Chapter 6

Understanding and Exploiting Peptide Fragment Ion Intensities Using Experimental and Informatic Approaches*

Ashley C. Gucinski, Eric D. Dodds, Wenzhou Li, and Vicki H. Wysocki

Abstract

Tandem mass spectrometry is a widely used tool in proteomics. This section will address the properties that describe how protonated peptides fragment when activated by collisions in a mass spectrometer and how that information can be used to identify proteins. A review of the mobile proton model is presented, along with a summary of commonly observed peptide cleavage enhancements, including the proline effect. The methods used to elucidate peptide dissociation chemistry by using both small groups of model peptides and large datasets are also discussed. Finally, the role of peak intensity in commercially available and developmental peptide identification algorithms is examined.

Key words: Peptide fragmentation, Data mining, Tandem mass spectrometry, Mobile proton model, Intensity-based algorithms, Dissociation pattern, Intensity, Statistical analysis

1. Introduction

Mass spectrometry (MS), which allows for measurement of peptide, protein, and fragment ion mass-to-charge ratios (*m/z*), is widely used in studies that aim to identify peptides and proteins. Often, these studies involve high-throughput, large-scale identification of proteins from complex mixtures (1, 2). MS is expected to continue serving an important function in this arena for many years to come due to the sensitivity, selectivity, and speed of MS-based analyses (3). The further optimization and enhancement of MS

* This chapter is dedicated to the memory of Katheryn A. Resing: colleague, collaborator and friend, who left us January 8th, 2009 after a courageous battle with cancer.

Simon J. Hubbard and Andrew R. Jones (eds.), *Proteome Bioinformatics*, Methods in Molecular Biology, vol. 604
DOI 10.1007/978-1-60761-444-9_6, © Humana Press, a part of Springer Science+Business Media, LLC 2010

technology and data analysis capabilities for proteomics remain a highly active area of research (4–6).

While single stage mass spectrometry does play a role in protein identification, many protein identifications are performed by tandem mass spectrometry (MS/MS) of peptides derived from protein digests (7–9). In a common "bottom-up" MS/MS approach to proteomics for large-scale protein identification, peptides are produced by enzymatic digestion of a mixture of proteins. The specificity of the protease determines the sites at which peptide bonds are hydrolyzed and thus dictates the numbers, lengths, and terminal residue identities of peptides produced from a given protein. The peptides produced by digestion of a mixture of proteins are commonly separated by one or two stages of high-performance liquid chromatography (HPLC), ionized (typically by electrospray ionization, ESI) (10), and mass-selected for MS/MS fragmentation analysis. After peptide ion activation and subsequent dissociation, product ions are analyzed by *m/z* and relative intensity. This MS/MS spectral information must then be converted into peptide sequence information and in turn, protein identification. A schematic for this process is shown in Fig. 1.

Several algorithms are available that perform peptide sequencing and protein identification from MS/MS data (11–14), and additional software tools have been developed to help users consolidate and interpret database search results (15). These various protein identification algorithms have differing success rates, and current algorithms assign sequence matches to only a minority of acquired spectra. Therefore, it would be appealing to obtain sequence matches for a larger percentage of peptide spectra submitted to a given algorithm. This would allow additional proteins to be identified from a given dataset and would also

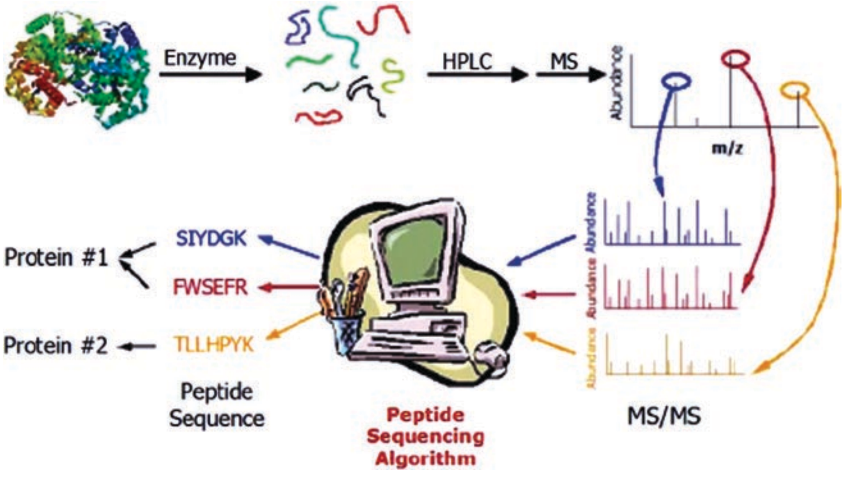

Fig. 1. Schematic of tandem mass spectrometry based protein analysis

provide a larger number of matching peptides per identified protein. Together, these improvements would lend greater confidence to protein identifications while minimizing the potential for false positive associations. It should also be noted that simplistic proteomic approaches are impractical in certain situations for a variety of reasons. For instance, genome data may not be available for a particular species (16), posttranslational modifications may require characterization (17), or the peptides being analyzed may not be protein derived (e.g., neuropeptides or peptide hormones) (18).

