

1

Mass Spectrometry of Amino Acids and Proteins

Simin D. Maleknia and Richard Johnson

1.1

Introduction

1.1.1

Mass Terminology

Like most matter (with the exception of, say, neutron stars), proteins and peptides are mostly made of nothing – an ephemeral cloud of electrons with very little mass surrounding tiny and very dense atomic nuclei that contain nearly all of the mass (i.e., peptides and proteins are made of atoms). Atoms have mass and the unit of mass that is most convenient to use is called the atomic mass unit (abbreviated amu or u) or in biological circles a Dalton (Da). Over the years, physicists and chemists have argued about what standard to use to define an atomic mass unit, but the issue seems to have been settled in 1959 when the General Assembly of the International Union of Pure and Applied Chemistry defined an atomic mass unit as being exactly 1/12 of the mass of the most abundant carbon isotope (^{12}C) in its unbound lowest energy state. Therefore, one atom of ^{12}C has a mass of 12.0000 u. Using this as the standard, one proton has a measured mass of 1.00728 u and one neutron is slightly heavier at 1.00866 u. One ^{12}C atom contains six protons and six neutrons, the sum of which is clearly more than the mass of 12.0000 u. A carbon atom is less than the sum of its parts, and the reason is that the protons and neutrons in a carbon nucleus are in a lower energy state than free protons and neutrons. Energy and mass are interchangeable via Einstein's famous equation ($E = mc^2$), and so this "mass defect" is a result of the nuclear forces that hold neutrons and protons together within an atom. This mass defect also serves as a reminder of why people like A. Q. Khan are so dangerous [1].

Each element is defined by the number of protons per nucleus (e.g., carbon atoms always have six protons), but each element can have variable numbers of neutrons. Elements with differing numbers of neutrons are called isotopes and each isotope possesses a different mass. In some cases, the additional neutrons result in stable isotopes, which are particularly useful in mass spectrometry (MS) in a method called

isotope dilution. Examples in the proteomic field that employ isotope dilution methodology include the use of the stable isotopes ^2H , ^{13}C , ^{15}N , and ^{18}O , as applied in methods such as ICAT (isotope-coded affinity tags) [2], SILAC (stable isotope labeling with amino acids in cell culture) [3], or enzymatic incorporation of ^{18}O water [4]. Whereas some isotopes are stable, others are not and will undergo radioactive decay. For example, hydrogen with one neutron is stable (deuterium), but if there are two additional neutrons (a tritium atom) the atoms will decay to helium (two protons and one neutron) plus a negatively charged β -particle and a neutrino. Generally, if there are sufficient amounts of a radioactive isotope to produce an abundant mass spectral signal, the sample is likely to be exceedingly radioactive, the instrumentation would have become contaminated, and the operator would likely come to regret having performed the analysis. Therefore, mass spectrometrists will typically concern themselves with stable isotopes. Each element has a different propensity to take on different numbers of neutrons. For example, fluorine has nine protons and always 10 neutrons; however, bromine with 35 protons is evenly split between possessing either 44 or 46 neutrons. There are most likely interesting reasons for this, but they are not particularly relevant to a description of the use of MS in the analysis of proteins.

What is relevant is the notion of “monoisotopic” versus “average” versus “nominal” mass. The monoisotopic mass of a molecule is calculated using the masses of the most abundant isotope of each element present in the molecule. For peptides, this means using the specific masses for the isotopes of each element that possess the highest natural abundance (e.g., ^1H , ^{12}C , ^{14}N , ^{16}O , ^{31}P , and ^{32}S as shown in Table 1.1). The “average” or “chemical” mass is calculated using an average of the isotopes for each element, weighted for natural abundance. For elements found in most biological molecules, the most abundant isotope contains the fewest neutrons

Table 1.1 Mass and abundance values for some biochemically relevant elements.

Element	Average mass	Isotope	Monoisotopic mass	Abundance (%)
Hydrogen	1.008	^1H	1.00783	99.985
		^2H	2.01410	0.015
Carbon	12.011	^{12}C	12	98.90
		^{13}C	13.00335	1.10
Nitrogen	14.007	^{14}N	14.00307	99.63
		^{15}N	15.00011	0.37
Oxygen	15.999	^{16}O	15.99491	99.76
		^{17}O	16.99913	0.04
		^{18}O	17.99916	0.200
Phosphorus	30.974	^{31}P	30.97376	100
Sodium	22.990	^{23}Na	22.98977	100
Sulfur	32.064	^{32}S	31.97207	95.02
		^{33}S	32.97146	0.75
		^{34}S	33.96787	4.21
		^{36}S	35.96708	0.02

and the less abundant isotopes are of greater mass. Therefore, the monoisotopic masses calculated for peptides are less than what are calculated using average elemental masses. The term “nominal mass” refers to the integer value of the most abundant isotope for each element. For example, the nominal masses of H, C, N, and O are 1, 12, 14, and 16, respectively. A rough conversion between nominal and monoisotopic peptide masses is shown as [5]:

$$M_c = 1.000495 \cdot M_n \quad (1.1)$$

$$D_m = 0.03 + 0.02 \cdot M_n/1000 \quad (1.2)$$

where M_c is the estimated monoisotopic peptide mass calculated from a nominal mass, M_n . D_m is the estimated standard deviation at a given nominal mass. For example, peptides with a nominal mass of 1999 would be expected, on average, to have a monoisotopic mass of around 1999.99 with a standard deviation of 0.07 u. Therefore, 99.7% of all peptides (3 standard deviations) at a nominal mass 1999 would be found at monoisotopic masses between 1999.78 and 2000.20 (Figure 1.1).

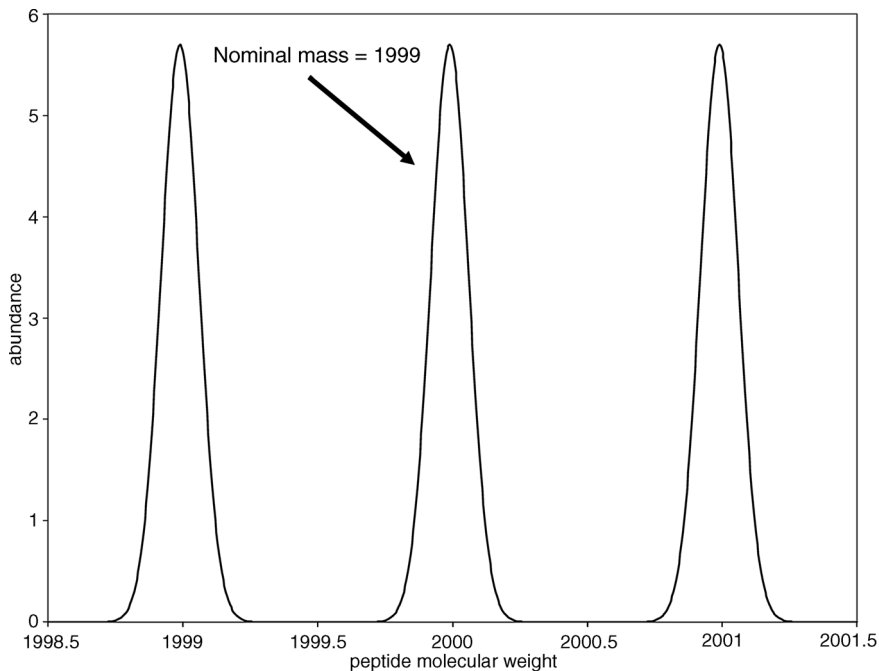


Figure 1.1 Predicting monoisotopic from nominal molecular weights. Using the equations from Wool and Smilansky [5], peptides with nominal molecular weights of 1998, 1999, and 2000 would on average be expected to have monoisotopic molecular

weights of 1998.99, 1999.99, and 2000.99 with standard deviations of 0.07. The difference between monoisotopic and nominal masses is called the mass defect and this value scales with mass.

As can be seen, at around mass 2000, the mass defect in a peptide molecule is just about one whole mass unit. Most of this mass defect is due to the large number of hydrogen atoms present in a peptide of this size. The mass defect associated with nitrogen and oxygen tends to cancel out, and carbon by definition has no mass defect. The other important observation that can be made from this example is that 99.7% of peptides with a nominal mass of 1999 will be found between 1999.78 and 2000.20. Therefore, a molecule that is accurately measured to be 2000.45 cannot be a standard peptide and must either not be a peptide at all or is a peptide that has been modified with elements not typically found in peptides.

1.1.2

Components of a Mass Spectrometer

At minimum, a mass spectrometer has an ionization source, a mass analyzer, an ion detector, and some means of reporting the data. For the purposes here, there is no need to go into any detail at all regarding the ion detection and although there are many historically interesting methods of recording and reporting data (photographic plates, UV-sensitive paper, etc.), nowadays one simply uses a computer. The ionization source and the mass analyzer are the two components that need to be well understood.

Historically, ionization was limited to volatile molecules that were amenable to gas phase ionization methods such as electron impact. Over time, other techniques were developed that allowed for ionization of larger polar molecules – techniques such as fast atom bombardment (FAB) or field desorption ionization. However, these had relatively poor sensitivity requiring 0.1–1 nmol of peptide and, with the exception of plasma desorption ionization – a technique that used toxic radioactive californium, were generally not capable of ionizing larger molecules like proteins. Remarkably, two different ionization methods were developed in the late 1980s that did allow for sensitive ionization of larger molecules – electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI). Posters presented at the 1988 American Society for Mass Spectrometry conference by John Fenn's group showed mass spectra of several proteins [6, 7], which revealed the general nature of ESI of peptides and proteins. Namely, a series of heterogeneous multiply protonated ions are observed, where the maximum number of charges is roughly dependent on the number of basic sites in the protein or peptide. Conveniently, this puts the ions at mass-to-charge (m/z) ratios typically below 4000, which is a range suitable for just about all mass analyzers (see below). In a series of papers between 1985 and 1988, Hillenkamp and Karas described the essentials of MALDI [8–10]. Also, Tanaka presented a poster at a Joint Japan–China Symposium on Mass Spectrometry in 1987 showing a pentamer of lysozyme using laser desorption from a glycerol matrix containing metal shavings [11]. These early results showed the general nature of MALDI – singly charged ions predominate and therefore the mass analyzer must be capable of measuring ions with very high m/z ratios.

It is desirable for users to have some basic understanding of the different types of mass analyzers that are available. At one time multiselector analyzers [12] were well-liked (back when FAB ionization was popular), but quickly became dinosaurs for protein work after the discovery of ESI. It was too difficult to deal with the electrical arcs that tended to arise when trying to couple kiloelectronvolt source voltages with a wet acidic atmospheric spray. ESI was initially most readily coupled to quadrupole mass filters, which operated at much lower voltages. Quadrupole mass filters [13], as the name implies, are made from four parallel rods where at appropriate frequency and voltages, ions at specific masses can oscillate without running into a rod or escaping from between the rods. Given a little push (a few electronvolts potential) the oscillating ions will pass through the length of the parallel rods and be detected at the other end. Both quadrupole mass filters and multiselector instruments suffer from slow scan rates and poor sensitivity due to their low duty cycle. Instrument vendors have therefore been busy developing more sensitive analyzers. The ion traps [14, 15] are largely governed by the same equations for ion motion as quadrupole mass filters, but possess a greater duty cycle (and sensitivity). For those unafraid of powerful super cooled magnets, and who possess sufficiently deep pockets to pay for the initial outlay and subsequent liquid helium consumption, Fourier transform ion cyclotron resonance (FT-ICR) provides a high-mass-accuracy and high-resolution mass analyzer [16]. In this case, the ions circle within a very high vacuum cell under the influence of a strong magnetic field. The oscillating ions induce a current in a pair of detecting electrodes, where the frequency of oscillation is related to the m/z ratio. Detection of an oscillating current is also performed in Orbitrap instruments [17, 18], except in this case the ions circle around a spindle-shaped electrode rather than magnetic field lines. The time-of-flight hybrid (TOF) mass analyzer [19, 20] is, at least in principle, the simplest analyzer of all – it is an empty tube. Ions are accelerated down the empty tube and, as the name implies, the TOF is measured and is related to the m/z ratio (big ions move slowly and little ones move fast).

Tandem MS is a concept that is independent of the specific type of mass analyzer, but should be understood when discussing mass analyzers. As the name implies, tandem MS employs two stages of mass analysis, where the two analyzers can be scanned in various ways depending on the experiment. In the most common type of experiment, the first analyzer is statically passing an ion of a specific mass into a fragmentation region, where the selected ions are fragmented somehow (see below) and the resulting fragment ions are mass analyzed by the second mass analyzer. These so-called daughter, or product, ion scans are usually what are meant when referring to an “MS/MS spectra.” However, there are other types of tandem MS experiments that are occasionally performed. One is where the first mass analyzer is statically passing a precursor ion (as in the aforementioned product ion scan) and the second analyzer is also statically monitoring one, or a few, specific fragment ions. This so-called selected reaction monitoring (SRM) experiment is particularly useful in the quantitation of known molecules. There are other less frequently used tandem MS scans (e.g., neutral loss scans) and it should be noted

that only certain combinations of specific analyzers are capable of performing certain kinds of scans.

There are various combinations of mass analyzers used in different mass spectrometers. One of the more popular has been the quadrupole/TOF hybrid (quadrupole/time-of-flight hybrid Q-TOF) [21], which uses the quadrupole as a mass filter for precursor selection and the TOF is used to analyze the resulting fragment ions. Ion trap/time-of-flight hybrids are also sold and provide additional stages of tandem MS compared to the quadrupole/linear ion trap hybrid (Q-trap). The Q-TOF hybrid [22] is a unique instrument in that it can be thought of as a triple-quadrupole instrument where the third quadrupole can alternatively be used as a linear ion trap. There is consequently a great deal of flexibility in the types of experiments that can be done on such a mass spectrometer. The tandem TOF (TOF-TOF) [20] is an instrument that allows acquisition of tandem mass spectra or single-stage mass spectra of MALDI-generated ions. A timed electrode is used for precursor selection, which sweeps away all ions except those passing at a certain time (i.e., m/z) when the electrode is turned off momentarily. The selected packet of ions is then slowed down, possibly subjected to collision-induced dissociation (CID), and reaccelerated for the final TOF mass analysis of the fragments. The Orbitrap analyzer is purchased as a linear ion trap/Orbitrap hybrid and the same vendor sells their ion cyclotron resonance ICR instrument as a linear ion trap/ICR hybrid. It is beyond the scope of this chapter to go into any further details regarding the operation of the mass analyzers. Furthermore, it seems likely that the field will continue to change in the coming years, where instrument vendors will make further changes.

1.1.3

Resolution and Mass Accuracy

Regardless of the mass spectrometer, the user needs to understand their capabilities and limitations. Sensitivity has been a driving force for the development of many of the newer mass spectrometers. It is also a difficult parameter to evaluate, and one has to be careful not to simply evaluate the ability and tenacity of each vendor's application chemist when sending test samples out. Dynamic range is a parameter that is useful in the context of quantitative measurements and for most instruments it is around 10^4 . Some instruments can perform unique scan types (e.g., the Q-trap), or are more sensitive at performing SRM quantitative experiments (triple-quadrupole and Q-trap instruments). The scan speed or rate of MS/MS spectra acquisition is an instrument parameter that is relevant when attempting a deeper analysis of a complex mixture in a given amount of time. This latter issue is particularly important when analyzing complex proteomic samples.

Two analyzer-dependent parameters are particularly important – mass accuracy and resolution. Resolution is defined as a unit-less ratio of mass divided by the peak width and is typically measured halfway up the peak. Figure 1.2 shows the peak shapes calculated for the peptide glucagon at various resolution values. At this mass, a resolution of 10 000 is sufficient to provide baseline separation of

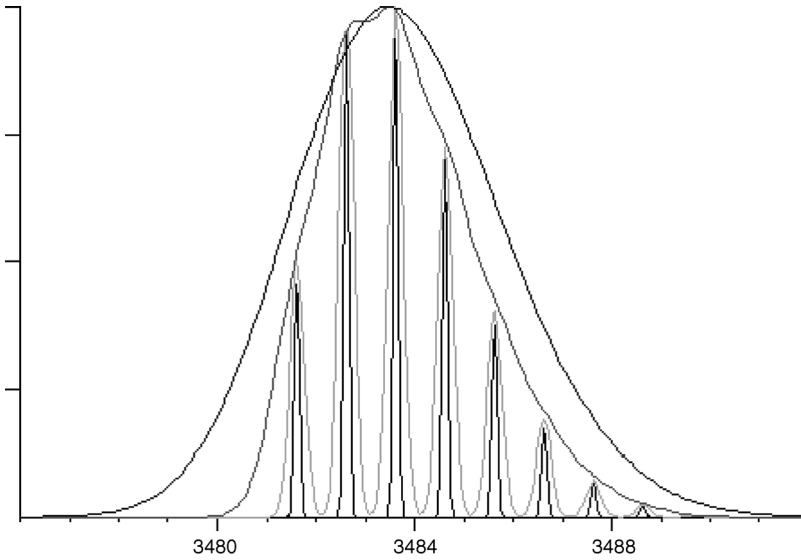


Figure 1.2 Effect of mass spectrometric resolution on peak shape. Shown are the calculated peak shapes for the $(M + H)^+$ ion of porcine glucagon (monoisotopic mass of 3481.62 Da and average mass of 3483.8 Da) at various resolution values: 30 000 (inner most narrow peaks), 10 000 (outer most broad peak), 3000 (outer most broad peak), and 1000 (outer most broad peak).

each isotope peak and the higher resolution of 30 000 results in the narrowing of each isotope peak. As the resolution drops below 10 000 the valley between each isotope becomes higher until at 3000 the isotope cluster becomes a single broad unresolved peak. As the resolution drops further (blue), the single broad peak gets even fatter. Resolution is important to the extent that one needs to know if it is sufficient to separate the isotope peaks of a particular sample. If not, then a centroid of a broad unresolved peak (e.g., 1000 or 3000 for glucagon) is going to be closest to the peptide mass calculated using average elemental mass. Alternatively, if the resolution is sufficient to resolve the isotope peaks, and it is possible for the data system to accurately and consistently identify the monoisotopic ^{12}C peak, then this observed peptide mass will be closest to that calculated using monoisotopic elemental masses.

