific cleavage. Alternative splice variants will only be established unambiguously if an amino acid sequence corresponding to the actual splice point is observed in a peptide.

No structural or sequence inferences can be based on the failure to observe a specific peptide; such a peptide may arise from proteolysis but simply not be detected, it may have an unexpected mass due to one or more modifications, or the residues may indeed be present in the sequence but the proteolysis may not proceed as anticipated. It is equally important to note that the absence of evidence for the presence of a particular protein does not definitively establish the absence of the protein. Even though a protein may be found in one sample but not another, it is not correct to describe this as being present in sample A but not sample B. This is something that simply cannot be proved by MS.

Then there is the question of supporting information. Until recently, most mass spectrometric experiments were directed toward identifying small numbers of proteins. In such cases, it would be conventional to confirm any identification by obtaining an appropriate antibody and running a Western blot. Finally, the creation of a knockout mouse might be necessary to ascertain whether there was any biological relevance, which for the identification of neural cell adhesion molecule 1 as a binding partner of the prion protein proved not to be the case (52). Such confirmations are impossible in high-throughput proteomic experiments, therefore researchers have a responsibility to either discard marginal results or to present the full details of the search output so that others can make their own evaluations.

**PROTEIN QUANTITATION**

Although many experiments in proteomics only require that proteins be identified, quantitation can also be important, particularly for differential analysis, such as in comparisons of samples from diseased and healthy states. Prior to the widespread adoption of MS, comparison of the intensity of staining of spots on two-dimensional gels by densitometry was a common approach to this problem. The use of gels continued with mass spectrometric analysis and was extensively automated by some workers (37). Nevertheless, this approach to quantitation may suffer from limited dynamic range for staining, uncertainties in protein identifications, and the danger that the darkening of a spot could be due to the presence of a new protein rather than up-regulation of the protein believed to be responsible. Although it is now commonplace to use mass spectrometric analysis of spots that show differential staining, problems with the identification of low-level proteins in mixtures may persist. With differential gel electrophoresis (53), which is commercially available as the DIGE system (Amersham Biosciences, Piscataway, NJ), the use of spectrally resolvable fluorescent dyes may separate and quantitate up to three samples on a single two-dimensional gel. This employs software that automatically locates and analyzes protein spots, assigning statistical confidence to observed differences.

A different approach relies on the power of MS to separate and identify differentially isotopically labeled species. Some researchers have compared proteins from cells grown in different isotopic environments, e.g. incorporating $^{15}$N or $^{13}$C into one population of proteins, the peptides from that could be distinguished from analogs containing the lighter isotope $^{14}$N or $^{12}$C by their mass differences (54, 55). Another isotope used in the MS analysis of peptides for at least 20 years is $^{18}$O (56). Digestion of a protein in $^{18}$O-enriched water results in the
heavy isotope being incorporated at the C terminus of each new peptide formed, with the exception of the pre-existing protein C terminus. The use of a 1:1 mixture of heavy and light water results in mass spectral doublets for all CID fragments containing a peptide C terminus, e.g. y ions, allowing them to be distinguished from b ions and other N-terminal fragments. Comparative proteomics using digestion in $^{18}$O-water was given a new twist with a method described as “inverse labeling.” By carrying out two parallel but isotopically reversed experiments then subtracting one spectrum from the other, only peptides derived from proteins differing in abundance remained to be analyzed (57). A complicating factor is further exchange of isotopically labeled oxygen catalyzed by the protease that can lead to the presence of two $^{18}$O per carboxyl group, rather than one. However, this can be decoupled from proteolysis (58) and is advantageous as it gives a mass difference of 4 Da rather than 2, thereby ensuring more effective separation of the natural isotopic clusters.