Some types of MS/MS scoring routines involve production of a list of expected fragment ions or generation of a predicted MS/MS spectrum. These theoretical predictions are then used to rank potentially matching sequences that lie within a given *m/z* tolerance of known sequences derived from genomic data. To date, knowledge of residue- or peptide-specific dissociation chemistry has been only sparingly incorporated into the process of spectrum prediction and match scoring. Moreover, those algorithms that do include chemically relevant criteria involve only the most simplistic implementations. For example, experimentally observed fragment ions corresponding to the neutral loss of ammonia would require the presence of arginine, lysine, glutamine, or asparagine in the fragment ion. The inclusion of these very simple and qualitative chemical dissociation rules is typically the only extent to which knowledge of peptide ion chemistry informs the informatic aspect of a proteomic experiment.

At present, fragment ion *intensity* information is disregarded or only minimally accounted for by proteomic database search algorithms. The overwhelming majority of these algorithms are based on *m/z* values only, with none of the popular approaches to database searching presently employing a sophisticated model of relative peak intensities among peptide dissociation products. Generally, this means that ion abundance information, including strong enhancement or suppression of particular ions, is not used by the algorithms. Thus, the current paradigm for MS/MS database searching in proteomics is based on only one dimension of inherently two dimensional datasets. The incomplete use of the available spectral information is largely attributed to the fact that it is not yet fully known how to most appropriately determine and exploit peptide product ion intensity information. Considering that sequence information is also encoded within the intensity dimension of an MS/MS spectrum, a chemically meaningful incorporation of fragment ion abundance into tools for proteome informatics has significant potential to improve the success rate and confidence level of sequence and protein identifications. The development of this type of platform is expected to provide a rich and thus far relatively untapped source of sequence relevant information.

A large body of research has established that the relative intensity of peptide fragment ions is remarkably sensitive to peptide composition, sequence, charge state, and the location of charges, as well as the type of instrument and activation method used (19, 20). This complex and nuanced behavior presents major challenges for the design of rigorous predictive models for peptide product ion abundances. Because our research and the research of others has shown that certain structural motifs lead to enhanced or diminished MS/MS cleavage, it is logical to consider whether inclusion of selective cleavage information for particular structural motifs into protein identification algorithms might improve identification rates. Recently, we and several other authors have made the suggestion that greater knowledge of gas-phase peptide dissociation patterns and the underlying chemical reasons for the dissociation patterns might lead to the development of improved algorithms. In order to realize the potential benefits of relative intensity information in a proteomic context, multifaceted and interdisciplinary research will be essential. First, understanding of the chemical basis for cleavage selectivity and fragment ion abundance must be advanced and refined through systematic study of model peptide systems. Second, large databases of peptide MS/MS data must be explored for distinctive spectral features that can be related to peptide sequence. Finally, these insights must be used to inform the design and implementation of improved sequencing algorithms. This chapter will address each of these areas in turn.

2. The Mobile Proton Model of Peptide Dissociation

Peptides are usually analyzed by MS as singly protonated (i.e., $[M+H]^{+}$) and multiply protonated (i.e., $[M+nH]^{n+}$) molecules. The most common method of dissociating peptides in MS/MS is collision-induced dissociation (CID), which involves the conversion of peptide ion kinetic energy into vibrational energy upon impact with neutral, inert target gas atoms or molecules. Peptides may also be subjected to tandem mass spectrometry using surface-induced dissociation (SID), which deposits vibrational energy into precursor ions by means of colliding them with a surface. Although this chapter is primarily focused on peptide ion dissociation as a result of vibrational activation, it is important to note some important alternative activation methods. In recent years, electron capture dissociation (ECD) and electron transfer dissociation (ETD) have proven to be effective dissociation methods for proteomics (21, 22). These activation techniques involve the capture of a low-energy electron by a multiply protonated peptide (in the case of ECD) or transfer of a low-energy electron from an anionic reagent to a multiply protonated peptide (in the case of ETD). While CID and SID MS/MS spectra contain predominantly b and y sequence ions, ECD and ETD MS/MS spectra contain mainly c and z ions. These collisional and electronic activation methods produce very different MS/MS spectra with a high level of complementarity. The combination of complementary activation methods, such as CID and ETD, can often provide more protein identifications than either method alone (21, 23).