Why do high resolution and high mass accuracy go hand in hand? One does not hear of low-resolution, high-mass-accuracy instruments, for instance. There are at least two reasons. First, it is not possible to determine a very accurate average elemental mass, which is weighted for isotope abundance. Chemical and physical fractionation processes occurring in nature result in variable amounts of each isotope in different samples. For example, the different photosynthetic processes (e.g., C3 and C4) will fractionate ^{13}C slightly differently. Hence, corn will tend to have a slightly higher percentage of ^{13}C than a tree. Therefore, in contrast to monoisotopic masses, average elemental masses come with fairly substantial error bars. The second reason

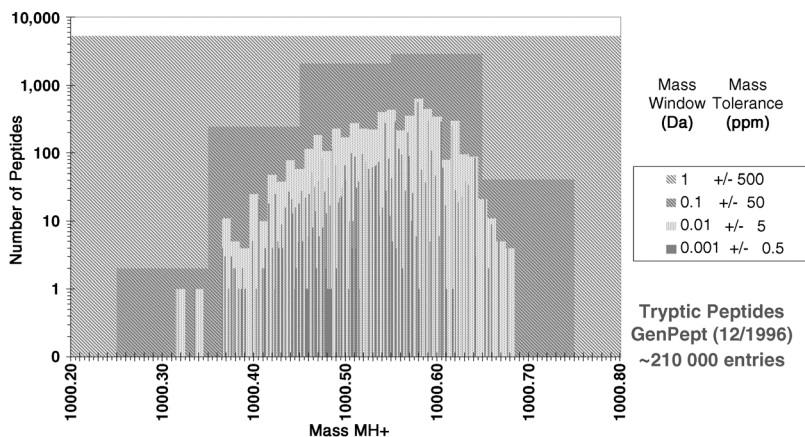


Figure 1.3 Role of high mass accuracy in reducing false-positives from database searches. This histogram (from [23]) shows the number of tryptic peptides at different mass accuracies for a 1996 GenPept database. For a nominal molecular weight of 1000, there are around 5000 tryptic peptides if the measurements are accurate to 0.5 Da

(500 ppm). If the mass measurements are accurate to 0.01 Da (5 ppm), which are routinely available for Orbitrap and certain Q-TOF instruments, the number of possible tryptic peptides in the database drops to one to a couple hundred, depending on the specific mass window.

why higher resolution usually results in higher mass accuracy is that as a mono-isotopically resolved peak becomes narrower, any slight variation in the peak position is also reduced. Due to factors such as overlapping peaks and ion statistics, it is not possible to consistently and accurately measure a much wider unresolved isotope cluster at low resolution. Hence, the type of mass analyzer will determine the resolution and mass accuracy.

There are three types of resolution (and mass accuracy) for tandem MS that are associated with the precursor ion, precursor selection, and fragment ions. The precursor and fragment ion resolution and accuracy may be identical (e.g., for Q-TOF or ion traps) or different (e.g., for ion trap-FT-MS hybrids or TOF-TOF). The importance of being able to more accurately determine peptide masses was clearly demonstrated by Clauser *et al.* [23] as shown in Figure 1.3, which depicts a histogram of the number of tryptic peptides at different mass accuracies. For a 1996 GenPept database, there are around 5000 tryptic peptides at a nominal mass of 1000 with a tolerance of ± 0.5 Da (500 ppm). However, if the tolerance is tightened up to ± 0.05 Da, then the number of tryptic peptides drops by an amount that is dependent on the mass. There are fewer peptides at either the low- or high-mass end of the histogram, such that there are only two tryptic peptides in the database at a measurement of 1000.3 ± 0.05 Da. Likewise, there are only 30–40 peptides with a mass of 1000.7 ± 0.05 Da. Most of the tryptic peptides at a nominal mass of 1000 are in the range of 1000.45–1000.65, so a tolerance of ± 0.05 Da in the middle of this histogram will reduce the number of possible tryptic peptides from 5000 to 2000–3000. When using a database search program that identifies peptides from

their MS/MS spectra, a tighter precursor mass tolerance will result in fewer candidate sequences, which has the desirable effect of reducing the chances of an incorrect identification.

Database search programs (e.g., Mascot [24] or SEQUEST [25]) assume that there is only a single precursor and that all of the fragment ions are derived from that one precursor ion. For more complicated samples it is quite possible that more than one precursor is selected at a time and the likelihood of this happening is dependent on the precursor selection resolution. Typical ion traps select the precursor using a window that is three or four m/z units wide, Q-TOFs are similar, and TOF-TOFs have a precursor resolution of around 400 (e.g., at m/z 1000, any peak at 997.5 will have its transmission reduced by half). The shape of this precursor selection window is also important – a sharp cutoff to zero transmission is good and a slow taper is not. Sometimes an extraneous low-intensity precursor is not a problem, as long as most of the fragment ion intensity is associated with the major precursor and the precursor mass that is associated with the resulting MS/MS spectrum is from the correct precursor ion. Search programs will still identify the major peptide, since there will only be a few low-intensity fragment ions left over. However, one can readily imagine several scenarios where mass selection of multiple precursors would be a problem. For example, suppose a minor precursor fragments really well, but the major precursor does not. In this case, the MS/MS spectrum contains fragment ions from the minor precursor, but the precursor mass that is used in the database search is derived from the major one. Or, a low-intensity precursor triggers a data-dependent MS/MS acquisition, but another very intense ion that is a few m/z units away contributes much of the fragment ion intensity. In such instances, where the fragment ions are derived from more than one precursor, search programs may get the wrong answer because the wrong precursor mass was used or there are too many leftover fragment ions and the scoring algorithm penalizes one of the correct sequences. Tighter selection windows with abrupt cutoffs (high precursor selection resolution) reduce the likelihood of this occurring. Improved database search algorithms would also help.

One of the major challenges in proteomics is high-throughput analysis. The high resolving power of FT-ICR instruments offers less than 1 ppm mass measurement accuracy and the peptide identification protocol of accurate mass tags (AMTs) now affords protein identification without the need for tandem MS/MS. Combining the AMT information with high-performance liquid chromatography (HPLC) elution times and MS/MS is referred to as peptide potential mass and time tags (PMTs) [26]. This approach expedites the analysis of samples from the same proteome through shotgun proteomics – a method of identifying proteins in complex mixtures by combining HPLC and MS/MS [27]. Once a peptide has been correctly identified through AMT and MS/MS with an assigned PMT, the information is stored in a database. This strategy greatly increases analysis throughput by eliminating the need for time-consuming MS/MS analyses. Accurate mass measurements are now routinely practiced in applications involving organisms with limited proteomes, including proteotyping the influenza virus [28], and the rapid differentiation of seasonal and pandemic stains [29].

1.1.4

Accurate Analysis of ESI Multiply Charged Ions

It is important to briefly describe the deconvolution algorithms used to translate m/z ratios of multiply charged ions generated during ESI to zero-charge molecular mass values. The accurate assignment of multiply charged ions is significant in proteomics applications, both in the analysis of the intact proteins and for the identification of fragment ions by MS/MS. For low-resolution mass spectra, algorithms were originally developed by assuming the nature of charge-carrying species or considering only a limited set of charge carrying species (i.e., proton, sodium) [30, 31]. For two ions (m_a/z_a and m_b/z_b) that differ by one charge unit and both contain the same charge-carrying species, the charge on ion a (z_a) is given by Eq. (1.3), where m_p is the mass of a proton, and the molecular weight is derived from Eq. (1.4):

$$z_a = (m_b/z_b - m_p) / (m_b/z_b - m_a/z_a) \quad (1.3)$$

$$\text{molecular weight} = z_a(m_a/z_a) - z_a m_p \quad (1.4)$$

The advantage of high-resolution electrospray mass spectra is that the ion charge can be derived directly from the reciprocal of the mass-to-charge separation between adjacent isotopic peaks ($1/\Delta m/z$) for any multiply charged ion – referred to as the isotope spacing method [32]. Although the isotope spacing method is direct, complexities arising from spectral noise and overlapping peaks may result in inaccurate ion charge determination; furthermore, distinguishing $1/z$ and $1/(z + 1)$ for high charge state ions ($z > 10$), would require mass accuracies of a few parts per million, which is not achieved routinely. To overcome some of these limitations, algorithms of Zscore [33] and THRASH [34] combined pattern recognition techniques to the isotope spacing method. For example, the THRASH algorithm matches the experimental abundances with theoretical isotopic distributions based on the model amino acid “averagine” ($C_{4.938} H_{7.7583} N_{1.3577} O_{1.4773} S_{0.0417}$) [35]; however, this requirement restricts its application to a specific group of compounds and elemental compositions (i.e., proteins). The AID-MS [36] and PTFT [37] algorithms further advanced the latter algorithms by incorporating peak-finding routines to locate possible isotopic clusters and to overcome the problems associated with overlapping peaks.

A unique algorithm, CRAM (charge ratio analysis method) [38–40], deconvolutes electrospray mass spectra solely from the m/z values of multiply charged ions. The algorithm first determines the ion charge by correlating the ratio of m/z values for any two (i.e., consecutive or nonconsecutive) multiply charged ions to the unique ratios of two integers. The mass, and subsequently the identity of the charge carrying species, is then determined from m/z values and charge states of any two ions. For the analysis of high-resolution electrospray mass spectra, CRAM correlates isotopic peaks that share the same isotopic compositions. This process is also performed through the CRAM process after correcting the multiply charged ions to their lowest common ion charge. CRAM does not require prior knowledge of the elemental composition of a

molecule and as such does not rely at all on correlating experimental isotopic patterns with the theoretical patterns (i.e., known compositions), and therefore CRAM could be applied to mass spectral data for a range of compounds (i.e., including unspecified compositions).

1.1.5

Fragment Ions

Although a considerable amount of work has been done in order to understand fragmentations of negatively charged peptide ions [41], the majority of protein identification work has employed positively charged peptide ions [42]. This is partially due to a general fear and ignorance of negatively charged peptides, but mostly because peptide signals are typically more abundant in the positive ion mode and the fragment ions are more likely to delineate a large portion of the peptide sequence. The following discussion is centered on fragmentation of peptide cations.

Depending on the type of mass spectrometer used, one can expect to generate fragment ions from three different processes – low-energy CID, high-energy CID, and electron capture (or transfer) dissociation (electron capture dissociation ECD or electron transfer dissociation ETD). Low-energy CID is the most common means of fragmenting peptide ions and occurs when the precursor ions collide with neutral collision gas with kinetic energies less than 500–1000 eV. This is the situation for any instrument with a quadrupole collision cell (triple-quadrupole, Q-trap, or Q-TOF), or any ion trap, including ion trap hybrids. A different process known as postsource decay (PSD) occurs in MALDI-TOF and MALDI-TOF-TOF instruments (when operated without collision gas). In PSD, precursor ions resulting from the MALDI process are sufficiently stable to stay intact during the initial acceleration into the flight tube, but they then fall apart in transit through the flight tube after full acceleration. These PSD-derived ions are largely identical to what is produced by low-energy CID. Figure 1.4(a) shows the peptide fragmentation nomenclature originally devised by Roepstorff and Fohlman [43], where the three possible bonds in a residue of a peptide are cleaved and the resulting fragment ion designated as X, Y, or Z (charge retained on the C-terminal fragments), or A, B, or C (N-terminal fragments). In addition to cleavage of the bond, different fragment ions also have variable numbers of hydrogen atoms and protons transferred to them. For a time there was considerable discussion as to whether the hydrogen transfer should be designated by tick marks (e.g., Y' for two hydrogen atoms transferred to a Y cleavage ion) or by “+2” (e.g., Y + 2) designations. Biemann [44] subsequently proposed a similar designation whereby the letters went to lower case and the proper number of hydrogen atom transfers was assumed, without ticks or anything else. These are high stake issues, since adopting a specific nomenclature could dramatically increase one's citation index.

At the most simplistic level, low-energy CID and PSD produce *b* and *y* ions. The structures shown in Figure 1.4(b) are not strictly accurate, but they illustrate how to go about calculating the masses of any fragment ion. The concept of a “residue mass” is

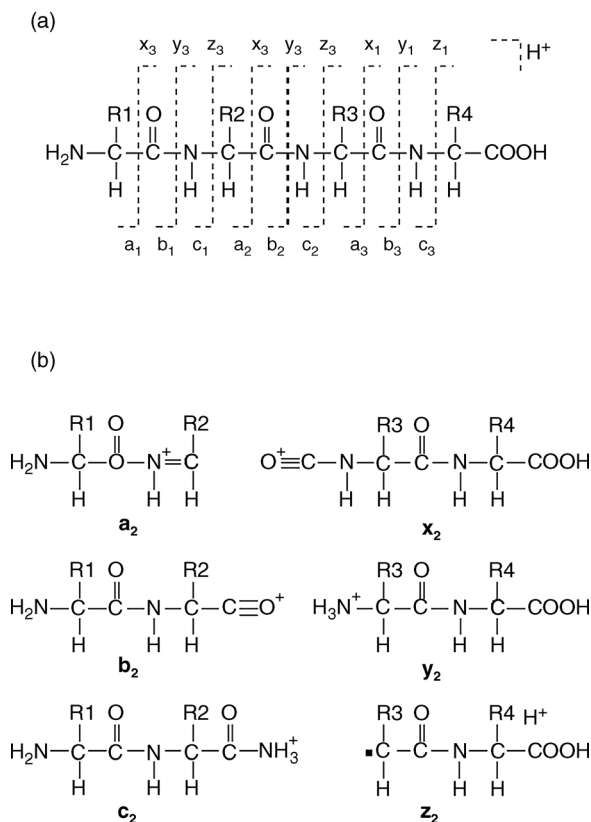
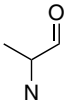
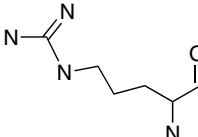
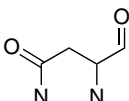
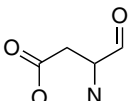
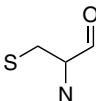
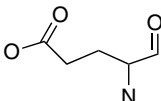
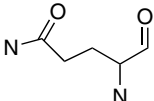


Figure 1.4 Nomenclature for positive ion peptide fragments. Roepstorff nomenclature [43] is shown in (a). X, Y, and Z denote C-terminal fragments and A, B, and C denote N-terminal fragments. Fragment ions also have variable numbers of hydrogen atoms and protons transferred to them, as shown in

(b), which uses the Biemann nomenclature [44]. Low-energy CID of peptides in positive mode generally produces *b*-type and *y*-type ions. ETD and ECD generally produce *c*-type and *z*-type ions. The *z*-type ions are odd-electron radical cations, whereas the others are all even-electron cations.

that this is the mass of an amino acid within a peptide (i.e., it is the mass of an amino acid minus the mass of water, which is lost when amino acids polymerize to form peptides). Table 1.2 gives the average and monoisotopic residue masses for the common amino acids. It can be seen from Figure 1.4 that a *b* ion would be calculated by summing the residue masses and adding the mass of a single hydrogen atom (assuming that the peptide has an unmodified N-terminus). Likewise, a *y* ion would be calculated by summing the appropriate residue masses and then adding the mass of water plus a proton. The formulae for calculating the various peptide fragment ions are summarized in Table 1.3. It is believed that the actual structure of a *y* ion is the same as a protonated peptide and what is shown in Figure 1.4(b) is probably an

Table 1.2 Amino acid residue masses.