Another method to introduce stable isotopes is to differentially label a specific amino acid such as cysteine with a chemical reagent containing light or heavy isotopes. Such reagents may incorporate an isotope-coded affinity tag (ICAT) to enable enrichment of the labeled peptides, as well as to permit measurement of relative abundance (59). The strength of the isotope labeling strategy is that it allows specific markers to be incorporated into two populations of proteins that can then be mixed together before any extraction, digestion, peptide separation, and analysis is undertaken. As long as peptides incorporating the different isotopes have virtually identical physical and chemical properties, the ratios of peak heights or areas from MS experiments will be proportional to the relative amounts of each protein. In practice some isotopes are better than others, e.g. differences in polarity between hydrogen and deuterium affect the retention times in HPLC whereas the incorporation of $^{12}$C/$^{13}$C labels in peptide between hydrogen and deuterium affect the retention times in the relative amounts of each protein. In practice some iso-


d…

Note that the methods described above yield only “relative” quantities, whereas “absolute” protein quantitation requires a technique such as stable isotope dilution in which known amounts of isotopically labeled peptides are added to the digest (63). Such peptides are designed to mimic the peptides formed by proteolysis of the proteins of interest and can carry chemical modifications to simulate modified proteins such as phosphoproteins (64).

CHEMICAL AND POSTTRANSLATIONAL MODIFICATIONS

MS is ideal for the study of covalent modifications as all such modifications involve a change in molecular mass, which is reflected in the mass of any peptide carrying the modified amino acid(s). Sometimes this is uniquely attributable to the chemical transition, such as -2 Da for disulfide bond formation or +42 Da for acetylation. Some changes are easily identified, such as +16 Da for the addition of oxygen, but it may be unclear whether this is the common formation of methionine sulfoxide or something more unusual. It is also uncertain whether this is a posttranslational modification or a chemical change occurring during sample processing. Other mass changes may be more ambiguous, e.g. +80 Da could represent phosphorylation or sulfation, although these may be distinguished by the detection of an anion at m/z 79 versus 80, respectively, and a very accurate mass measurement. Specialized MS/MS techniques have been developed for the identification of the precursor ions for peptides that carry specific modifications such as phosphorylation or glycosylation (65–67). Some modifications may be quite heterogeneous, as with N-glycosylation, although enzymatic removal of the modified group can give a more defined change, e.g. peptide N-glyco-
sidase F removes N-linked sugars and converts asparagine to aspartic acid (+1 Da). The major difficulties arise with the search for transient modifications of low stoichiometry, such as phosphorylation occurring in a regulatory role. Affinity methods have been employed to enrich phosphopeptides such as immobilized metal ion chromatography but with mixed success (68, 69). The difficulties are exacerbated when the modification is labile, either during chemical separation or in the mass spectrometer. Thus CID and MS/MS of the modified peptide should potentially reveal the site of modification, but a facile loss of phosphoric acid from the molecular ion may prevent the detection of any sequence-specific ions. In general, peptide mass mapping alone cannot be used for identification of posttranslational modifications. Based on MS data alone, a search can be made for phosphopeptide mass values that are 80 Da higher in mass than calculated for the amino acid sequence (or 160 Da, etc.), but MS/MS to confirm the sequence is essential, even if the precise site of modifi-
cation cannot be defined. Other more labile modifications requiring special techniques for detection are sulfation (70) and O-GlcNAc (71, 72).

CONCLUSIONS

It is clear that the field of proteomic analysis by MS con-tinues to be in a very dynamic state, making it presently difficult
to specify absolute standards for analytical protocols and subsequent interpretative judgments. Nevertheless, authors submitting MS-based manuscripts to Molecular & Cellular Proteomics must employ methodology that meets currently acceptable standards. Thus, protein identification based solely on peptide mass fingerprinting should no longer be acceptable and must be complemented by CID and MS/MS. Clearly, it is essential that researchers avoid the “black box” mentality and take responsibility for understanding the methods that they are using. Furthermore, it is strongly recommended that authors take the risk of presenting too much information in describing their experiments, their data, and its interpretation. It will then be incumbent on journal editors and reviewers to determine what fraction of this information is relevant and is required for publication. Authors should report the software they use, including the version, the database relevant and is required for publication. Authors should report

References

Issues in Protein Identification by Mass Spectrometry