Because peptides are polyfunctional molecules, the charge-carrying proton or protons may potentially occupy a number of basic sites on the side chains of amino acid residues (e.g., the side chain guanidino group of arginine residues) or along the peptide backbone (e.g., carbonyl oxygen atoms). Given a sufficient internal energy, an activated peptide ion will undergo unimolecular decay to yield fragment ions. In CID and SID, these are most commonly sequence ions of the b and y types, which are formed through dissociation mechanisms that involve the participation of a charge-carrying proton. Thus, the location of protons exerts a strong influence on the sites of cleavage (24–27). While some potential protonation sites are more favored than others, it should not be overlooked that at a given point in time and for a given distribution of internal energies, a population of ostensibly identical protonated peptides is actually a collection of variously protonated isoforms. That is, a population of protonated peptides can, in reality, be a collection of distinct ions, with the proton or protons occupying different sites. Moreover, a given protonated peptide is not static; rather, protons can be intramolecularly transferred to a number of potential sites.

The foregoing considerations serve to illuminate a general qualitative framework for describing peptide fragmentation behavior on the basis of proton mobility. While the mobile proton model alone does not provide for quantitative prediction of fragment ion intensities, the model does furnish sound chemical rationale for several well known types of enhanced and diminished cleavage. One influence of proton mobility on peptide fragmentation can be dramatically demonstrated by comparing the collision energies required to dissociate peptide ions having differing numbers of charge-carrying protons in relation to the number of basic amino acid side chains (28, 29). Those peptide ions with a number of protons greater than the number of basic amino acid residues tend to dissociate at relatively low collision energies. In these cases, each basic residue is considered to harbor a proton, leaving at least one additional, mobile proton. Dissociation of these precursor ions generally yields product ions with good sequence coverage, as under such circumstances, there are many roughly equivalent sites of protonation that may be occupied by the mobile proton. By contrast, peptide ions with a number of charge-carrying protons less than or equal to the number of basic amino acid residues (particularly, arginine residues) require significantly greater collision energies in order to efficiently dissociate. In these cases, all available protons are most favorably localized at the basic side chains, thus not allowing for a readily mobile proton. In this case, additional energy is required to mobilize these sequestered protons or to reduce basicity by an intermediate neutral loss and thus allow the participation of these protons in backbone cleavage mechanisms.

Proton mobility not only plays a role in the overall activation energy required to bring about peptide ion dissociation but also serves to explain some well-known types of selective cleavage. For example, cleavage C-terminal to aspartic acid residues (and, to a lesser extent, glutamic acid residues) is highly favored in the absence of mobile protons (30, 31). This type of enhanced cleavage has been attributed to the participation of an acidic side chain proton in the dissociation mechanism. Because the proton participating in the dissociation chemistry is not the charge-carrying proton, this type of cleavage is often described as a charge-remote pathway. When mobile protons are available, cleavage C-terminal to acidic residues becomes an essentially nonselective process. Selective cleavage is also commonly observed at the C-terminus of histidine residues, although the behavior of this cleavage is different from that seen at the C-terminus of acidic residues (32). For these peptide ions, the fragmentation occurs preferentially only in the presence of mobile protons. This observation has been interpreted as evidence that a charge-carrying proton must occupy the histidine side chain imidazole group in order to bring about the selective cleavage. By contrast, histidine-containing peptide ions with no mobile protons cleave in a nonselective manner. While these examples do not constitute an exhaustive discussion of mobile proton related selective cleavage types, they do serve to illustrate the exquisite sensitivity of peptide dissociation patterns to the chemistry of each specific ion.

3. Elucidation of Chemical Trends from Collections of Fragmentation Spectra

As mentioned previously, proteomics experiments use algorithms, such as Sequest or Mascot, to assign peptide sequences to peptide fragmentation spectra in order to identify the corresponding proteins present in a sample (12, 14). While these programs have greatly enabled progress in proteomics, they are still limited from both a practical and chemical perspective. Of the thousands of tandem mass spectra acquired in a given experiment, only a small percentage of the spectra are identified by the algorithms (33–36). This may be due in part to the simplicity of the chemical fragmentation models these algorithms use, as mentioned in the previous section (12, 14). One limitation of the fragmentation models used is that cleavages are predicted to occur almost exclusively at the amide bond between neighboring residues, regardless of amino acid residues present. As many groups have identified several reproducible residue-dependent cleavage enhancements (19, 31, 37, 38), it is clear that the algorithms do not take into account all of the chemical information available to describe a peptide fragmentation spectrum. Incorporating more chemically detailed information may help to improve the ability of an algorithm to correctly identify a peptide based on a fragmentation spectrum if a robust, fast, and sophisticated model can be developed.

A wide variety of chemical properties have been shown to affect the fragmentation pattern of a peptide. Some of those explored include size, charge state, and residues present (28, 30, 37–42). The way in which all of these factors act together to give a certain fragmentation spectrum is complex and not yet fully understood. Two main approaches have been taken in order to understand the effect of different characteristics on peptide fragmentation: systematic studies using model peptides and data mining applied to large datasets.