Residue	Three-letter code	One-letter code	Monoisotopic mass	Average mass	Structure
Alanine C ₃ H ₅ NO	Ala	A	71.03712	71.08	
Arginine C ₆ H ₁₂ N ₄ O	Arg	R	156.10112	156.19	
Asparagine C ₄ H ₆ N ₂ O ₂	Asn	N	114.04293	114.10	
Aspartic acid C ₄ H ₅ NO ₃	Asp	D	115.02695	115.09	
Asn or Asp	Asx	B			
Cysteine C ₃ H ₅ NOS	Cys	C	103.00919	103.14	
Glutamic acid C ₅ H ₇ NO ₃	Glu	E	129.04260	129.12	
Glutamine C ₅ H ₈ N ₂ O ₂	Gln	Q	128.05858	128.13	
Glu or Gln	Glx	Z			

(Continued)

Table 1.2 (Continued)

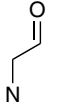
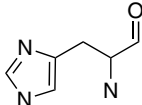
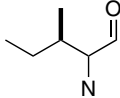
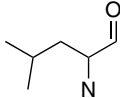
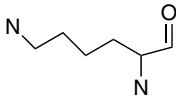
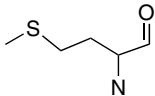
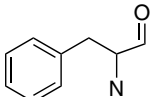

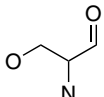
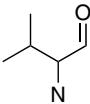
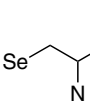
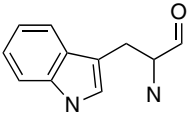
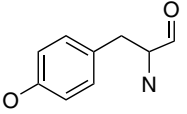
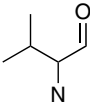
Residue	Three-letter code	One-letter code	Monoisotopic mass	Average mass	Structure
Glycine C ₂ H ₃ NO	Gly	G	57.02147	57.05	
Histidine C ₆ H ₇ N ₃ O	His	H	137.05891	137.14	
Isoleucine C ₆ H ₁₁ NO	Ile	I	113.08407	113.16	
Leucine C ₆ H ₁₁ NO	Leu	L	113.08407	113.16	
Lysine C ₆ H ₁₂ N ₂ O	Lys	K	128.09497	128.17	
Methionine C ₅ H ₉ NOS	Met	M	131.04049	131.19	
Phenylalanine C ₉ H ₉ NO	Phe	F	147.06842	147.18	
Proline C ₅ H ₇ NO	Pro	P	97.05277	97.12	
Serine C ₃ H ₅ NO ₂	Ser	S	87.03203	87.08	

Table 1.2 (Continued)

Residue	Three-letter code	One-letter code	Monoisotopic mass	Average mass	Structure
Threonine C ₄ H ₇ NO ₂	Thr	T	101.04768	101.10	
Selenocysteine C ₃ H ₅ NOSe	SeC	U	150.95364	150.03	
Tryptophan C ₁₁ H ₁₀ N ₂ O	Trp	W	186.07932	186.21	
Tyrosine C ₉ H ₉ NO ₂	Tyr	Y	163.06333	163.18	
Unknown	Xaa	X			
Valine C ₅ H ₉ NO	Val	V	99.06842	99.13	

accurate depiction of that type of fragment ion, although the site of protonation will vary. In contrast, the *b* ion structure in Figure 1.4(b) is almost certainly incorrect and instead is probably a five-membered ring structure [45]. The mechanism of formation of *b*-type ions most likely involves the carbonyl oxygen of the residue N-terminal to the cleavage site, which explains why one never observes *b*₁ ions in peptides with free N-termini. Acylated peptides will produce *b*₁ ions, since there is an N-terminal carbonyl available to induce the cleavage reaction.

The concept of a “mobile proton” provides a useful framework for understanding the low-energy CID peptide fragmentation process [46]. In solution, the sites of peptide protonation are likely to be the N-terminal amino group, the lysine amino group, the histidine imidazole side-chain, or the guanidino group on arginine. In the gas phase, however, the peptide backbone amides are of comparable basicity to all but

Table 1.3 Calculating the masses of positively charged fragment ions.

Ion type	Neutral molecular weight of the fragment
<i>a</i>	$[N] + [M] - CO - H$
<i>a</i> -H ₂ O	$a - 18.0106$
<i>a</i> -NH ₃	$a - 17.0266$
<i>b</i>	$[N] + [M] - H$
<i>b</i> -H ₂ O	$b - 18.0106$
<i>b</i> -NH ₃	$b - 17.0266$
<i>c</i>	$[N] + [M] + NH_2$
<i>d</i>	<i>a</i> - partial side-chain
<i>x</i>	$[C] + [M] + CO - H$
γ	$[C] + [M] + H$
γ -H ₂ O	$\gamma - 18.0106$
γ -NH ₃	$\gamma - 17.0266$
<i>z</i>	$[C] + [M] - NH$
<i>v</i>	γ - complete side-chain
<i>w</i>	<i>z</i> - partial side-chain

[N] is the mass of the N-terminus (e.g., 1.0078 Da for unmodified peptides and 43.0184 Da for acetylated N-terminus). [C] is the mass of the C-terminus (e.g., 17.0027 Da for unmodified peptides and 16.0187 Da for amidated C-terminus). [M] is the sum of the amino acid residue masses (see Table 1.1) that are contained within the fragment ion. CO is the combined mass of oxygen plus carbon atoms (27.9949 Da) and H is the mass of a proton (1.0078 Da). To calculate the *m/z* value of a fragment ion, add the mass of the protons to the neutral mass calculated from the table and divide by the number of protons added.

the arginine guanidino group. Therefore, in the absence of arginine, it takes only a little bit of collisional energy to scramble the site of protonation such that the ionized peptide is actually a population of ions that differ in the site of protonation (e.g., protonation occurring at any of the backbone amides or the side-chains). Protonation of the backbone amide is required for the production of *b*- or γ -type fragment ions and such cleavages that require protonation are called “charge promoted” fragmentations. Hence, as long as there is a mobile proton that can be sprinkled across the peptide backbone, one can expect to see a fairly contiguous series of *b*- and/or γ -type ions (e.g., Figure 1.5a). A major snag in this simplified view of low-energy CID of peptides is that the arginine guanidino group has such high gas-phase basicity that it essentially immobilizes a single proton. If there are at least as many arginine residues as protons, then to create *b*- or γ -type fragments, additional energy is required to “mobilize” one of the protons that would otherwise prefer to be stuck to the guanidino group. This additional energy will also result in the production of new and undesirable fragment ion types, such that the resulting spectra no longer possess the anticipated contiguous *b*- and γ -type fragment ion series (Figure 1.5b). One can see why low-energy CID of electrospray ionized tryptic peptides has been so successful, since most tryptic peptides will have no more than one arginine at the C-terminus, yet

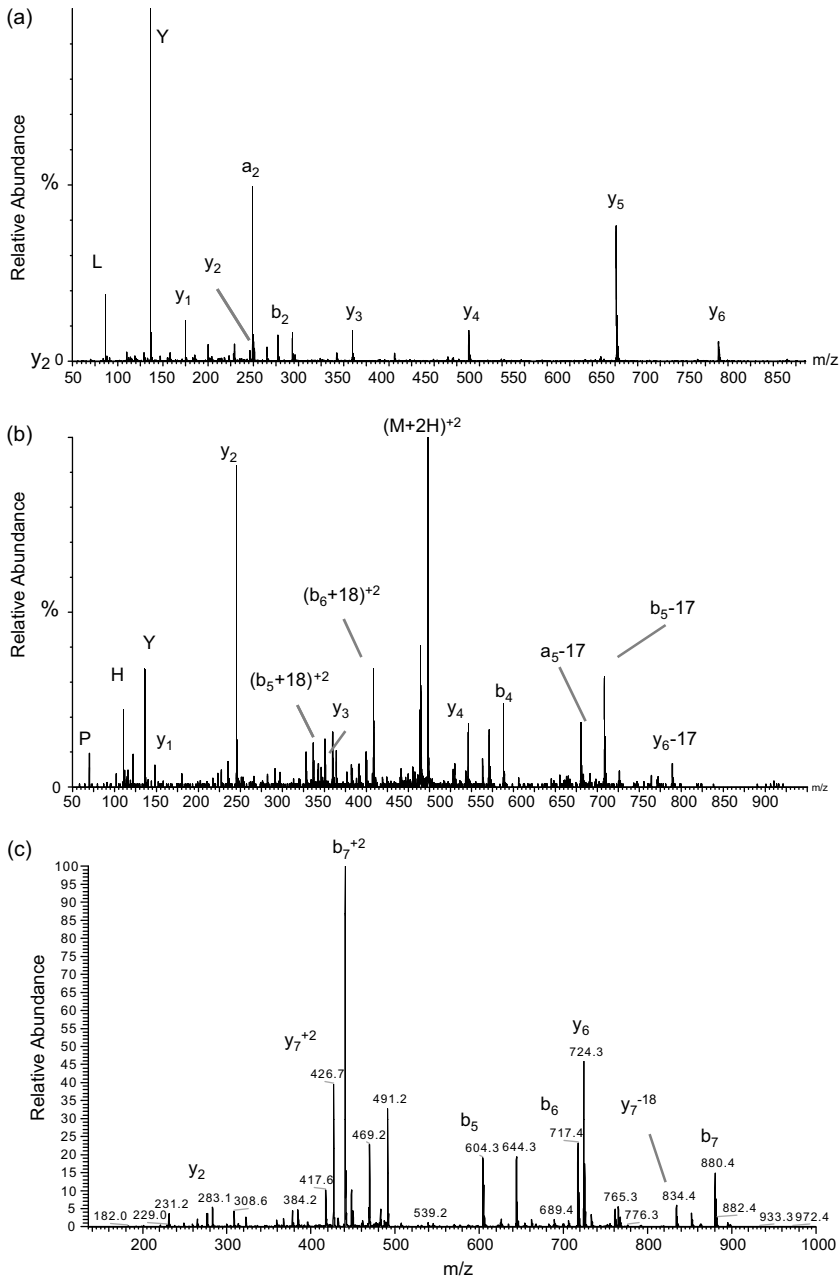


Figure 1.5 Effect of arginine on fragment ion formation. (a) CID of $(M + H)^{2+}$ precursor ion of the tryptic peptide YLYEIAR, where one of the protons is “mobile” and induces a contiguous series of γ -type ions plus some b -type ions. (b) CID of $(M + H)^{2+}$ precursor ion of the peptide YSRRHPE, which has two arginine residues and therefore no “mobile” proton.

Atypical fragmentations are seen and the sequence is impossible to determine. (c) CID of $(M + H)^{2+}$ precursor ion of the peptide FKGRDIYT, which has a mobile proton that induces b - and γ -type fragmentations. However, the arginine in the middle of the peptide prevents formation of a contiguous series of ions.

be able to take on two protons – one for the arginine side-chain and one “mobile” proton to produce the b/γ fragment ions. Even for cases where there is a mobile proton, the presence of arginine in the middle of a peptide sequence can have adverse consequences as illustrated in Figure 1.5(c). Here, the mobile proton allows the production of b - and γ -type fragments; however, cleavages near the arginine are of reduced intensity and overall sequence coverage is sparse.

Low-energy CID produces a few additional fragment ion types and the resulting spectra possess certain characteristics that are useful to note. Under “mobile proton” conditions, the presence of proline in a peptide typically results in intense γ -type (and sometimes the corresponding b -type) ions resulting from cleavage on the N-terminal side of proline. Concomitantly, cleavage on the C-terminal side of proline is nonexistent or very much reduced. These effects are due to a combination of increased gas-phase basicity of the proline nitrogen and the unusual ring structure of the proline side-chain that inhibits the attack of the carbonyl on the N-terminal side of the proline. Under “mobile proton” conditions, histidine promotes fragmentation at its C-terminal side, resulting in enhanced abundance of the corresponding b/γ fragment ions. Sometimes a b/γ cleavage will occur twice in the same molecule, resulting in a fragment ion that contains neither the peptide’s original C- or N-terminus (Figure 1.6a). These “internal fragment ions” usually only contain a few residues and are often present if one of the two required b/γ fragmentations is particularly abundant. For example, cleavage at the N-terminal side of proline is sometimes so facile that this fragment will often fragment again, resulting in “internal fragment ions” that have the proline at the N-terminal side of the internal fragment ion. The b - and γ -type fragment ions often undergo an additional neutral loss of a molecule of water or ammonia. These ions are often designated as $b-17$ or $b-18$, and so on. Under mobile proton conditions, these ions are usually less abundant than their corresponding b - or γ -type ion. The exceptions are when the N-terminal amino acid is glutamine or carbamidomethylated cysteine, in which case cyclization of the N-terminal amino acid and loss of ammonia occurs quite readily, resulting in abundant $b-17$ ions. Likewise, an N-terminal glutamic acid can cyclize and lose water, and the $b-18$ ions can be more abundant than the corresponding b fragment ions. In some cases, a b -type fragment ion can lose a molecule of carbon monoxide to form an a -type ion (28 Da less than the b -type fragment ion), although these seem to be more prominent for the lower mass fragments (e.g., it is not uncommon to find a_2 ions that are of comparable intensity to the b_2 ion in low-energy CID). Single amino acid immonium ions (Figure 1.6b) are often seen when MS/MS spectra acquisition includes this low mass region. Certain immonium ions are particularly diagnostic for the presence of their corresponding amino acid – leucine and isoleucine (m/z 86), methionine (m/z 104), histidine (m/z 110), phenylalanine (m/z 120), tyrosine (m/z 136), and tryptophan (m/z 159).

For peptide ions undergoing low-energy CID that lack a mobile proton, there are some additional fragment ions that become more prominent. Abundant ions resulting from cleavage at the C-terminal side of aspartic acid were first noticed in

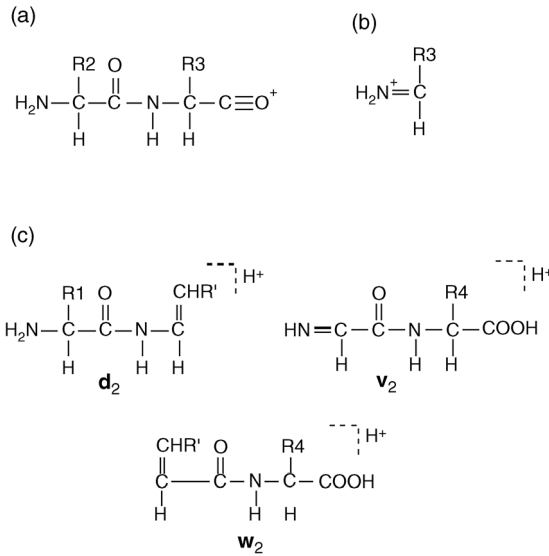


Figure 1.6 Additional ion types. (a) Internal ions are formed when a *b*/ γ -type fragmentation occurs twice in the same molecule. R2 and R3 denote side-chains of the second and third amino acids in the original peptide sequence.

(b) Single amino acid immonium ions are observed if data acquisition includes lower mass regions. (c) Additional ions have been observed in high-energy CID (above 1 keV), but not at low energy.

MALDI-PSD spectra [47]. It later became clear that in the absence of a mobile proton, the side-chain carboxylic protons from aspartic acid (and to a lesser extent glutamic acid) can provide the necessary proton to catalyze a *b*/ γ fragmentation [46]. This was first observed in the MALDI-PSD spectra, since the MALDI-derived singly charged ions need only a single arginine residue to lose the mobility of its one proton. Low-energy CID of peptide ions lacking a mobile proton also seem to be subject to the formation of a fragment ion that is sometimes called “*b* + 18” [45]. This is a rearrangement that occurs where the C-terminal residue is lost, but the C-terminal -OH group, plus a proton, are transferred to the ion. The designation “*b* + 18” refers to the fact that these have the mass of a *b*-type fragment ion plus the mass of water; however, the mechanism that gives rise to them is not related to the *b*-type fragmentation mechanism. Finally, it should be mentioned that low-energy CID of “nonmobile” peptide ions will often give more abundant neutral losses of water and ammonia (e.g., example, one might observe a γ - 17 ion in the absence of the corresponding γ -type fragment ion). For low-energy CID, MS/MS spectra from peptides with a mobile proton will exhibit the standard *b*- and γ -type fragment ions, and are most readily identified using database search programs. Likewise, spectra from peptides containing aspartic or glutamic acid in the absence of a mobile proton are also fairly readily interpreted. However, a “nonmobile proton” MS/MS spectrum of a peptide lacking aspartic or glutamic acid can be the most difficult type of peptide

to identify in a database search. This is especially true when the arginine is in the middle of the peptide.

The old multisection instruments were capable of subjecting peptide ions to much higher collision energy than the currently popular quadrupole collision cell and ion-trap instruments. At collision energies above 1 keV peptide ions can undergo alternative fragmentation pathways. In addition to the *b*/ γ fragments seen for low-energy CID, high-energy CID can induce some additional “charge remote” fragmentations (Figure 1.6c), including the *d*- and *w*-type fragment ions that allows for the distinction between leucine and isoleucine [48, 49]. In general, these high-energy CID fragmentations seemed not to be influenced by the presence or absence of a mobile proton, which made it easier to derive sequences *de novo* directly from the spectra without recourse to searching a sequence database [50]. As already mentioned, these instruments are not used much anymore, but high-energy collisions are still relevant for one of the more modern instruments. If collision gas is used in a MALDI-TOF-TOF instrument [20], the collision energies can be as high as a couple of kiloelectronvolts, and the resulting MS/MS spectra will contain the *d*-, ν -, and *w*-type fragment ions.

ECD is a process whereby an isolated multiply charged peptide ion captures a low-energy thermal electron, and the resulting radical cation becomes sufficiently unstable and fragments to produce *c*- and *z*-type fragment ions (Figure 1.4) [51]. Of key importance is that ECD induces fragmentation in a manner that does not result in intramolecular vibrational energy redistribution. In contrast, the additional energy acquired in CID is redistributed across the many vibrational modes of the entire molecule with the end result being that the weakest bonds break first, which often leaves insufficient energy for further peptide backbone cleavages. For example, low-energy CID of peptides containing phospho-serine or phospho-threonine usually results in a facile neutral loss of phosphoric acid. Sometimes the phosphate group stays attached, but usually not. The problem with this is that low-energy CID spectra of phosphopeptides typically exhibit a very abundant phosphoric acid neutral loss, but have tiny *b*/ γ -type fragment ions that may not rise above the noise. Hence, the user is left knowing that they have a phosphopeptide, but not which one. Glycopeptides behave similarly. In contrast, the ECD fragmentation process leaves the phosphate or carbohydrate attached to the *c*- and *z*-type fragment ions, which allows for one to identify the protein and pinpoint the site of phosphorylation or glycosylation [52].