4. Model Peptide Studies

Several groups have used small subsets of model peptides to demonstrate trends in peptide fragmentation spectra. Tsaprailis et al. used a small set of angiotensin peptide analogs to systematically explore the effect of the neighboring residue on enhanced cleavage at histidine residues (32). Dongre et al. demonstrated the role of residue basicity, peptide length, and peptide sequence on fragmentation patterns using systematically modified leucine enkephalin analogs, polyalanine analogs, and des-Arg bradykinin derivatives (28). Figure 2 shows the fragmentation efficiency curves for a series of singly protonated polyalanine analogs with different N-termini. As the gas phase basicity of the first residue increases, additional collision energy is required to achieve the same fragmentation efficiency. The increase in energy required to achieve fragmentation within the given timescale demonstrates the ability of more basic residues to more tightly sequester the ionization proton, a result that played a role in development of the mobile proton model.

Vaisar and Urban used a similar method to examine the proline effect on peptide fragmention by looking at a series of five different peptides of the sequence Ala-Val-X-Leu-Gly (43). These studies and others clearly indicate that multiple factors are responsible for the overall fragmentation behavior of a peptide. While each of these examples can describe differences in fragmentation behavior in relation to other peptides in the study that have been varied with a systematic intent, it is not possible to either fully elucidate all of the contributions to the fragmentation spectrum, nor is it possible to draw more general conclusions of how these

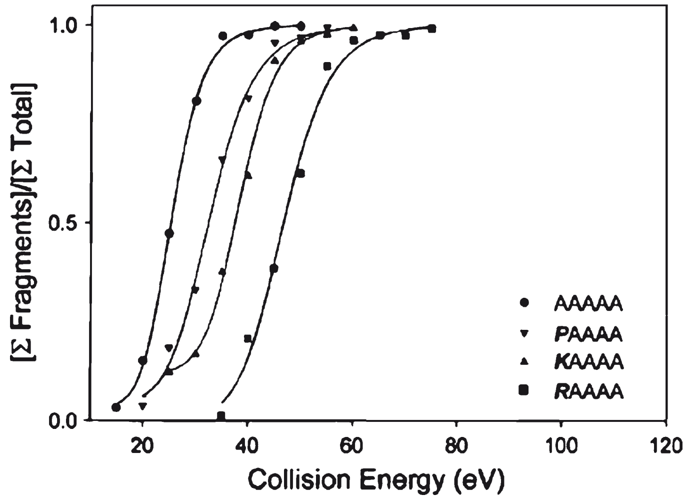

Fig. 2. Influence of gas-phase basicity on fragmentation efficiency. As gas-phase basicity increases (from A to P to K to R), the ionizing proton is more easily sequestered so that more energy is required to achieve the same fragmentation efficiency. Reproduced with permission from *J. Am. Chem. Soc.* 1996, *118*, 8365–8374. Copyright 1996 Am. Chem. Soc

factors can be applied to larger sets of spectra. Because proteomics readily generates a large number of spectra to be interpreted, and because large numbers of spectra are needed to achieve statistically valid numbers of combinations of various residues, methods that seek to discern fragmentation patterns from large sets of data may be more appropriate tools.

5. Introduction to Data Mining

Tandem mass spectrometry data are aptly suited for data mining as a typical proteomics experiment will quickly generate several thousands of widely varied MS/MS spectra. The goal of data mining is to identify underlying patterns from the spectra that can ideally be correlated to chemical phenomena that will help describe the ways in which peptides fragment. Generally, data mining can be broken down into two approaches after data acquisition: classification and pattern analysis, and/or clustering and pattern analysis.

It is important to note here that a major requirement of data mining is the availability of large, high quality datasets in which there is great certainty that the peptide sequences are correctly identified based on the corresponding fragmentation spectra. Datasets consisting of a few thousands to a few million spectra have been studied via data mining in order to elucidate trends

(33, 35, 36, 38, 40, 42, 44). Small sets of model peptides have an advantage in terms of the ease of assembling the data set because peptides and their desired analogs can readily be synthesized and easily characterized using basic MS and MS/MS measurements. Assembling a dataset with thousands of spectra in the same manner would be extremely time intensive and lacking in practicality. Rather than synthesizing thousands of peptide analogs, proteolytic digests of complex protein mixtures are analyzed via LC-MS and the corresponding MS/MS spectra are collected. As stated previously, in a given experiment of this type, as few as 10–35% of the spectra can be correctly identified. In order to trim these data sets to include only spectra that have had their sequences identified with high certainty, the data are first run through an algorithm, and the spectra that are matched to a peptide/protein with an acceptable cutoff score are saved (33, 35, 38, 42). In order to further validate a dataset, Smith and coworkers ran a complex digest through two types of mass spectrometers, an FT-ICR and an ion trap, which were coupled with identical chromatographic conditions (45). The combination of the accurate mass measurements from the FT-ICR and the fragmentation spectra from the ion-trap was paired with the use of Sequest; when Sequest identified the peptide that was within 1 ppm of the accurate mass and correlated to the fragmentation spectrum at the same retention time within a margin of error, then the spectrum was considered to be identified with very high confidence. However, this approach necessarily introduces bias because those spectra and sequences that are not identified are not represented in the database. While this method would not eliminate all incorrectly assigned peptide fragmentation spectra, it would identify a large number of high quality spectra in a relatively small amount of time.