The trapping of thermal electrons for use in ECD has only been possible in FT-ICR instruments, which happen to be the most expensive type of mass spectrometer. Avoiding this expense provided some of the impetus in the development of ETD [53], where anionic molecules are trapped in a linear ion trap (using radiofrequency electrical fields) and are mixed with multiply charged cationic peptide analyte ions. Given the appropriate anion (one with low electron affinity), an electron is transferred to the peptide cation in an exothermic process that induces the production of the same *c*- and *z*-type fragment ions observed in ECD (Figure 1.4). ETD is sufficiently rapid that it can be used in conjunction with LC-MS/MS, and is sometimes used along with CID (i.e., data-dependent analysis might trigger the acquisition of both an

ETD and a CID spectrum from the same precursor ion). Similar to ECD, ETD seems to be particularly useful for the analysis of post-translational modifications (PTMs) that are otherwise labile under CID conditions (e.g., phosphorylation) [54]. For shotgun protein identifications, CID and ETD appear to be complementary in that ETD tends to be more successful at identifying peptide precursor ions with higher charge density, whereas CID is better at precursors with one to three protons [55].

1.2

Basic Protein Chemistry and How it Relates to MS

1.2.1

Mass Properties of the Polypeptide Chain

Proteins are linear chains of monomers made up of 20 standard amino acids (Table 1.2) that can be as massive as a few mega-Daltons (e.g., titin), but are typically in the range of 10–100 kDa. Proteins can exhibit a fairly wide range of physical properties such as solubility and hydrophobicity, which can make it difficult to find a universal means of separating and isolating them. Although most proteins are soluble in the buffers used for sodium dodecylsulfate–polyacrylamide gel electrophoresis (SDS–PAGE), the resulting separation leaves the proteins within a polyacrylamide gel matrix from which it is difficult and inefficient to extract the intact proteins. Proteolytic digestion and release of peptides derived from proteins entrained in gel slices is relatively efficient (in-gel digestion) [56]. In contrast to intact proteins, peptides derived from proteins via proteolysis tend to have more uniform distributions of physical properties that make them amenable to standard peptide separation techniques such as HPLC. There will often be a subset of proteolytic peptides for each protein that exhibit favorable properties with respect to chromatography and ionization. Therefore, most proteomics involves the so-called bottom-up approach of first ravaging proteins with one protease or another, analyzing the resulting peptide bits, and then trying to deduce which proteins were present in the first place.

1.2.2

In Vivo Protein Modifications

The standard amino acids can be decorated with a variety of biologically significant modifications. There are a couple of Web resources that list the various modifications that have been observed and these should be used whenever unexpected mass shifts are observed (www.unimod.org and <http://www.abrf.org/index.cfm/dm.home>). Some of these modifications will alter more than just the residue mass, possibly making the modified peptide more or less readily ionized, hydrophilic, or soluble. In some cases, the modification introduces a chemical bond that is particularly labile to mass spectrometric fragmentation. Therefore, interpreting spectra from modified

peptides is often a tricky business that involves more than just adding the right masses together. What follows is a brief description of just a few of the more common protein modifications, focusing on chemical and physical properties that are relevant to their analysis by MS.

Glycosylation is one of the more common modifications, which can be subdivided into at least four categories – *N*-linked, *O*-linked, *C*-mannosylation, and cytosolic *O*-GlcNAc modifications (GlcNAc = acetylglucosamine). *C*-Mannosylation is a modification of tryptophan in a WXXW motif, where the first tryptophan is modified by mannose via a carbon–carbon bond [57]. This bond is stable to low-energy CID, and can therefore be readily pinpointed using standard tandem MS methods. The *O*-GlcNAc modification is a very interesting modification involved in signal transduction pathways that occurs on nuclear and cytoplasmic proteins in eukaryotic cells. Specific serine and threonine residues are modified by *N*-acetylglucosylaminyl-transferase and *O*-GlcNAcase, which are the two enzymes that dynamically attach and remove this single monosaccharide [58, 59]. Low-energy CID of *O*-GlcNAc-peptides tends to produce abundant fragment ions resulting from loss of the monosaccharide leaving the modified serine intact, whereas ETD preferentially cleaves peptide bonds thereby leaving the *O*-GlcNAc attached to the modified residue [60]. In contrast to *O*-GlcNAc modification, the more standard extracellular *N*- and *O*-linked glycosylation are polymeric in nature, and the carbohydrate structures are typically large (above 2000 Da for *N*-linked) and heterogeneous at any given site of modification. Sites of *N*-linked glycosylation are determined relatively easily by use of *N*-glycosidase F, which removes the entire carbohydrate from the side-chain of asparagine and in the process converts it to aspartic acid [61, 62]. This modification only occurs at a specific sequence motif consisting of asparagine, followed by any residue except proline, which is then followed by serine, threonine, or cysteine. Thus, identification of aspartic acid in place of asparagine in such a motif after treatment by *N*-glycosidase F is sometimes considered to be sufficient for identifying a site of *N*-linked glycosylation. Given that asparagine is capable of chemically deamidating [63], absolute proof can be obtained by performing the enzymatic deglycosylation reaction in the presence of ^{18}O water, which is incorporated into the deamidated aspartic acid side-chain. Extracellular *O*-linked carbohydrates are smaller than *N*-linked (only a few carbohydrate monomers) and are attached to serine or threonine, but within no clear sequence motif. Unlike *N*-glycosylation, there is no robust means of removing the glycan prior to mass spectrometric analysis, and determinations of *O*-linked glycopeptides and proteins can be quite challenging [64]. Enrichment of glycopeptides and glycoproteins is typically accomplished using lectin-affinity chromatography [65–68]. The analysis of carbohydrate heterogeneity for a specific glycopeptide site is best performed in a stepwise manner by treating the proteolytic sample with appropriate enzymes [69]. For example, *N*-acetyl neuraminic acid residues are easily removed with neuraminidase (Figure 1.7). This stepwise enzymatic treatment is beneficial for revealing fine structural details of complex glycopeptides mixtures. CID of glycopeptides results in fragmentation of glycosidic bonds allowing for characterization of the carbohydrate portion, but fragmentation of peptide bonds is usually absent. For determination of the site of glycosylation, ETD can provide peptide fragmentation. For isolation of

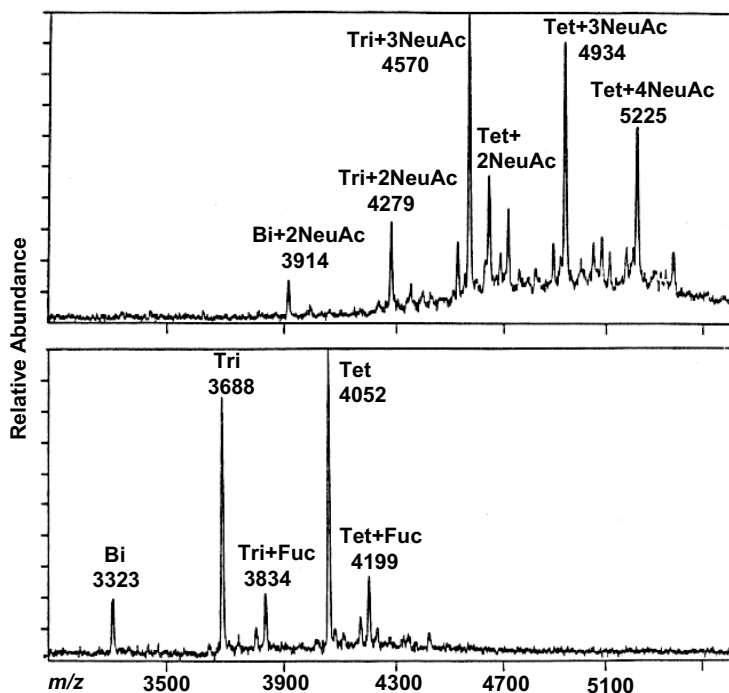


Figure 1.7 Negative-ion MALDI-TOF mass spectra displaying carbohydrate heterogeneity (bi-, tri-, and tetra-antennary) for glycopeptide site III (IQATFFYFTPN⁺KTE) obtained from Staph V8 digest of human α_1 -acid glycoprotein; (top) before and (bottom) after treatment with

neuraminidase that removes *N*-acetyl neuraminic acid (NeuAc) residues while retaining other sugar moieties including Fucose (Fuc). The matrix solution was a 2 : 1 mixture of 2-aminobenzoic acid: nicotinic acid and a nitrogen laser at 337 nm was used.

N-linked peptides, peptide identification, and determining sites of *N*-linked modification, the “glycocapture” technique is particularly useful [70, 71].

Phosphorylation is a dynamic modification that most often occurs on serine, threonine, and tyrosine. Low-energy CID of peptides containing phospho-Ser/Thr tends to produce abundant fragment ions resulting from the neutral loss of phosphoric acid (loss of 98 Da), where the sequence-specific ions (*b*- and γ -type) are much less intense. In contrast, ETD tends to leave the phosphate intact while still promoting sequence-specific fragment ions (*c*- and *z*-type), which makes pinpointing the site of phosphorylation more reliable [54]. Phospho-tyrosine is more stable, and sequence specific ions that still possess the phosphate are prominent in these CID spectra. There is evidence that during CID, phosphate groups can migrate from one site to another within a peptide molecule [72] and this is particularly pronounced in the absence of a mobile proton. Multiple phosphorylations of individual proteins seem to be quite common, which adds to the difficulty of analysis. Moreover, not all protein phosphorylation appears to be functionally relevant and, more importantly, different phosphorylation sites on the same protein may regulate different processes.

Thus, the challenge for understanding how phosphorylation modulates a given biological pathway is to discover which phosphorylation sites on a given protein are the relevant ones and how phosphorylation at those sites changes in response to various stimuli. The ability to derive quantitative information on specific phosphorylation sites is imperative to this goal. A confounding effect of phosphorylation analysis is that phosphorylated peptides can behave differently from their nonphosphorylated counterparts (e.g., changing solubility, ability to be ionized, chromatographic behavior, or tendency to adsorb to surfaces). The presence of phosphate near a predicted proteolytic site may inhibit proteolysis, which can make it difficult to make direct quantitative comparisons between phosphorylated and nonphosphorylated peptides encompassing the same site of modification. Even if the physical properties were identical, the oftentimes low stoichiometry of phosphorylation means that much larger amounts of sample needs to be analyzed in order to detect the phosphorylated peptides. For this reason, phosphopeptides and phosphoproteins are often enriched prior to mass spectral analysis (e.g., [73–77]). Of course, by enriching a phosphopeptide and removing its unphosphorylated counterpart, it becomes impossible to determine stoichiometry.

There is a class of proteins called ubiquitin-like modifiers (UBLs; proteins including ubiquitin, NEDD8, ISG15, SUMO1, etc.) that are all used by cells to tag other protein substrates on specific lysine residues [78]. This tagging of protein substrates by UBLs serves a variety of purposes ranging from targeting the substrate for degradation to signaling functions. In some cases, a single UBL will modify a particular lysine residue; in other cases, long chains of polymerized UBLs are attached to a substrate lysine. Although structurally more complicated than phosphorylation, this PTM is similar in that it can be dynamic. There are enzymes that put UBLs on substrates and others that take them off. Like phosphorylation, the stoichiometry of the modification can be quite low and care must be taken that the deubiquitinating enzymes do not remove the UBLs during sample preparation. From the standpoint of MS, it is important to note that all of these UBLs are attached to the substrate lysine ϵ -amino group via an amide linkage to the UBL's C-terminus that always ends with Gly–Gly. In the case of ubiquitin itself, the Gly–Gly sequence is preceded by an Arg residue, which upon tryptic digestion of the substrate leaves the formerly ubiquitinated lysine tagged with a Gly–Gly [79, 80]. This often forms the basis for the identification of ubiquitination sites, and it should be pointed out that the amide linkage of Gly–Gly to the ϵ -amino group of lysine is quite stable to CID (in contrast to Ser/Thr phosphorylation or *N*- and *O*-glycosylation). Other UBLs may not have this arginine residue. For example, SUMO1-modified proteins have a 19-amino-acid peptide appended to the lysine ϵ -amino group, which makes the identification via CID a bit of a challenge [81].

Acylation of protein N-terminal amino groups and lysine ϵ -amino groups is a common PTM. Acetylation of the protein N-terminus occurs on over half of eukaryotic cytosolic proteins; myristoylation of N-terminal glycine is also found in a small number of cytoplasmic proteins. Acylation produces an amide bond that is stable to low-energy CID, which makes MS/MS analysis considerably easier than the more labile modifications described above (e.g., glycosylation and phosphorylation).

Acetylation of lysine side-chains is a reversible modification that appears to be involved in a variety of cellular processes [82]. Acetylation of the lysine side-chain prevents proteolytic cleavage by trypsin, which therefore makes it difficult to make quantitative comparisons with the unmodified form if trypsin is used. Acetylation also reduces both the solution and gas-phase basicity of the lysine side-chain, which would likely influence peptide ionization and charge state, as well as retention time in cation-exchange chromatography. Enrichment of acetylated peptides can be accomplished using anti-acetyl lysine antibodies [82, 83].

Disulfide bonds are one of a few protein modifications that result in a loss of mass (two hydrogen atoms per cysteine pair). In principal, the determination of disulfide bonds is a simple matter of measuring the mass of proteolytic peptides before and after reduction, where one looks for ions that disappear after reduction as well as the appearance of the corresponding peptide ions containing reduced cysteine. In practice, disulfide determination is rarely this simple. In order to unambiguously assign disulfide linkages, proteolytic cleavage sites must be located between every cysteine, which they often are not. Moreover, proteins with intact disulfide bonds are often refractory to proteolytic degradation, so the intended cleavages often do not occur. A typical outcome for a disulfide experiment is to identify some of the reduced peptides, but not the original disulfide-linked peptide (or vice versa). Or sometimes one of the reduced ions is observed, but not the other (perhaps it chromatographs poorly or is not ionized). Or in a protein with several disulfide bonds, a few of them might be determined, but the others are refractory. Sometimes this can be overcome by using different proteases, whereby the resulting peptides might have more favorable chemical and physical properties for analysis. One aspect of disulfide chemistry that is often forgotten is the fact that once a protein starts losing secondary and tertiary structure (via proteolysis or addition of denaturant), both acid- and base-catalyzed scrambling can occur [84]. For proteolysis at $\text{pH} > 7$ one needs to add a small amount of alkylating reagent to eliminate any catalytic amounts of thiol that might be present in the sample. The base-catalyzed scrambling drops off 10-fold per pH unit, which is why pepsin cleavage at pH 3 is often used for these purposes. The acid-catalyzed scrambling occurs in the presence of above 6 M HCl and is not typically a condition used in protein chemistry. Finally, it should be noted that although CID does not typically break a disulfide bond, ETD favors cleavage of this bond [85, 86].

Proteolysis is normally thought of as something that is done to samples during a bottom-up proteomic analysis; however, it is also an important PTM that occurs in a variety of settings and has numerous biological purposes. One of the most common proteolytic events is the removal of the N-terminal initiator methionine. Secreted proteins and certain classes of membrane proteins possess secretory signal sequences at the N-terminus that are proteolyzed while entering the endoplasmic reticulum. There are cell surface proteases that clip other membrane proteins to release the extracellular domain; tumor necrosis factor being one of the more famous examples of a shed membrane protein [87]. There are many other examples of proteolysis occurring in a wide variety of biological processes ranging from blood clotting to processing polypeptide chains into smaller peptide hormones. It can be

relatively easy to establish that a proteolytic event occurred by, for example, identifying a transmembrane protein extracellular domain in some cultured cell supernatant [88]. Or sometimes one might identify a protein from a SDS-PAGE gel slice that is at a much lower molecular weight than would be predicted from the full sequence. However, identifying the specific site of proteolysis can be difficult if peptides from that region are hard to ionize or have unfortunate chromatographic properties, especially if the peptide containing the endogenous cleavage site is further modified, for example, by *O*-linked glycosylation. Obviously, the endogenous cleavage site has to be different from the protease specificity used to create peptides for LC-MS/MS (e.g., the C-terminus of a protein cannot be either arginine or lysine if trypsin was used). There are a few methods available that enrich for N- or C-terminal peptides that might be useful for these purposes [89–94].

1.2.3

Ex Vivo Protein Modifications

Aside from the numerous chemical modifications researchers do to proteins on purpose (e.g., reduction of disulfide bonds, alkylation of thiols, or various reactions that incorporate stable isotopes into modified peptides), there are several that occur by accident during sample handling. These modifications typically add mass and the corresponding shifts are measured by MS. What follows is a brief list of the more common ones.

Denaturation in urea is often accompanied by carbamylation of amino groups (either N-terminal or lysine), as well as other functional groups on other side-chains to a lesser extent [95]. Urea is in equilibrium with ammonium cyanate and it is the latter that is reactive with amines. Carbamylated peptides, like acetylated peptides, are stable to low-energy CID, which means that fragmentation of the peptide bonds will not result in loss of the carbamyl group. The key to limiting carbamylation is trying to limit the concentration of cyanate anion by making urea solutions fresh from solid urea, avoiding elevated temperatures, and using ion-exchange resins to deplete cyanate from the neutral urea solutions. In addition, use of amine-containing buffers (e.g., Tris) should also help scavenge cyanate. Acidification is often done to halt a tryptic digestion, but it will also limit further carbamylation by protonating amino groups.