The motivation for using a larger dataset as opposed to a set of systematically altered model peptides is that a larger distribution and variability of peptides and their corresponding fragmentation spectra will be present. With a greater distribution, the goal is to identify underlying trends in the fragmentation spectra that can be universally applied to future systems. However, many subsets are limited to one charge state or one type of peptide. Many have focused their studies on doubly charged tryptic peptides, as they are a common type of peptide ion seen (35, 38, 40, 42). Only a few researchers, including Wysocki and Zhang, have investigated the role of a variety of charge states (41, 46). While some other charge states and nontryptic peptides are less common in proteomics experiments, it is nonetheless important to acknowledge the specific bias a given dataset may contribute to the outcome of a data mining effort.

Once the dataset is assembled, data mining may proceed through two main approaches: classification and pattern analysis or clustering and pattern analysis. One common approach is to first include a preclassification step. Based on previously understood chemical principles, Huang et al. preliminarily separated data from 28,311 spectra into nine subsets based on structural features, such as proline content and basic residue content, and the charge state (41). In each subset of this study, pairwise fragmentation maps were generated to describe cleavages between all possible residue pairs. An example of this fragmentation map is shown in Fig. 3, which illustrates the y (top) and b (bottom) ion intensity patterns among doubly charged arginine (left) and lysine (right) terminated peptides. These fragmentation maps yield a plethora of information that may be integrated into future peptide identification algorithms.

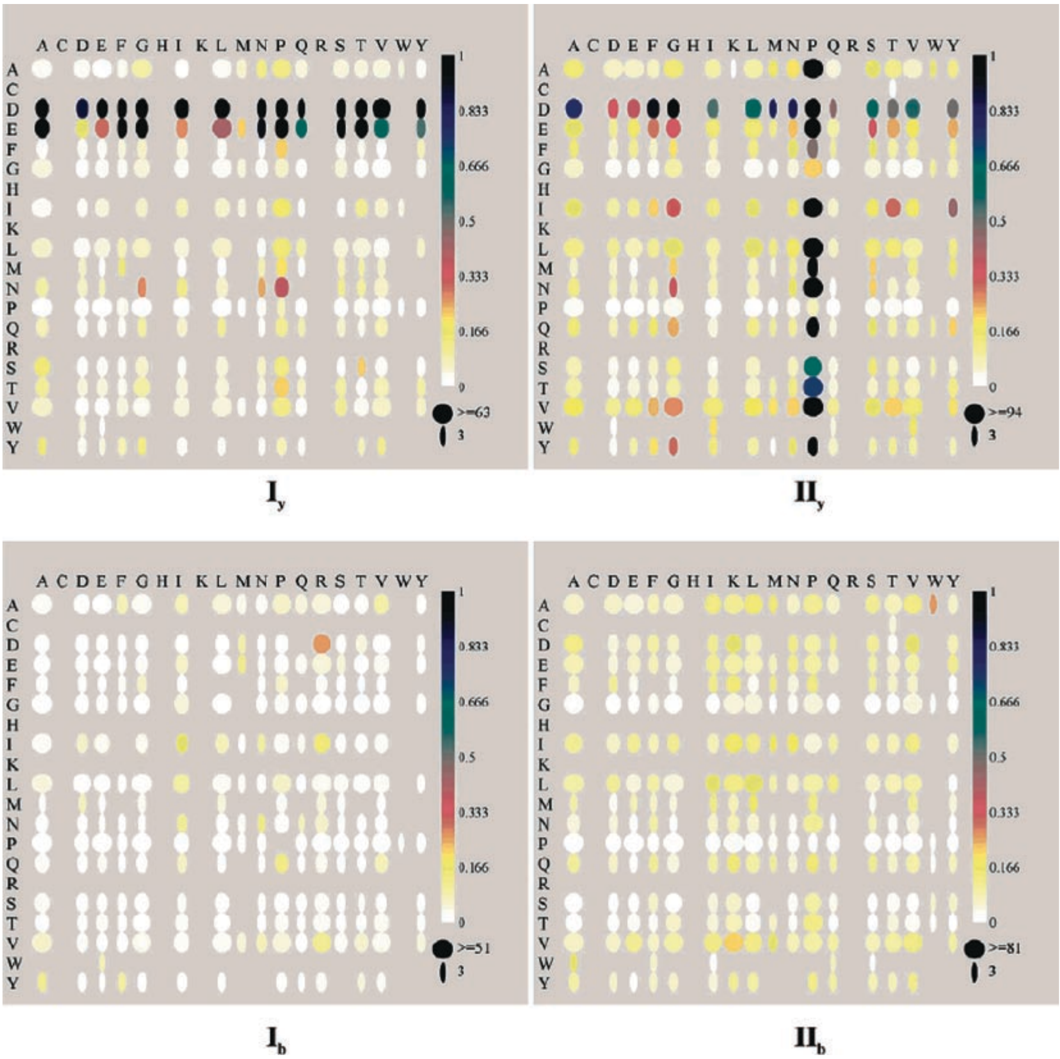

Fig. 3. Pairwise fragmentation map for singly charged peptides ending in arginine (Iy and Ib) or lysine (IIy and IIb). Reproduced with permission from *Anal. Chem.* 2005, *77*, 5800–5813. Copyright 2005 Am. Chem. Soc

In a similar approach, Tabb et al. examined trends in a database of 1,465 doubly charged tryptic peptides (35). Initially, they refined their dataset to include only doubly charged ions whose spectra contained at least 50% of the theoretically predicted ions that were fully tryptic; that is, ending in Arg or Lys without any internal Arg or Lys residues. They then examined the relationship between fragment intensity and ion series origin, fragment mass, residue type and effect on the neighboring amide bond cleavage, and the link between peptide amino acid composition and neutral fragment loss. In another study by Tabb et al., proteinase K was used to generate 2,568 nontryptic doubly charged peptides so that the role of basic residue location in a peptide could be correlated to fragmentation efficiency (42). A similar method was used by Kapp et al. to investigate trends using a dataset of 5,500 peptides. The authors demonstrated that the incorporation of a proton mobility factor could greatly improve algorithm identification success (36).

Others have used data mining to focus on specific fragmentation patterns, such as Huang's investigation of the influence of internal basic residues on the fragmentation C-terminal of the acidic residues Asp and Glu and Breci's look at fragment ion intensities due to cleavage N-terminal to Pro (37, 38). Through an examination of the b and y fragment ion intensity C-terminal to Asp when an internal His was present, Huang and coworkers were able to demonstrate that cleavage C-terminal to Asp was enhanced because of the ability of a basic His internal residue to sequester protons for doubly charged tryptic peptides. Breci et al. used a measure of the relative bond cleavage, which compares the intensity of the ions from cleavage at Pro to the intensity of all ions present in the spectrum, to determine that while cleavage N-terminal to Pro is reproducible for a certain residue, there is not enough chemical understanding as of yet to fully elucidate the entire fragmentation mechanism.

An alternative approach taken by Huang et al. was to use a penalized K-means algorithm to allow for unsupervised clustering of 28,330 spectra (47). This allowed for the peptide fragmentation spectra to cluster into four groups without the introduction of any prior chemical knowledge into the algorithm, as shown in Fig. 4. After the clustering, a decision tree was used in order to correlate the clusters to specific chemical properties. A fifth cluster for noise and outlier peaks was also generated using a method developed by Tseng, to allow for cleaner clustering. This method is important because it bypasses the need to introduce any prior assumptions and instead provides a relatively unbiased overview of the fragmentation behavior observed in the dataset as a whole.

Whittaker and coworkers have employed an alternative data mining technique that they refer to as statistical modeling, which