Although there may be functional roles for *in vivo* oxidation of tryptophan and methionine, most often these modifications are observed as a result of sample handling. Exposure to oxidants can occur while running a SDS-PAGE gel [96, 97] or even from reaction of ozone from outside air with thin dry layers of samples during MALDI preparation [98]. The extent of modification can be limited by reducing exposure to oxygen (e.g., purging samples with argon before extensive digestion periods). Obviously, exposure to oxidizing chemicals such as sodium periodate (used in the so-called glyco-capture method [71]) or performic acid (used for cleaving disulfide bonds) will cause extensive oxidation [99]. Oxidation of methionine typically adds a single oxygen, and in low-energy CID neutral losses of 64 Da (HSOCH_3) are often observed as satellite peaks below any fragment ion that contains the oxidized

methionine [96]. Even more extensive oxidation can lead to an additional oxygen (+32 Da) added onto the sulfur, but this is not seen as frequently. Oxidation of tryptophan is more complicated, and can result in mass increases of 3.9949, 15.9949, 19.9898, and 31.9898 Da [97].

Deamidation can occur at asparagines, glutamine, and carbamidomethylated cysteine residues. When glutamine is located at the N-terminus of a peptide (or protein) the alpha-amino group undergoes a nucleophilic attack of the side-chain amide resulting in the loss of ammonia and formation of a cyclic five-membered ring (pyroglutamic acid) [100, 101]. In buffers typically used for tryptic digestion, the half-life of this reaction is in the range of several hours to a day, so for a standard overnight tryptic digestion a substantial fraction of peptides with N-terminal glutamine will have converted. In a very similar fashion, N-terminal carbamidomethylated cysteine will also cyclize and lose ammonia to form (*R*)-5-oxoperhydro-1,4-thiazine-3-carbonyl residue [102]. The half-life for this reaction is also on the order of hours to days and is often seen in tryptic digests. N-Terminal asparagine does not undergo this reaction, since it would result in an unfavorable four-membered ring structure. However, when asparagine is not located at the terminus it can undergo a nucleophilic attack of the amide nitrogen on the C-terminal side of asparagine forming a succinimidyl intermediate that can then re-open as aspartic acid or isoaspartic acid [103]. The rate at which this reaction occurs is dependent on the steric hindrance introduced by the residue located C-terminal to the asparagine. Sequences containing Asn–Gly are particularly prone to this *ex vivo* modification. Internal glutamines can also deaminate, but the rate of reaction is orders of magnitude slower [63].

Even if purified “proteolytically correct” peptides enter a mass spectrometer, the mass spectrometer source may generate ions other than the desired intact protonated species. Either by design or accident, in-source CID [104] can occur when ions are accelerated with higher energy through regions of high pressure (e.g., use of high cone voltages for certain source designs). When these fragment ions are detected in a data-dependent scan mode, MS/MS spectra of these in-source fragments can be collected and a database search identifies them as “proteolytically incorrect” peptides. The most labile bonds are preferentially cleaved via in-source CID; for example, MS/MS are often collected on fragments containing an N-terminal proline (i.e., production of a γ ion via in-source cleavage at proline). Protons typically provide the positive charge for peptide ions; however, contaminated solvents or incomplete desalting can result in peptide charging via sodium, or other adventitious cations. Not only will this lead to incorrect mass determinations, but the MS/MS spectra will exhibit atypical fragment ions [105–107].

Proteolysis was mentioned earlier in the context of an *in vivo* post-translational event that is often of considerable biological interest; however, inadvertent proteolysis can also occur *ex vivo* through experimental mishandling. It is well known that cell lysis can release proteases from subcellular compartments and one typically disrupts cells only in the presence of a variety of protease inhibitors where the sample is worked-up at reduced temperature. In the case of trypsin it is thought that autolysis results in a protease that is still active, but with reduced specificity for arginine and lysine. Partial methylation of lysine side-chains within trypsin eliminates some of

these cleavage sites, thereby allowing for a prolonged use of trypsin. Even with precautions, a low level of nonspecific cleavages can occur [108]. The goal of achieving complete tryptic digestion has to be balanced against the increased level of nonspecific cleavage.

1.3

Sample Preparation and Data Acquisition

1.3.1

Top-Down Versus Bottom-Up Proteomics

Bottom-up MS/MS methods are based on matching a single peptide to a single MS/MS spectrum. Of course, a given protein is likely to be digested into many different peptides and many of these will be identified, all of them pointing to the identification of the same gene product. The difficulty is that a single gene can give rise to many different proteins, either through gene splicing, proteolytic processing, or a variety of other PTMs. However, since the intact protein structure has been destroyed by proteolysis (trypsin), there is no way of reassembling the peptides into a 100% accurate determination of the protein present originally. This fact is one of the more compelling reasons for the promotion of the so-called top-down approach to proteomics [109]. Here, the intact protein is analyzed – measuring the masses and relative amounts of all of the protein variants and acquire structural data on each one individually. Clearly, this is the most logical route to take; however, the technical difficulties are significant and in many cases insurmountable. Typically, these experiments can only be done with the most expensive instrumentation (i.e., FT-MS), and one can only apply the technique to the most well-behaved proteins (soluble abundant ones that can be chromatographed in buffers suitable for ESI). Despite the difficulties, the top-down approach is becoming more popular and the identification of thyroglobulin extended the upper mass limit of the top-down to 669 kDa [110]. The bottom-up methods, where proteolytic peptides are analyzed, are likely to be applicable to the majority of biological problems; however, one needs to understand the limitations when attempting to jump from peptide identifications to protein identifications.

1.3.2

Shotgun Versus Targeted Proteomics

Data-dependent shotgun analysis is the process whereby MS/MS spectra are acquired for the more abundant precursor ions over time as they elute from an HPLC column. These spectra are then analyzed as described below (Section 1.4.1), where the goal is to identify previously unknown proteins present in a sample. The problem with shotgun analysis for complex proteomic samples is that only the more abundant proteins are identified. To identify lower abundance proteins one needs to fractionate the proteins using, for example, SDS-PAGE [56], multi-dimensional

HPLC [27], gas-phase fractionation [111], isoelectric focusing [112], or extended gradients [113]. Sometimes combinations of these fractionation techniques are used such that a single sample will be subject to mass spectrometric analysis for several days to weeks. The goal is to increase the dynamic range over which protein identifications can be made; however, this process is subject to diminishing returns and the sample throughput is very slow.

In contrast, targeted proteomics is a much more sensitive method that has the goal of verifying the presence and quantity of known proteins within a sample. For each target, one needs to specifically monitor a few tryptic peptides that serve as surrogate measurements for the protein. Ideally these tryptic peptides would readily form from tryptic digestion, exhibit sharp chromatographic peaks, ionize easily, and not contain any confounding amino acid residues or sequences that could lead to variable quantitative results (no methionine or Asn–Gly, for example, that can variably oxidize or deamidate). The most sensitive way to perform targeted proteomics is to carryout SRM using a triple quadrupole (see Section 1.1.2). Setting up an SRM assay for targeted proteins involves determining which peptides are formed by tryptic cleavage and identifying those peptides that produce the most abundant precursor ions and their charge states, and then subjecting those precursors to CID and acquiring the MS/MS data. From these spectra one would choose the product ions to monitor – usually the more abundant γ -type ions, preferably at higher m/z than the precursor ion where there is less background. The assay is ready to use once several transitions (precursor–product ion pairs) have been established for each peptide to be monitored in a set of samples. Development of these SRM assays is time-consuming; however, there is an initiative called the SRM Atlas that has the goal of predetermining assays for every open reading frame from various species [114]. Such an atlas has already been completed for yeast [115], which allows for targeted SRM experiments to be performed without extensive assay development time.

1.3.3

Enzymatic Digestion for Bottom-Up Proteomics

The protease most frequently used is trypsin, which cleaves on the C-terminal side of arginine and lysine. This sounds like a simple rule, but there are a number of nuances. Usually the rule for trypsin also includes the prohibition of cleavages N-terminal to proline; however, there is growing evidence [116] that this cleavage reaction can occur with very slow kinetics. Sometimes trypsin will not cleave at certain arginine or lysine sites, which may be due to having stopped the proteolysis too soon – a process called “limited proteolysis” where the most susceptible bonds are cleaved first (at the “hinges” and “fringes” of a folded protein). Or sometimes trypsin cleavage is slowed or prevented by the presence of surrounding acidic residues. Also, it needs to be kept in mind that trypsin is not an exopeptidase. When there is a short series of contiguous arginine or lysine residues (e.g., the sequence ELVISKKRISQ-ING), trypsin will cleave at one of the sites, thereby producing two new peptides that contain additional cleavage sites at the N- and C-termini (e.g., ELVISKK and

RISQING). However, these potential cleavage sites at or near the termini of the resulting peptides are not amenable to further cleavage. Finally, it should be noted that trypsin is capable of nonspecific cleavages at a very low level [108], which is a problem that becomes worse with prolonged incubation times. Trypsin autolysis (self-digestion) results in a slightly damaged protease with reduced specificity. To prevent this there are a number of vendors that sell trypsin that has been partially methylated on lysine – the sites of self-immolation. Nonspecific cleavages can be a significant source of background when working with samples that possess a wide dynamic range of protein concentration (e.g., blood plasma). In addition to the very high abundance fully tryptic peptides (cleavage at arginine or lysine at each end of the peptide), the low level of semi-tryptic peptides (one end produced by nonspecific cleavage) will still be more abundant than the fully tryptic peptides derived from low-level proteins. Trypsin is not perfect.

Other imperfect, but useful, enzymes include Lys-C, Lys-N, Asp-N, and Glu-C, which as their names imply, cleave on the C-terminal side of Lys, the N-terminal side of Lys, the N-terminal side of Asp, and the C-terminal side of Glu. Lys-C enzymes are commercially available from at least two biological sources (*Achromobacter lyticus* and *Lysobacter enzymogenes*) [117] and will generally produce larger peptides than trypsin. Similarly, Lys-N is an enzyme isolated from the mushroom *Grifola frondosa*, which cleaves on the N-terminal side of Lys [118, 119]. Asp-N protease cleaves at aspartic residues around 200 times faster than at glutamic acid, which means that some Glu-N activity will be seen, especially at higher enzyme/substrate ratios and with prolonged incubation time. Likewise, Glu-C will exhibit some Asp-C activity [120]. As with trypsin, one would expect to find low levels of nonspecific cleavages using these or any other enzyme. Most other enzymes, such as pepsin, chymotrypsin, subtilisin, or thermolysin, do not have reliable cleavage specificity. They can cleave at many different residues and will often produce peptides with ragged ends. There are chemical cleavage methods available, too, but the ones with the greatest specificity are those that cleave at the rarest amino acids (methionine, cysteine, and tryptophan), and therefore on average produce larger peptides. Large peptides can be good or bad. Production of a few larger peptides for each protein results in less complex mixtures for analysis, and theoretically would improve the ability to identify lower abundance proteins. On the other hand, large peptides can be more difficult to chromatograph, fragment, analyze, and identify.

1.3.4

Liquid Chromatography and Capillary Electrophoresis for Mixtures in Bottom-Up

The analysis of peptide mixtures obtained from enzymatic digests of proteins is best performed by coupling liquid chromatography with MS (liquid chromatography mass spectrometry LC-MS). The most common approach utilizes reverse-phase (C18) columns with ESI for online analysis. Commercial columns are available ranging in size from narrow bore (1–2 mm inner diameter) to capillary (above 50 μm inner diameter). Thus, users can match the column loading capacity with the sample size [121]. Greater overall sensitivity is achieved using the narrowest bore columns;

however, larger bore columns tend to be more robust and easier to use (more reproducible retention times, less plugging, and flow rates that are easier to manage). Fortunately, HPLC manufacturers have come out with suitable pumps and fittings that make it much easier to work with packed capillaries. Coupling capillary electrophoresis to MS (capillary electrophoresis mass spectrometry CE-MS) has been less popular due to limited sample loading compared to liquid chromatography. The advantages of capillary electrophoresis are lower sample consumption, shorter analysis time, and higher separation efficiencies. These benefits were shown in the analysis of a tryptic digest of human cerebrospinal fluid [122]. The high-throughput digestion of proteins is achieved by coupling of immobilized enzyme columns in tandem with the reverse-phase columns [123]. The interaction time of proteins with the immobilized enzyme phase is controlled by varying the flow rate through the enzyme column, which could be useful for digesting proteins resistant to proteolysis (Figure 1.8).

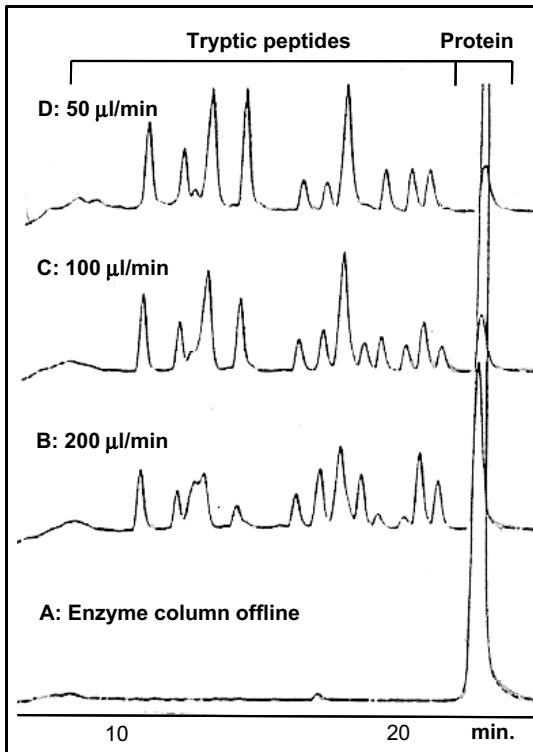


Figure 1.8 HPLC tryptic map of horse cytochrome *c* using a 2.1×150 mm Vydac-C18 column at a flow rate of 200 ml/min over 10 column volumes. A 2.1×30 mm Trypsin-POROS column was equilibrated at 50°C . A

digestion buffer of 25 mM Tris-HCl, pH 8.5 containing 10 mM CaCl_2 was used with flow rates of (B) 200, (C) 100, (D) 50 ml/min. The protein digestion increases by decreasing the flow rate of buffer through the enzyme column.

Although ESI is usually used, it is possible to couple liquid chromatography or capillary electrophoresis to MS using MALDI [124]. Peptides eluting from a reverse-phase capillary column are deposited off-line on a MALDI sample stage, and subsequently analyzed using appropriate software and robotics. This is potentially a high-throughput method where several HPLC and MALDI plate spotters could prepare plates to be analyzed by a single MALDI mass spectrometer. By decoupling the HPLC from the mass spectrometer in this manner, it is possible to interrogate an HPLC run several times, possibly performing MS/MS on various precursors, all at a leisurely pace. In practice, LC-MALDI-MS has not been very popular, due to the technical difficulty of making homogeneous sample-matrix spots. It can also be difficult to troubleshoot HPLC problems when the detector is off-line.

1.4

Data Analysis of LC-MS/MS (or CE-MS/MS) of Mixtures

1.4.1

Identification of Proteins from MS/MS Spectra of Peptides

Mass mapping was the first high-throughput MS method developed for protein identification, where the general idea was to compare observed molecular weights of tryptic peptides with those calculated from a protein sequence database [125]. This procedure generally requires purified proteins (e.g., in-gel digests from two-dimensional gel spots) and is most rapidly performed using a simple MALDI-TOF instrument. However, for more complex samples containing more than two or three proteins, data-dependent shotgun analyses acquiring many MS/MS spectra have become typical. In order to analyze all of this data, each of the MS vendors has developed (or licensed) their own software for seamlessly moving from raw data files to protein identifications; however, most research groups do not find this solution satisfactory. Either the software is inadequate, not portable to different operating systems, or a single workflow is desired that can encompass data from different mass spectrometers regardless of the vendor. Hence, there has been a move towards open-source software solutions. The description that follows is based on the general flow used by one of the open-source packages, the Trans-Proteomic Pipeline (TPP), which involves (i) extracting MS/MS spectra from raw binary data files, (ii) performing database searches, (iii) validating the peptide to spectrum matches, and finally (iv) validating the protein identification. For more in-depth details on how to use the TPP, a tutorial has recently been published [126] and for the casual user an Internet version of the TPP is being developed at the Australian Proteomics Computational Facility (www.apcf.edu.au).

Data files produced by mass spectrometers from different vendors have proprietary formats that need to be converted to a common open format. Some formats are flexible enough that they can capture most of the information contained in the original raw file (e.g., mzXML [127]), which is useful if subsequent processing of the

data involves MS spectra, in addition to MS/MS spectra. However, the file sizes for these open formats tend to be larger than the original raw binary file, so some laboratories favor smaller and simpler text file formats (e.g., *dta* or *mgf* files) that only contain fragment ion m/z and intensity values with headers containing limited information such as scan number, precursor m/z , and charge state. Several conversion programs have been written for each vendor's data files (e.g., ReAdW for conversion of Thermo raw files); however, there has been progress made within the ProteoWizard set of open-source tools and libraries [128] to support reading of multiple vendor files. In most cases, the converted files tend to be faithful reproductions of the original raw file; however, it seems likely that in the future conversion programs will optionally be able to perform some level of data enhancement. First, for low-resolution MS/MS spectra one could remove some of the noise via a moving window filter that, for example, only retains the four most intense ions within a 60 m/z window. Using such a filter, an *mgf* file can be reduced in size by as much as 90%, and provide improved search results [129]. An exact conversion of the raw file to an open-source format associates the MS/MS spectra with the low intensity precursor mass measurement that triggered the MS/MS acquisition. A better approach is to take an intensity weighted average of the m/z measurements for all of the isotope peaks, for all precursor charges that are present, and for all of the mass spectra acquired across the chromatographic peak. This recalculation of the precursor mass is particularly useful when the single-stage mass spectrum level (MS^1) is acquired at high resolution and mass accuracy [130]. Also, for high-resolution and high-mass-accuracy MS/MS spectra it seems that a significant improvement in database search results could be obtained by deisotoping the fragment ions (transforming isotope clusters into a single value corresponding to the ^{12}C peak). Ideally, all of this data manipulation could occur at the point where the raw files are converted to an open format.