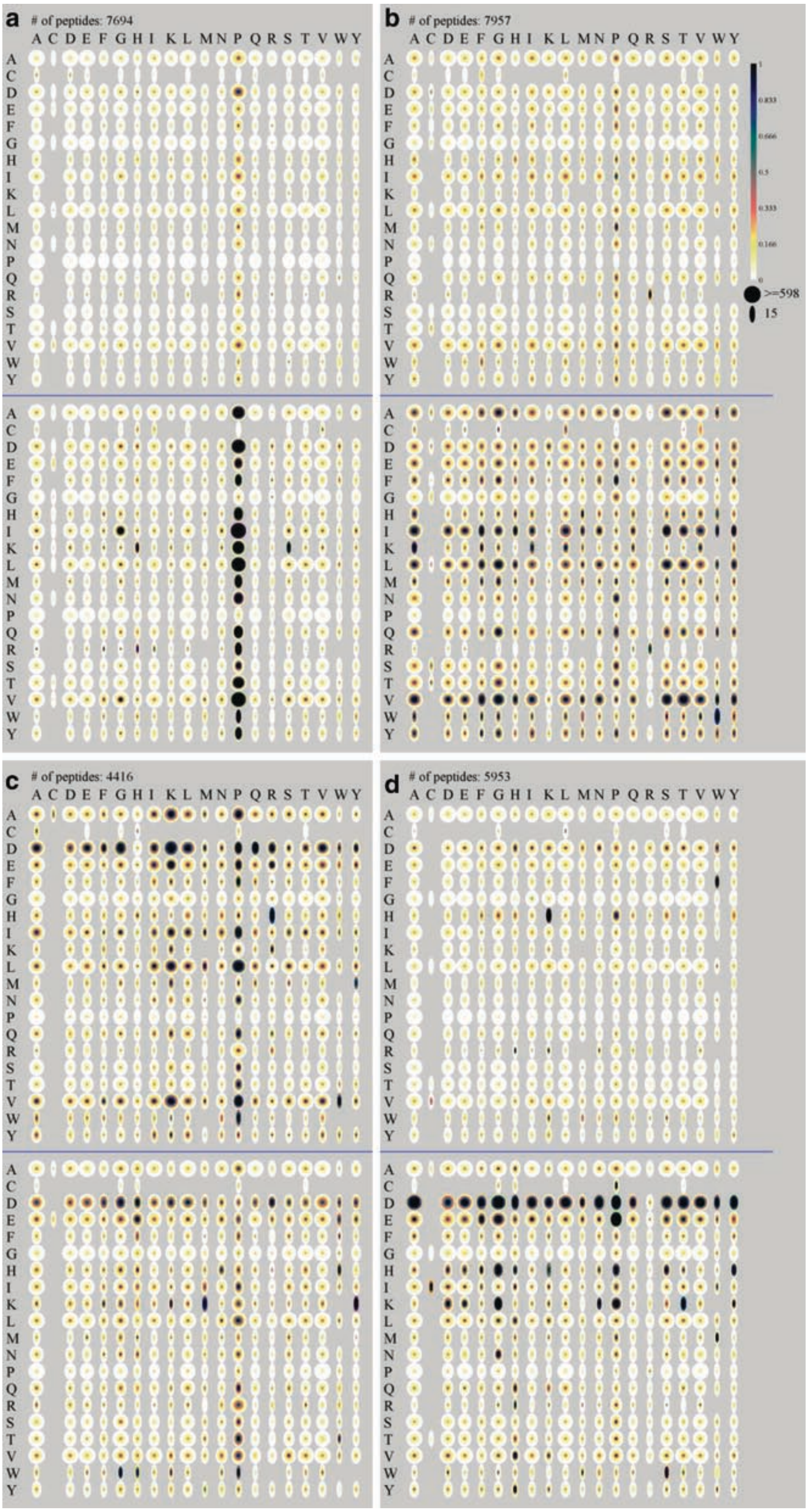

Fig. 4. Quantile maps of b (*above*) and y (*below*) ions for the four clusters identified from Huang's study using a penalized K-means algorithm for unsupervised clustering. The four clusters of spectra are characterized by the dominant cleavages patterns seen: (**a**) X–P, (**b**) I/L/V–X, (**c**) both D–X and X–P, and (**d**) D/E–X. Reproduced with permission from Proteome Res. 2008, 7, 70–79. Copyright 2008 Am. Chem. Soc

uses probabilistic models relating trends in fragmentation spectra to multiple predictor variables (39, 48). The key advantage of statistical modeling is in the ability to consider each factor simultaneously as opposed to independently. This is ideally suited for the interpretation of tandem mass spectra, as the factors dictating a particular fragmentation pattern are complex and multivariate in nature. For example, Barton et al. used models to describe b and y ion formation (separately, as they regarded different factors to influence the formation of each ion type) involving fragment ion mass, cleavage location and neighboring residues, and peptide residue composition (48).

Elias et al. used a machine learning approach to examine the ion intensities of 27,000 high quality fragmentation spectra to develop a model that can describe how likely it is that certain fragments would appear with a predicted relative intensity (33). They compared these predictions to a set of peptides that were either matched or mismatched to determine how the incorporation of ion intensity information could improve the success of the peptide identification algorithm. They saw improvements in peptide identification from 50 to 96%, suggesting that the incorporation of intensity is crucial to the improvement of these algorithms. This will be further discussed in the following section.

6. Incorporation of Fragment Ion Intensity in Peptide Sequencing Algorithms

As mentioned previously, various factors, including size, charge state, amino acid content, and charge location, can contribute to the process of gas phase peptide dissociation, making the resulting fragmentation spectra difficult to fully predict or interpret (19). This problem is compounded by the fact that most current algorithms rely on models that oversimplify the fragmentation process, thus causing valuable spectral information to be discarded. Introducing more of the available chemical information and fragmentation patterns into a sequencing algorithm could therefore allow the algorithm to more efficiently and more accurately match a peptide fragmentation spectrum to its correct matching peptide. This section will examine how several popular algorithms use the available peptide fragmentation information to predict spectral matches.