The next step is to perform a database search (also see Chapter 14). SEQUEST [25] was the first to perform a database search without any user-derived interpretation. The University of Washington, where SEQUEST was invented, gave Finnigan (now Thermo Corporation) an exclusive license to sell the software, along with the requirement that they vigorously defend the intellectual property. This briefly held up the development of alternative database search software, but others eventually came along. In addition to some for-profit software such as Mascot [24] and Phenyx [131], there are now a variety of freely available programs (e.g., Tandem [132], MyriMatch [133], OMSSA [134], and InsPecT [135]). With few exceptions, these programs use the precursor mass and tolerance as a filter to derive a list of candidate sequences from a protein sequence database. Mock spectra are made for each candidate sequence, these are compared to the real MS/MS spectrum, and scores are assigned to how well they match. The top-scoring match is the winner. Despite having identical purposes, and often similar algorithms, each program will produce slightly different results. There are at least two reasons for these differences. (i) Each program might process the real MS/MS spectrum differently (e.g., by de-noising or eliminating some of the fragment ions in various ways.) (ii) Each program will score the match between mock and real spectra with different equations, models, or

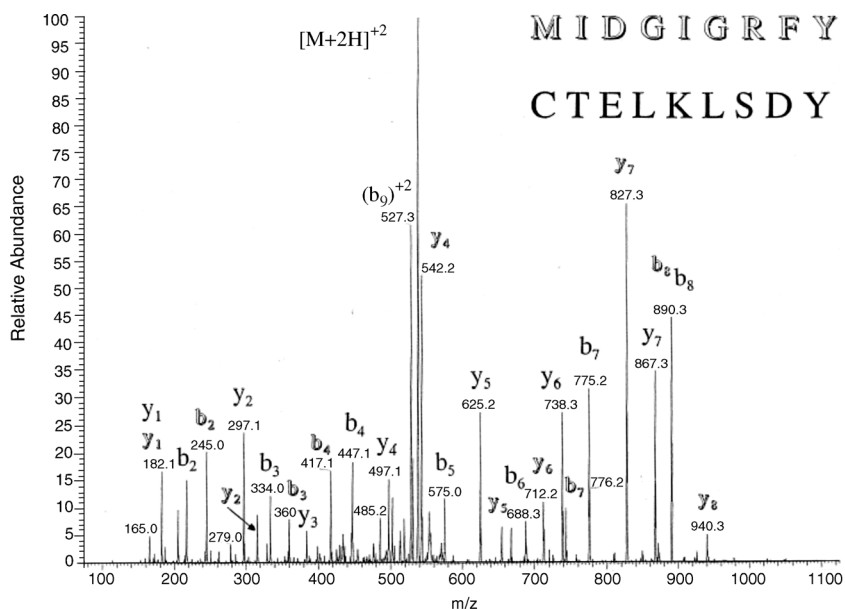


Figure 1.9 Simultaneous CID of $(M + H)^{2+}$ precursor ion containing two isobaric peptides with sequences of MIDGIGRFY and CTELKLSDY recorded on an ion-trap mass spectrometer. Both peptides were correctly identified by the PepSearch program and were correlated to the influenza A virus nucleoprotein.

methods. Each search engine has its own search result output format that needs to be converted to a single common format (e.g., pepXML) in order to be accessed in the next step – validation. To illustrate the accuracy of protein sequence identification through automated database searching, MS/MS spectrum obtained for a mixture containing two isobaric peptides was analyzed [136]. The CID mass spectrum shown in Figure 1.9 contains the *b*- and *y*-type ions for two different sequences, and despite this complexity, the program correctly identified both sequences associated with the isobaric peptides.

As just described, all MS/MS spectra will each have a top-scoring candidate sequence; however, the difficulty is determining whether it is correct [137]. Early on, validation was done manually by expert review or by using a simple score threshold that was assumed to reliably bisect correct assignments from incorrect ones. This approach was refined by the use of metrics that show how much better the top score for a given MS/MS spectrum was from all of the other candidate sequences. For SEQUEST, this was simply a score difference between the first and second ranked sequence candidates; later, expectation values were calculated. The latter were meant as estimates for how many times one would expect to achieve the first ranked score by chance. More rigorous validation methods were later developed for large MS/MS data sets (e.g., LC-MS/MS), which use either target-decoy or empirical Bayesian methods. It is now relatively common for proteomics researchers to estimate error rates by searching reversed or randomized databases along with the

targeted database [138]. The search results will then contain a number of matches to the randomized database, which are assumed to be false, and database search result scores can be matched to estimated error rates. Alternatively, the idea behind the TPP computer program Peptide Prophet [139] is that a histogram plot of the top scores for all spectra in a LC-MS/MS run (or any large collection of MS/MS spectra) is made from a composite of two distributions. The assumption is that there are two distributions – one for incorrectly identified spectra with a range of low scores and another for correctly identified spectra with a range of high scores. The mathematical best fit of two distributions is then used to determine error rates and probabilities. A good combination of the two approaches is to use the search results of a randomized database to model the distribution of scores for incorrect identifications, which is a concept that has been implemented within current versions of Peptide Prophet. In general, it is essential to have software that can objectively validate database search results, since expert reviews tend to vary with the physiological state of the expert (not to mention that some experts have delusions of adequacy).

The empirical Bayesian approach used by Peptide Prophet can incorporate additional information in order to modify the final probability determination. For example, those peptides that are formed with the anticipated tryptic cleavage specificity are more likely to be correct than those derived from completely nontryptic cleavages. High-mass-accuracy measurements of precursors permit a postsearch evaluation of how the calculated candidate sequence molecular weights cluster. Those sequences whose calculated molecular weights deviate by more than the average are less likely to be correct. Results from multiple search engines [140], presence of an anticipated motif (e.g., *N*-linked glycosylation), and HPLC retention times can all be included in a final probability determination to help with automated validation.

Shotgun bottom-up proteomics is intrinsically a peptide identification technique and protein identification can only be inferred from these identifications. At first, this step sounds like it should be simple, but database redundancies and protein homology often make it difficult to be certain. If a peptide sequence is shared between different proteins, how should one apportion peptide-spectrum match probabilities among the possible protein choices? In general, most protein validation software does this by using principles of parsimony to create the simplest and shortest protein list possible [141–143]. Although reality is rarely simple, this is really the only choice available.

1.4.2

***De Novo* Sequencing**

De novo sequencing refers to the process of deriving a peptide sequence directly from the MS/MS spectrum without recourse to any sequence database. Manual *de novo* sequencing can be mentally diverting (see <http://www.abrf.org/ResearchGroups/MassSpectrometry/EPosters/ms97quiz/abrfQuiz.html>); however, this is not practical when confronted with more than a handful of spectra. For larger numbers of

spectra one needs to use automated *de novo* sequencing programs. One of the first was Lutefisk [144], which has since been used to benchmark other *de novo* sequencing programs (PepNovo being a notable open-source example [145]). For a variety of reasons, deriving a single correct sequence exclusively from a MS/MS spectrum is often not possible, either manually or with a computer program:

- i) Some amino acids have identical or nearly identical masses – leucine and isoleucine, glutamine and lysine, and phenylalanine and oxidized methionine.
- ii) Cleavages may be absent between adjacent amino acids. Absence of cleavage between the first and second amino acids of a tryptic peptide is very common.
- iii) Some amino acids have the same mass as pairs of other amino acids (e.g., Gly–Gly is exactly the same as Asn).
- iv) If one can identify a series of ions whose mass differences delineate an amino acid sequence, it remains unclear whether the derived sequence is going from the N- to C-terminus or the other way around. In other words, it is not always clear whether a series of ions are all *b*- or *y*-type fragment ions.

For these reasons, *de novo* sequencing typically results in a short list of candidate sequences that each account for the data to varying degrees.

Why bother performing a *de novo* sequence determination when database search programs and validation tools are so fast and easy to use? Obviously, one reason would be if one was working with a species whose genome has not yet been sequenced [146]. Generally, this method involves generating a list of *de novo* sequences, and then submitting them to a homology search engine where the parameters have been optimized to account for the vagaries and problems associated with MS/MS-derived sequences (e.g., inability to distinguish leucine and isoleucine) [144]. A second reason for performing *de novo* sequencing is that it potentially provides further validation of a database search result. Database searching and *de novo* sequencing are quite orthogonal approaches, and agreement between the two should boost the likelihood of a correct identification [147]. A third reason is that one might be wondering about all of the unmatched spectra (often around 90%) in a typical LC-MS/MS experiment. For example, one could find that many peptides have been carbamylated due to bad urea or that the autosampler exhibits severe carryover problems from prior users studying a different species not present in the database that was searched. A fourth application of *de novo* sequencing is to help identify high-quality spectra, particularly ones that had not been matched to a database sequence. Finding the “sequenceable” spectra is the same as finding the “high-quality” spectra.

1.5

MS of Protein Structure, Folding, and Interactions

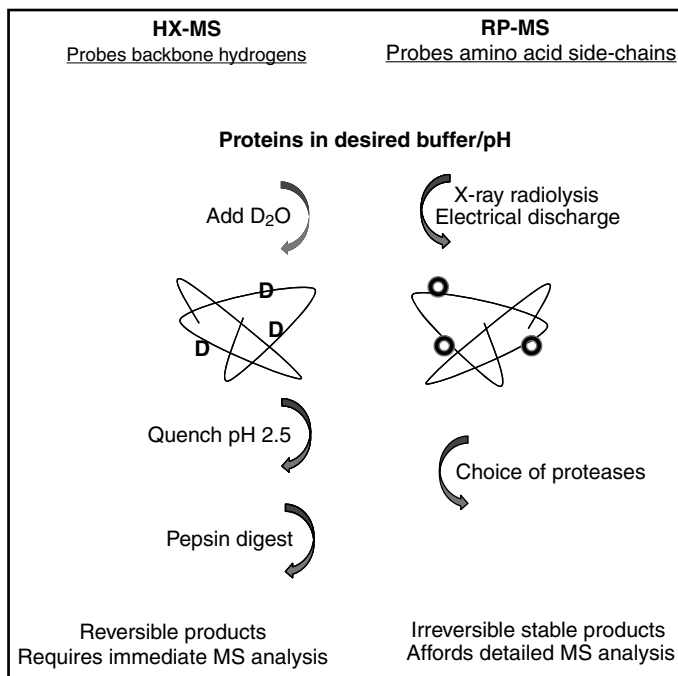
The study of structural dynamics of proteins and noncovalent interactions ideally requires analytical methods that enable capturing events occurring over timescale from nanoseconds to seconds, while simultaneously monitoring specific sites

within the structure of each participant. Current analytical techniques do not provide these capabilities individually. The fast timescale for monitoring macromolecular motions are achieved by spectroscopy-based methods, which provide global structural information. Other techniques such as the X-ray diffractometry or nuclear magnetic resonance (NMR) spectroscopy with capabilities of defining the location of individual atoms do not provide the fast timescale required to capture many events. The process of forming quality crystals, and the need for high-purity samples with good solubility properties further limit the applicability of X-ray and NMR for many studies. The need for solution-based structural characterization of proteins and protein interactions prompted considerable recent attention to the development of chemical probes combined with the benefits of MS analysis (i.e., high-throughput identification of proteins at low levels). A recent book [148] provides a thorough and updated review of a wide range of techniques applied for structural elucidation of proteins and their interactions, and a brief summary of two methods providing high structural resolution is presented below.

1.5.1

Methods to Mass-Tag Structural Features

Hydrogen–deuterium exchange MS (hydrogen–deuterium exchange mass spectrometry HX-MS) is routinely used to probe protein structure, conformation, and dynamics [149–151]. This method measures hydrogen atoms located at peptide amide linkages (i.e., backbone amide hydrogens) with an exchange half-life of seconds to several weeks depending on their solvent accessibility. The HX-MS approach starts with proteins in a physiological buffer at room temperature to conserve their native structure (Scheme 1.1). Deuterium oxide (D_2O) is added in excess of 10- to 20-fold and the onset of the exchange reaction is recorded. The exchange reaction is quenched at set time points by adjusting the pH to 2.5 (using an excess of protonated buffers) where the exchange rate drops to its minimum. Further reduction of the exchange rate is achieved by cooling the solution. The quenched timepoints can be snap-frozen in liquid nitrogen to be analyzed later, as no measureable back-exchange can be detected from frozen samples stored at $-70^\circ C$. At $0^\circ C$ and pH 2.5, the deuterium label reverts to hydrogen in a back-exchange process with a half-life of approximately 1 h, providing sufficient time for MS analysis. After the hydrogen exchange is completed and the reaction is quenched, the protein solution is analyzed by LC-ESI to measure the overall uptake of deuterium. A portion of protein is digested with pepsin to monitor the localized deuterium uptake. The incorporation of deuterium as a function of time is plotted for the intact protein as well as the proteolytic peptides. The main disadvantages of the HX-MS are that the exchanged products are reversible with a limited lifetime and the possibility of the back exchange reactions introduces some error in measurements. Further, the low pH condition of the quench reaction limits the choice of proteases. The resolution of hydrogen exchange is limited to the size of the peptic peptides that are generated, as residue resolution based on CID is not



Scheme 1.1 Experimental procedures for HX-MS and RP-MS.

possible due to proton scrambling [152]. However, there have been recent results indicating that scrambling does not occur when ETD is used for peptide fragmentation [153, 154], which could make it possible to measure hydrogen exchange rates for individual residues within a protein.

The protein structure could alternatively be evaluated through the solvent accessibility of the amino acid side-chains. The requirement for high structural resolution and the fast timescale of reactions prompted the use of hydroxyl radical as the ideal chemical probe that is similar in size to the water molecule with a diameter of 2.5 \AA^2 . Time-resolved hydroxyl radical protein footprinting employing MS was developed over a decade ago by applying synchrotron X-ray radiolysis [155, 156] or an electrical discharge source [157] to effect the oxidation of proteins on millisecond timescales. These approaches, which are referred to as radical probe MS (radical probe mass spectrometry RP-MS), have since been successfully applied to the analysis of protein structure, protein folding, and protein–protein interactions [158]. Hydroxyl radicals induce oxidative modification of a number of amino acid side-chains in the range of 10^9 to $10^{10} \text{ M}^{-1} \text{ s}^{-1}$, which is sufficiently fast for studies of protein folding and interaction dynamics. Further, the reactive hydroxyl radical probe originates in water at physiological pH without the need for other chemicals. Other advantages of the oxidative labeling are that the products are

stable, and this affords the use of a wide range of proteases and the application of a number of MS experiments.

The RP-MS approach identifies the site of amino acid oxidation and this information combined with the quantitative measure of the level of oxidation is used to map the solvent accessibility of the side-chains across a protein's surface. For structural and conformational studies, oxidation is kept to approximately 30–50% for the whole protein, in order to avoid forming degraded and cross-linked products. The timescale and the extent of reactions could then be used to monitor the onset of oxidative damage of proteins in relation with various diseases and aging. The application of RP-MS to studies of the onset of damage was first reported in 2005 for the protein α -crystallin, which demonstrated that different regions of a protein could exhibit different levels of susceptibility to oxidative damage [159]. These types of structural information are important in designing targeted therapeutics to prevent or control oxidative damage associated with a range of diseases.

A significant amount of information about a protein structure (i.e., solvent accessibility surface (SAS) of backbone hydrogen atoms and amino acid side-chains) is obtained through chemical labeling and MS protocols. These types of information prompted the development of a docking algorithm, PROXIMO, to propose structures for protein complexes based on those for their component molecules using RP-MS data [160]. The performance of the algorithm was successfully validated for a series of protein complexes, including the ribonuclease S-complex with several correctly identified conformers that deviated from the X-ray crystal structure with root mean square deviation values of 0.45 and 1.26 Å² (i.e., all within the 2.5 Å² SAS resolution of the RP-MS experimental structure). It could be envisioned that the application of chemical labeling in conjunction with computer algorithms will be valuable for solution-based structural analysis of proteins.

While the discovery of ESI has revolutionized the analysis of proteins and their noncovalently bound complexes, the question of whether or not the solution-based structures of proteins are preserved during the ESI process has been subject to numerous studies. A recent perspective [161] supports evidence for retention of native structure for some large proteins [162, 163]. Meanwhile, the authors propose that after the initial desolvation steps during the ESI process, structures of globular proteins of cytochrome *c* and ubiquitin undergo several transitions within picoseconds to seconds, which include collapse of the side-chains, unfolding and refolding steps that result in multiple conformers in the gas phase. Obviously, future studies for a range of proteins are required in order to validate this approach as a structural analysis tool.

As just discussed, various protein conformers could be generated during the ESI and these intermediate conformers are rarely isolated in the solution phase. Therefore, the gas-phase environment provides an ideal opportunity to investigate the subtle changes in protein structures during folding or binding transitions. Major advances in ion mobility MS (ion mobility mass spectrometry IMS) since the 1990s have made it possible to study small differences in structures of conformers in the gas phase based on their mobilities through a gas [164]. IMS has emerged as a powerful method for structural analysis of proteins and their complexes.

1.6

Conclusions and Perspectives

MS of proteins has made major strides over the past few decades. Of key importance was the development of new methods for the ionization of peptides and proteins (FAB, MALDI, and ESI), as well as new high-accuracy and high-resolution mass analyzers. Improvements in computer speed and data storage capacity, plus the rapid accumulation of protein and DNA sequence databases over this time was critical for enabling what has become known as “shotgun proteomics.” The latter was also dependent on enhanced understanding of gas-phase peptide ion fragmentation and software tools for matching mass spectral data to database-derived sequences. All of this led to a considerable amount of irrational exuberance (e.g., claims of sequencing the human proteome [165]), which has now largely subsided to a point where mostly what remains are serious people studying real problems. For proteomics, the next big area seems to be targeted proteomics, which has an improved dynamic range over shotgun analysis. However, targeted proteomics is also in danger of being excessively promoted and has a number of problems that need to be solved (e.g., how to estimate false discovery rates, how to handle targeted peptides that elute in more than one peak, how to target peptides that may or may not be modified, how to resolve contradictory quantitative results from different peptides from the same protein, etc.). In short, the shotgun proteomics wave has crashed on the beach, the targeted proteomics wave is coming, but regardless of the level of enthusiasm, these tools will continue to be useful for those who know their limitations and how to use them properly.

The identification of gene products is only one facet of proteomics. Identification and quantitation of PTMs is important for full characterization of a protein or proteome. Noncovalent structural aspects of proteins (folding, solvent accessibility, binding sites, etc.) can also be determined using MS. For the most part, PTM and noncovalent analysis is simply a matter of basic protein chemistry that has been considerably enabled by MS. For example, in the olden days protein chemists measured hydrogen exchange rates by measuring tritium incorporation; with mass spectrometers one can now more easily and safely measure the incorporation of the stable heavy isotope of hydrogen instead. Instead of determining disulfide bonds by comparing electrophoresis migration before and after bond cleavage, one now measures the molecular weights. To summarize, protein chemistry requires MS.

References

- 1 Langewiesche, W. (2005) The wrath of Khan. *The Atlantic Magazine*, (Nov), 62–85.
- 2 Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 17, 994–999.
- 3 Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular and Cellular Proteomics*, 1, 376–386.

- 4 Yao, X., Freas, A., Ramirez, J., Demirev, P.A., and Fenselau, C. (2001) Proteolytic ^{18}O labeling for comparative proteomics. *Analytical Chemistry*, **73**, 2836–2842.
- 5 Wool, A. and Smilansky, Z. (2002) Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting. *Proteomics*, **2**, 1365–1373.
- 6 Meng, C.K., Mann, M., and Fenn, J.B. (1988) Electrospray ionization of some polypeptides and small proteins. Proceedings 36th ASMS Conference, San Francisco, CA, pp. 771–772.
- 7 Mann, M., Meng, C.K., and Fenn, J.B. (1988) Parent mass information from sequences of peaks of multiply charged ions. Proceedings 36th ASMS Conference, San Francisco, CA, pp. 1207–1208.
- 8 Karas, M., Bachmann, D., and Hillenkamp, F. (1985) Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Analytical Chemistry*, **57**, 2935–2939.
- 9 Karas, M., Bachmann, D., Bahr, U., and Hillenkamp, F. (1987) Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International Journal of Mass Spectrometry and Ion Processes*, **78**, 53–68.
- 10 Karas, M. and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, **60**, 2299–2301.
- 11 Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T. (1988) Protein and polymer analyses up to m/z 100,000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, **2**, 151–153.
- 12 McLafferty, F.W., Todd, P.J., McGilvery, D.C., and Baldwin, M.A. (1980) High-resolution tandem mass spectrometry (MS/MS) of increased sensitivity and mass range. *Journal of the American Chemical Society*, **102**, 3360–3363.
- 13 Yost, R.A. and Enke, C.G. (1978) Selected ion fragmentation with a tandem quadrupole mass spectrometer. *Journal of the American Chemical Society*, **100**, 2274–2275.
- 14 Douglas, D.J., Frank, A.J., and Mao, D. (2004) Linear ion traps in mass spectrometry. *Mass Spectrometry Reviews*, **24**, 1–29.
- 15 March, R.E. (1997) An introduction to quadrupole ion trap mass spectrometry. *Journal of Mass Spectrometry*, **32**, 351–369.
- 16 Marshall, A.G., Hendrickson, C.L., and Jackson, G.S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrometry Reviews*, **17**, 1–35.
- 17 Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical Chemistry*, **72**, 1156–1162.
- 18 Perry, R.H., Cooks, R.G., and Noll, R.J. (2008) Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrometry Reviews*, **27**, 661–699.
- 19 Cotter, R.J. (1994) Time-of-flight mass spectrometry, in *Basic Principles and Current State*, American Chemical Society, Columbus, OH, pp. 16–48.
- 20 Vestal, M.L. and Campbell, J.M. (2005) Tandem time-of-flight mass spectrometry. *Methods in Enzymology*, **402**, 79–108.
- 21 Morris, H.R., Paxton, T., Panico, M., McDowell, R., and Dell, A. (1997) A novel geometry mass spectrometer, the Q-TOF, for low-femtomole/attomole-range biopolymer sequencing. *Journal of Protein Chemistry*, **16**, 469–479.
- 22 Hager, J.W. and Yves Le Blanc, J.C. (2003) Product ion scanning using a Q-q-Q_{linear} ion trap (QTRAPTM) mass spectrometer. *Rapid Communications in Mass Spectrometry*, **17**, 1056–1064.
- 23 Clauser, K.R., Baker, P., and Burlingame, A.L. (1999) Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry*, **71**, 2871–2882.
- 24 Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999) Probability-based protein identification by searching

- sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
- 25 Eng, J.K., McCormack, A.L., and Yates, J.R. III, (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, **5**, 976–989.
 - 26 Lipton, M.S., Pasa-Tolic, L., Anderson, G.A., Anderson, D.J., Auberry, D.L., Battista, J.R., Daly, M.J., Fredrickson, J., Hixson, K.K., Kostandarithes, H., Masselon, C., Markillie, L.M., Moore, R.J., Romine, M.F., Shen, Y., Strittmatter, E., Tolic, N., Udseth, H.R., Venkateswaran, A., Wong, K.-K., Zhao, R., and Smith, R.D. (2002) Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11049–11054.
 - 27 Washburn, M.P., Wolters, D., and Yates, J.R. III, (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, **19**, 242–247.
 - 28 Schwahn, A.B., Wong, J.W.H., and Downard, K.M. (2009) Subtyping of the influenza virus by high resolution mass spectrometry. *Analytical Chemistry*, **81**, 3500–3506.
 - 29 Schwahn, A.B., Wong, J.W.H., and Downard, K.M. (2010) Rapid differentiation of seasonal and pandemic H1N1 influenza through proteotyping of viral neuraminidase with mass spectrometry. *Analytical Chemistry*, **82**, 4584–4590.
 - 30 Covey, T.R., Bonner, R.F., Shushan, B.I., Henion, J., and Boyd, R.K. (1988) The determination of protein, oligonucleotide and peptide molecular weights by ion-spray mass spectrometry. *Rapid Communications in Mass Spectrometry*, **2**, 249–256.
 - 31 Mann, M., Meng, C.K., and Fenn, J.B. (1989) Interpreting mass spectra of multiply charged ions. *Analytical Chemistry*, **61**, 1702–1708.
 - 32 Senko, M.W., Beu, S.C., and McLafferty, F.W. (1995) Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *Journal of the American Society for Mass Spectrometry*, **6**, 52–56.
 - 33 Zhang, Z. and Marshall, A.G. (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *Journal of the American Society for Mass Spectrometry*, **9**, 225–233.
 - 34 Horn, D.M., Zubarev, R.A., and McLafferty, F.W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, **11**, 320–332.
 - 35 Senko, M.W., Beu, S.C., and McLafferty, F.W. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, **6**, 229–233.
 - 36 Chen, L., Sze, S.K., and Yang, H. (2006) Automated intensity descent algorithm for interpretation of complex high-resolution mass spectra. *Analytical Chemistry*, **78**, 5006–5018.
 - 37 Chen, L. and Yap, Y.L. (2008) Automated charge state determination of complex isotope-resolved mass spectra by peak-target Fourier transform. *Journal of the American Society for Mass Spectrometry*, **19**, 46–54.
 - 38 Maleknia, S.D. and Downard, K.M. (2005) Charge ratio analysis method: approach for the deconvolution of electrospray mass spectra. *Analytical Chemistry*, **77**, 111–119.
 - 39 Maleknia, S.D. and Downard, K.M. (2005) Charge ratio analysis method to interpret high resolution electrospray Fourier transform-ion cyclotron resonance mass spectra. *International Journal of Mass Spectrometry*, **246**, 1–9.
 - 40 Maleknia, S.D. and Green, D.C. (2010) eCRAM computer algorithm for implementation of the charge ratio analysis method to deconvolute electrospray ionization mass spectra. *International Journal of Mass Spectrometry*, **290**, 1–8.
 - 41 Bowie, J.H., Brinkworth, C.S., and Dua, S. (2002) Collision-induced

- fragmentations of the $(M - H)^-$ parent anions of underivatized peptides: an aid to structure determination and some unusual negative ion cleavages. *Mass Spectrometry Reviews*, **21**, 87–107.
- 42 Papayannopoulos, I.A. (1995) The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrometry Reviews*, **14**, 49–73.
- 43 Roepstorff, P. and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry*, **11**, 601.
- 44 Biemann, K. (1990) Appendix 5. Nomenclature for peptide fragment ions (positive ions). *Methods in Enzymology*, **193**, 886–887.
- 45 Schlosser, A. and Lehmann, W.D. (2000) Five-membered ring formation in unimolecular reactions of peptides: a key structural element controlling low-energy collision-induced dissociation of peptides. *Journal of Mass Spectrometry*, **35**, 1382–1390.
- 46 Wysocki, V.H., Tsapralis, G., Smith, L.L., and Breci, L.A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, **35**, 1399–1406.
- 47 Yu, W., Vath, J.E., Huberty, M.C., and Martin, S.A. (1993) Identification of the facile gas-phase cleavage of the Asp-Pro and Asp-Xxx peptide bonds in matrix-assisted laser desorption time-of-flight mass spectrometry. *Analytical Chemistry*, **65**, 3015–3023.
- 48 Johnson, R.S., Martin, S.A., and Biemann, K. (1988) Collision-induced fragmentation of $(M + H)^+$ ions of peptides. Side chain specific sequence ions. *International Journal of Mass Spectrometry and Ion Processes*, **86**, 137–154.
- 49 Johnson, R.S., Martin, S.A., Biemann, K., Stults, J.T., and Watson, J.T. (1987) Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Analytical Chemistry*, **59**, 2621–2625.
- 50 Johnson, R.S. and Biemann, K. (1987) The primary structure of thioredoxin from *Chromatium vinosum* determined by high-performance tandem mass spectrometry. *Biochemistry*, **26**, 1209–1214.
- 51 Zubarev, R.A., Kelleher, N.L., and McLafferty, F.W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society*, **120**, 3265–3266.
- 52 Stensballe, A., Jensen, O.N., Olsen, J.V., Haselmann, K.F., and Zubarev, R.A. (2000) Electron capture dissociation of singly and multiply phosphorylated peptides. *Rapid Communications in Mass Spectrometry*, **14**, 1793–1800.
- 53 Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J., and Hunt, D.F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9528–9533.
- 54 Chi, A., Huttenhower, C., Geer, L.Y., Coon, J.J., Syka, J.E.P., Bai, D.L., Shabanowitz, J., Burke, D.J., Troyanskaya, O.G., and Hunt, D.F. (2007) Analysis of phosphorylation sites on proteins from *Saccharomyces cerevisiae* by electron transfer dissociation (ETD) mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 2193–2198.
- 55 Good, D.M., Wirtala, M., McAlister, G.C., and Coon, J.J. (2007) Performance characteristics of electron transfer dissociation mass spectrometry. *Molecular and Cellular Proteomics*, **6**, 1942–1951.
- 56 Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature*, **379**, 466–469.
- 57 Furmanek, A. and Hofsteenge, J. (2000) Protein C-mannosylation: facts and questions. *Acta Biochimica Polonica*, **47**, 781–789.

- 58 Golks, A. and Guerini, D. (2008) The O-linked N-acetylglucosamine modification in cellular signalling and the immune system. "Protein modifications: beyond the usual suspects" review series. *EMBO Reports*, **9**, 748–753.
- 59 Kreppel, L.K., Blomberg, M.A., and Hart, G.W. (1997) Dynamic glycosylation of nuclear and cytosolic proteins: cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats. *Journal of Biological Chemistry*, **272**, 9308–9315.
- 60 Chalkley, R.J., Thalhammer, A., Schoepfer, R., and Burlingame, A.L. (2009) Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 8894–8899.
- 61 Dell, A. and Morris, H.R. (2001) Glycoprotein structure determination by mass spectrometry. *Science*, **291**, 2351–2356.
- 62 Helenius, A. and Aebi, M. (2001) Intracellular functions of N-linked glycans. *Science*, **291**, 2364–2369.
- 63 Robinson, N.E., Robinson, Z.W., Robinson, B.R., Robinson, A.L., Robinson, J.A., Robinson, M.L., and Robinson, A.B. (2004) Structure-dependent nonenzymatic deamidation of glutaminyl and asparaginyl pentapeptides. *Journal of Peptide Research*, **63**, 426–436.
- 64 Zaia, J. (2010) Mass spectrometry and glycomics. *Omic*s, **14**, 401–418.
- 65 Alvarez-Manilla, G., Warren, N.L., Atwood, J., Orlando, R., Dalton, S., and Pierce, M. (2010) Glycoproteomic analysis of embryonic stem cells: identification of potential glycomarkers using lectin affinity chromatography of glycopeptides. *Journal of Proteome Research*, **9**, 2062–2075.
- 66 Calvano, C.D., Zamboni, C.G., and Jensen, O.N. (2008) Assessment of lectin and HILIC based enrichment protocols for characterization of serum glycoproteins by mass spectrometry. *Journal of Proteomics*, **71**, 304–317.
- 67 Kaji, H., Saito, H., Yamauchi, Y., Shinkawa, T., Taoka, M., Hirabayashi, J., Kasai, K-ichi., Takahashi, N., and Isobe, T. (2003) Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nature Biotechnology*, **21**, 667–672.
- 68 McDonald, C.A., Yang, J.Y., Marathe, V., Yen, T.-Y., and Macher, B.A. (2009) Combining results from lectin affinity chromatography and glyco capture approaches substantially improves the coverage of the glycoproteome. *Molecular and Cellular Proteomics*, **8**, 287–301.
- 69 Maleknia, S.D., Treuheit, M.J., Carlson, J.E., Halsall, H.B., and Costello, C.E. (1993) Analysis of heterogeneity of native a1-acid glycoprotein by MADLI-TOFMS. Proceedings of the 41st ASMS Conference, San Francisco, CA, pp. 81a–81b.
- 70 Wollscheid, B., Bausch-Fluck, D., Henderson, C., O'Brien, R., Bibel, M., Schiess, R., Aebersold, R., and Watts, J.D. (2009) Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nature Biotechnology*, **27**, 378–386.
- 71 Zhang, H., Li, X.-j., Martin, D.B., and Aebersold, R. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nature Biotechnology*, **21**, 660–666.
- 72 Palumbo, A.M. and Reid, G.E. (2008) Evaluation of gas-phase rearrangement and competing fragmentation reactions on protein phosphorylation site assignment using collision induced dissociation-MS/MS and MS³. *Analytical Chemistry*, **80**, 9735–9747.
- 73 Alpert, A.J. (2008) Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides. *Analytical Chemistry*, **80**, 62–76.
- 74 Ficarro, S.B., McClelland, M.L., Stukenberg, P.T., Burke, D.J.,

- Ross, M.M., Shabanowitz, J., Hunt, D.F., and White, F.M. (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nature Biotechnology*, **20**, 301–305.
- 75 McNulty, D.E. and Annan, R.S. (2008) Hydrophilic interaction chromatography reduces the complexity of the phosphoproteome and improves global phosphopeptide isolation and detection. *Molecular and Cellular Proteomics*, **7**, 971–980.
- 76 Tsai, C.-F., Wang, Y.-T., Chen, Y.-R., Lai, C.-Y., Lin, P.-Y., Pan, K.-T., Chen, J.-Y., Khoo, K.-H., and Chen, Y.-J. (2008) Immobilized metal affinity chromatography revisited: pH/acid control toward high selectivity in phosphoproteomics. *Journal of Proteome Research*, **7**, 4058–4069.
- 77 Villén, J. and Gygi, S.P. (2008) The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nature Protocols*, **3**, 1630–1638.
- 78 Welchman, R.L., Gordon, C., and Mayer, R.J. (2005) Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nature Reviews Molecular Cell Biology*, **6**, 599–609.
- 79 Peng, J. (2008) Evaluation of proteomic strategies for analyzing ubiquitinated proteins. *BMB Reports*, **41**, 177–183.
- 80 Peng, J., Schwartz, D., Elias, J.E., Thoreen, C.C., Cheng, D., Marsischky, G., Roelofs, J., Finley, D., and Gygi, S.P. (2003) A proteomics approach to understanding protein ubiquitination. *Nature Biotechnology*, **21**, 921–926.
- 81 Matic, I., van Hagen, M., Schimmel, J., Macek, B., Ogg, S.C., Tatham, M.H., Hay, R.T., Lamond, A.I., Mann, M., and Vertegaal, A.C.O. (2008) *In vivo* identification of human small ubiquitin-like modifier polymerization sites by high accuracy mass spectrometry and an *in vitro* to *in vivo* strategy. *Molecular and Cellular Proteomics*, **7**, 132–144.
- 82 Kim, S.C., Sprung, R., Chen, Y., Xu, Y., Ball, H., Pei, J., Cheng, T., Kho, Y., Xiao, H., Xiao, L., Grishin, N.V., White, M., Yang, X.-J., and Zhao, Y. (2006) Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. *Molecular Cell*, **23**, 607–618.
- 83 Choudhary, C., Kumar, C., Gnad, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V., and Mann, M. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, **325**, 834–840.
- 84 Ryle, A.P. and Sanger, F. (1955) Disulphide interchange reactions. *Biochemical Journal*, **60**, 535–540.
- 85 Mikesch, L., Ueberheide, B., Chi, A., Coon, J.J., Syka, J.E.P., Shabanowitz, J., and Hunt, D.F. (2006) The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta*, **1764**, 1811–1822.
- 86 Chrisman, P.A., Pitteri, S.J., Hogan, J.M., and McLuckey, S.A. (2005) SO_2^{-} electron transfer ion/ion reactions with disulfide linked polypeptide ions. *Journal of the American Society for Mass Spectrometry*, **16**, 1020–1030.
- 87 Black, R.A., Rauch, C.T., Kozlosky, C.J., Peschon, J.J., Slack, J.L., Wolfson, M.F., Castner, B.J., Stocking, K.L., Reddy, P., Srinivasan, S., Nelson, N., Boiani, N., Schooley, K.A., Gerhart, M., Davis, R., Fitzner, J.N., Johnson, R.S., Paxton, R.J., March, C.J., and Cerretti, D.P. (1997) A metalloproteinase disintegrin that releases tumour-necrosis factor- α from cells. *Nature*, **385**, 729–733.
- 88 Guo, L., Eisenman, J.R., Mahimkar, R.M., Peschon, J.J., Paxton, R.J., and Black, R.A., and Johnson, R.S. (2002) A proteomic approach for the identification of cell-surface proteins shed by metalloproteases. *Molecular and Cellular Proteomics*, **1**, 30–36.
- 89 Abrahmsén, L., Tom, J., Burnier, J., Butcher, K.A., Kossiakoff, A., and Wells, J.A. (1991) Engineering subtilisin and its substrates for efficient ligation of peptide bonds in aqueous solution. *Biochemistry*, **30**, 4151–4159.
- 90 Staes, A., Van Damme, P., Helsen, K., Demol, H., Vandekerckhove, J., and Gevaert, K. (2008) Improved recovery of proteome-informative, protein

- N-terminal peptides by combined fractional diagonal chromatography (COFRADIC). *Proteomics*, **8**, 1362–1370.
- 91 Agard, N.J., Maltby, D., and Wells, J.A. (2010) Inflammatory stimuli regulate caspase substrate profiles. *Molecular and Cellular Proteomics*, **9**, 880–893.
- 92 Schilling, O., Barré, O., Huesgen, P.F., and Overall, C.M. (2010) Proteome-wide analysis of protein carboxy termini: C terminomics. *Nature Methods*, **7**, 508–511.
- 93 Samyn, B., Sergeant, K., Castanheira, P., Faro, C., and Van Beeumen, J. (2005) A new method for C-terminal sequence analysis in the proteomic era. *Nature Methods*, **2**, 193–200.
- 94 Kleifeld, O., Doucet, A., auf dem Keller, U., Prudova, A., Schilling, O., Kainthan, R.K., Starr, A.E., Foster, L.J., Kizhakkedathu, J.N., and Overall, C.M. (2010) Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nature Biotechnology*, **28**, 281–288.
- 95 Stark, G.R., Stein, W.H., and Moore, S. (1960) Reactions of the cyanate present in aqueous urea with amino acids and proteins. *Journal of Biological Chemistry*, **235**, 3177–3181.
- 96 Swiderek, K.M., Davis, M.T., and Lee, T.D. (1998) The identification of peptide modifications derived from gel-separated proteins using electrospray triple quadrupole and ion trap analyses. *Electrophoresis*, **19**, 989–997.
- 97 Perdivara, I., Deterding, L.J., Przybylski, M., and Tomer, K.B. (2010) Mass spectrometric identification of oxidative modifications of tryptophan residues in proteins: chemical artifact or post-translational modification? *Journal of the American Society for Mass Spectrometry*, **21**, 1114–1117.
- 98 Cohen, S.L. (2006) Ozone in ambient air as a source of adventitious oxidation. A mass spectrometric study. *Analytical Chemistry*, **78**, 4352–4362.
- 99 Hirs, C.H.W. (1956) The oxidation of ribonuclease with performic acid. *Journal of Biological Chemistry*, **219**, 611–621.
- 100 Khandke, K.M., Fairwell, T., Chait, B.T., and Manjula, B.N. (1989) Influence of ions on cyclization of the amino terminal glutamine residues of tryptic peptides of streptococcal PepM49 protein: resolution of cyclized peptides by HPLC and characterization by mass spectrometry. *International Journal of Peptide and Protein Research*, **34**, 118–123.
- 101 Sanger, F., Thompson, E.O.P., and Kitai, R. (1955) The amide groups of insulin. *Biochemical Journal*, **59**, 509–518.
- 102 Geoghegan, K.F., Hoth, L.R., Tan, D.H., Borzilleri, K.A., Withka, J.M., and Boyd, J.G. (2002) Cyclization of N-terminal S-carbamoylmethylcysteine causing loss of 17 Da from peptides and extra peaks in peptide maps. *Journal of Proteome Research*, **1**, 181–187.
- 103 Geiger, T. and Clarke, S. (1987) Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimidyl-linked reactions that contribute to protein degradation. *Journal of Biological Chemistry*, **262**, 785–794.
- 104 Williams, J.D., Flanagan, M., Lopez, L., Fischer, S., and Miller, L.A.D. (2003) Using accurate mass electrospray ionization-time-of-flight mass spectrometry with in-source collision-induced dissociation to sequence peptide mixtures. *Journal of Chromatography A*, **1020**, 11–26.
- 105 Russell, D.H., McGlohon, E.S., and Mallis, L.M. (1988) Fast-atom bombardment-tandem mass spectrometry studies of organo-alkali-metal ions of small peptides. Competitive interaction of sodium with basic amino acid substituents. *Analytical Chemistry*, **60**, 1818–1824.
- 106 Grese, R.P., Cerny, R.L., and Gross, M.L. (1989) Metal ion-peptide interactions in the gas phase: a tandem mass spectrometry study of alkali metal cationized peptides. *Journal of the American Chemical Society*, **111**, 2835–2842.
- 107 Lee, S.-W., Kim, H.S., and Beauchamp, J.L. (1998) Salt bridge chemistry applied to gas-phase peptide sequencing: selective fragmentation of

- sodiated gas-phase peptide ions adjacent to aspartic acid residues. *Journal of the American Chemical Society*, **120**, 3188–3195.
- 108 Picotti, P., Aebersold, R., and Domon, B. (2007) The implications of proteolytic background for shotgun proteomics. *Molecular and Cellular Proteomics*, **6**, 1589–1598.
- 109 Sze, S.K., Ge, Y., Oh, H., and McLafferty, F.W. (2002) Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 1774–1779.
- 110 Karabacak, N.M., Li, L., Tiwari, A., Hayward, L.J., Hong, P., Easterling, M.L., and Agar, J.N. (2009) Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. *Molecular and Cellular Proteomics*, **8**, 846–856.
- 111 Yi, E.C., Marelli, M., Lee, H., Purvine, S.O., Aebersold, R., Aitchison, J.D., and Goodlett, D.R. (2002) Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis*, **23**, 3205–3216.
- 112 Cargile, B.J., Talley, D.L., and Stephenson, J.L. (2004) Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides. *Electrophoresis*, **25**, 936–945.
- 113 Davis, M.T. and Lee, T.D. (1997) Variable flow liquid chromatography-tandem mass spectrometry and the comprehensive analysis of complex protein digest mixtures. *Journal of the American Society for Mass Spectrometry*, **8**, 1059–1069.
- 114 Moritz, R. (2010) The mechanics of detecting and quantifying the complete human proteome. Human Proteome World Congress 2010, Sydney, abstract OS073
- 115 Picotti, P., Lam, H., Campbell, D., Deutsch, E.W., Mirzaei, H., Ranish, J., Domon, B., and Aebersold, R. (2008) A database of mass spectrometric assays for the yeast proteome. *Nature Methods*, **5**, 913–914.
- 116 Rodriguez, J., Gupta, N., Smith, R.D., and Pevzner, P.A. (2008) Does trypsin cut before proline? *Journal of Proteome Research*, **7**, 300–305.
- 117 Jekel, P.A., Weijer, W.J., and Beintema, J.J. (1983) Use of endoproteinase Lys-C from *Lysobacter enzymogenes* in protein sequence analysis. *Analytical Biochemistry*, **134**, 347–354.
- 118 Taouatas, N., Heck, A.J.R., and Mohammed, S. (2010) Evaluation of metalloendopeptidase Lys-N protease performance under different sample handling conditions. *Journal of Proteome Research*, **9**, 4282–4288.
- 119 Hohmann, L., Sherwood, C., Eastham, A., Peterson, A., Eng, J.K., Eddes, J.S., Shteynberg, D., and Martin, D.B. (2009) Proteomic analyses using *Grifola frondosa* metalloendoprotease Lys-N. *Journal of Proteome Research*, **8**, 1415–1422.
- 120 Drapeau, G.R., Boily, Y., and Houmar, J. (1972) Purification and properties of an extracellular protease of *Staphylococcus aureus*. *Journal of Biological Chemistry*, **247**, 6720–6726.
- 121 Tomer, K.B., Moseley, M.A., and Deterding, L.J., and Parker, C.E. (1994) Capillary liquid chromatography/mass spectrometry. *Mass Spectrometry Reviews*, **13**, 431–457.
- 122 Wetterhall, M., Palmblad, M., Håkansson, P., Markides, K.E., and Bergquist, J. (2002) Rapid analysis of tryptically digested cerebrospinal fluid using capillary electrophoresis-electrospray ionization-Fourier transform ion cyclotron resonance-mass spectrometry. *Journal of Proteome Research*, **1**, 361–366.
- 123 Maleknia, S.D., Mark, J.P., Dixon, J.D., Elicone, C.P., McGuinness, B.F., Fulton, S.P., and Afeyan, N.B. (1994) Real-time protein mapping utilizing immobilized enzyme columns. Proceedings of the 42nd ASMS Conference, Chicago, IL, pp. 304–305.
- 124 Peters, E.C., Brock, A., Horn, D.M., Phung, Q.T., Ericson, C., Salomon, A.R., Ficarro, S.B., and Brill, L.M. (2002) An

- automated LC-MALDI FT-ICR MS platform for high-throughput proteomics. *LC GC Europe*, **15**, 423–428.
- 125 Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C., and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 5011–5015.
- 126 Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J.K., Martin, D.B., Nesvizhskii, A.I., and Aebersold, R. (2010) A guided tour of the Trans-Proteomic Pipeline. *Proteomics*, **10**, 1150–1159.
- 127 Pedrioli, P.G.A., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R., Cheung, K., Costello, C.E., Hermjakob, H., Huang, S., Julian, R.K., Kapp, E., McComb, M.E., Oliver, S.G., Omenn, G., Paton, N.W., Simpson, R., Smith, R., Taylor, C.F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology*, **22**, 1459–1466.
- 128 Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*, **24**, 2534–2536.
- 129 Renard, B.Y., Kirchner, M., Monigatti, F., Ivanov, A.R., Rappsilber, J., Winter, D., Steen, J.A.J., Hamprecht, F.A., and Steen, H. (2009) When less can yield more – computational preprocessing of MS/MS spectra for peptide identification. *Proteomics*, **9**, 4978–4984.
- 130 Cox, J. and Mann, M. (2009) Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. *Journal of the American Society for Mass Spectrometry*, **20**, 1477–1485.
- 131 Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics*, **3**, 1454–1463.
- 132 Craig, R. and Beavis, R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.
- 133 Tabb, D.L., Fernando, C.G., and Chambers, M.C. (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research*, **6**, 654–661.
- 134 Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., and Bryant, S.H. (2004) Open mass spectrometry search algorithm. *Journal of Proteome Research*, **3**, 958–964.
- 135 Tanner, S., Shu, H., Frank, A., Wang, L.-C., Zandi, E., Mumby, M., Pevzner, P.A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry*, **77**, 4626–4639.
- 136 Maleknia, S.D. (1996) Sequencing isobaric peptides by the application of MS^{nm} analysis on an ion trap mass spectrometer. Proceedings of the 44th ASMS Conference, Portland, OR, pp. 703–704.
- 137 Nesvizhskii, A.I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods*, **4**, 787–797.
- 138 Elias, J.E. and Gygi, S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, **4**, 207–214.
- 139 Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, **74**, 5383–5392.
- 140 Searle, B.C. (2010) Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics*, **10**, 1265–1269.
- 141 Zhang, B., Chambers, M.C., and Tabb, D.L. (2007) Proteomic parsimony

- through bipartite graph analysis improves accuracy and transparency. *Journal of Proteome Research*, **6**, 3549–3557.
- 142 Nesvizhskii, A.I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, **75**, 4646–4658.
- 143 Feng, J., Naiman, D.Q., and Cooper, B. (2007) Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data. *Analytical Chemistry*, **79**, 3901–3911.
- 144 Taylor, J.A. and Johnson, R.S. (1997) Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, **11**, 1067–1075.
- 145 Frank, A. and Pevzner, P. (2005) PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, **77**, 964–973.
- 146 Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W., and Standing, K.G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Analytical Chemistry*, **73**, 1917–1926.
- 147 Taylor, J.A. and Johnson, R.S. (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Analytical Chemistry*, **73**, 2594–2604.
- 148 Downard, K.M. (ed.) (2007) *Mass Spectrometry of Protein Interactions*, Wiley Interscience Series on Mass Spectrometry, John Wiley & Sons, Inc., Hoboken, NJ.
- 149 Zhang, Z. and Smith, D.L. (1993) Determination of amide hydrogen exchange by mass spectrometry: A new tool for protein structure elucidation. *Protein Science*, **2**, 522–531.
- 150 Smith, D.L., Deng, Y., and Zhang, Z. (1997) Probing the non-covalent structure of proteins by amide hydrogen exchange and mass spectrometry. *Journal of Mass Spectrometry*, **32**, 135–146.
- 151 Wales, T.E. and Engen, J.R. (2006) Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrometry Reviews*, **25**, 158–170.
- 152 Johnson, R.S., Krylov, D., and Walsh, K.A. (1995) Proton mobility within electrosprayed peptide ions. *Journal of Mass Spectrometry*, **30**, 386–387.
- 153 Rand, K.D., Zehl, M., Jensen, O.N., and Jørgensen, T.J.D. (2009) Protein hydrogen exchange measured at single-residue resolution by electron transfer dissociation mass spectrometry. *Analytical Chemistry*, **81**, 5577–5584.
- 154 Pan, J., Han, J., Borchers, C.H., and Konermann, L. (2008) Electron capture dissociation of electrosprayed protein ions for spatially resolved hydrogen exchange measurements. *Journal of the American Chemical Society*, **130**, 11574–11575.
- 155 Maleknia, S.D., Brenowitz, M., and Chance, M.R. (1999) Millisecond radiolytic modification of peptides by synchrotron X-rays identified by mass spectrometry. *Analytical Chemistry*, **71**, 3965–3973.
- 156 Maleknia, S.D., Ralston, C.Y., Brenowitz, M.D., Downard, K.M., and Chance, M.R. (2001) Determination of macromolecular folding and structure by synchrotron x-ray radiolysis techniques. *Analytical Biochemistry*, **289**, 103–115.
- 157 Maleknia, S.D., Chance, M.R., and Downard, K.M. (1999) Electrospray-assisted modification of proteins: a radical probe of protein structure. *Rapid Communications in Mass Spectrometry*, **13**, 2352–2358.
- 158 Maleknia, S.D. and Downard, K.M. (2001) Radical approaches to probe protein structure, folding, and interactions by mass spectrometry. *Mass Spectrometry Reviews*, **20**, 388–401.
- 159 Shum, W.-K., Maleknia, S.D., and Downard, K.M. (2005) Onset of oxidative damage in α -crystallin by radical probe mass spectrometry. *Analytical Biochemistry*, **344**, 247–256.
- 160 Gerega, S.K. and Downard, K.M. (2006) PROXIMO – a docking new algorithm to model protein complexes using data from

- radical probe mass spectrometry. *Bioinformatics*, **22**, 1702–1709.
- 161** Breuker, K. and McLafferty, F. (2008) Stepwise evolution of protein native structure with electrospray into the gas phase, 10^{12} to 10^2 s. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 18145–18152.
- 162** Heck, A.J.R. and van den Heuvel, R.H.H. (2004) Investigation of intact protein complexes by mass spectrometry. *Mass Spectrometry Reviews*, **23**, 368–389.
- 163** Benesch, J.L.P. and Robinson, C.V.R. (2006) Mass spectrometry of macromolecular assemblies: preservation and dissociation. *Current Opinion Structure Biology*, **16**, 245–251.
- 164** Bohrer, B.C., Merenbloom, S.I., Koeniger, S.L., Hilderbrand, A.E., and Clemmer, D.E. (2008) Biomolecule analysis by ion mobility mass spectrometry. *Annual Reviews of Analytical Chemistry*, **1**, 293–327.
- 165** Service, R.F. (2000) Proteomics: can Celera do it again? *Science*, **287**, 2136–2138.