Some of the popular algorithms that are used to perform peptide sequencing or protein identification from MS/MS data include MS-Tag, SEQUEST, MASCOT, X!Tandem, OMSSA, and Phenyx (14). MS-Tag is an algorithm that was originally developed for the interpretation of MS/MS spectra that do not contain a contiguous ion series; that is, not all characteristic b and y ions are present (11). Figure 5 shows an experimental spectrum

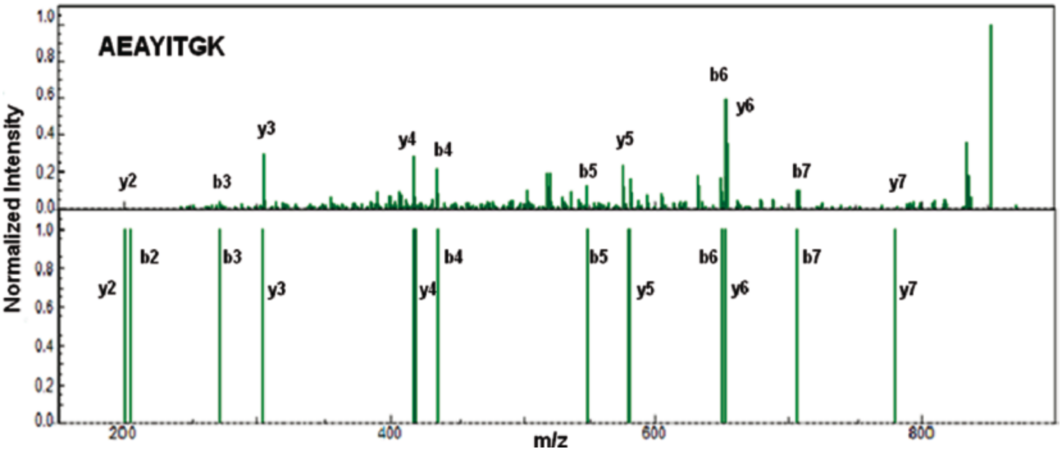

Fig. 5. Comparison of actual peptide fragmentation spectrum (*top*) to contiguous ion series (*bottom*)

and the theoretical contiguous ion series that would correspond to the sequence of the peptide AEAYITGK.

Assignment of a peptide sequence to a spectrum involves calculating the theoretical fragment ion *m/z* values for all candidate peptide sequences. MS-Tag ranks the candidate sequences in the order of increasing number of unmatched experimental fragment ions.

SEQUEST is an algorithm that correlates a given uninterpreted MS/MS spectrum with candidate sequences through the use of scoring and ranking methods based on spectral similarity by cross-correlation of the theoretically predicted spectra and the experimental spectrum (11). However, SEQUEST does not compare the raw spectra with predictions. Instead, it divides the spectrum into 10 bins and normalizes each to the most intense peak in the bin, effectively removing relative ion intensity across the entire fragmentation spectrum as a strong determinant of a match. This approach has been very successful in matching spectra to candidate sequences despite the lack of detailed rules for predicting fragment ion intensities.

MASCOT is an algorithm that contains multiple approaches to database searching, of which two use MS/MS data (MS/MS Ion Search and Sequence Query) (14). MS/MS Ion Search calculates theoretical fragment ion masses in a similar manner to that of MS-Tag before matching them to experimental spectra. Sequence Query requires some manual interpretation of the MS/MS data during which molecular weight, residue composition, and sequence qualifiers are determined for the candidate sequences. Both MASCOT strategies use the same probability-based scoring routine based on the MOWSE algorithm in which peptide size distributions (or peptide fragment size distributions) are considered with respect to protein masses (or peptide masses) in the searched database. A cutoff score for the probability that a match is a purely random event is given for each search.

X!Tandem, the most popular open source algorithm, uses intensity in its preliminary score, or hyperscore (49). This score is similar to ion intensity current, which is the sum of the intensities of all b and y ions found in the experimental spectra. This is not the same as using peak intensity information that reflects chemical fragmentation suppression or enhancement; it only acknowledges the presence of a peak. Through a statistical analysis of the hyperscore of each candidate sequence, an expectation value (E-value) describing the significance of the difference between the top match and other matches is generated and used as the main score of X!Tandem. Because this idea is common to several algorithms, the use of a hyperscore alone is not enough to significantly improve the success of X!Tandem when compared to other algorithms that use additional information and scoring stages to assign peptide spectra.

OMSSA (Open Mass Spectrometry Search Algorithm) is another example of an open source algorithm that uses expectation values as criteria, similar to X!Tandem. The older version of OMSSA only uses intensity as a threshold to filter noisy peaks (13), while the newer version has improved how intensity is used (50). In the newer edition, each peak in the experimental spectrum is ranked. The sum of the ranks of the matched peaks is compared with a normal distribution of ranks of random peak sums to calculate an expectation value. Like X!Tandem, OMSSA is complementary to Sequest because it gives an identification a probability component, whereas Sequest matches do not include probability.

Lastly, Phenyx is a platform that generates its score based on an extended match, which matches a peptide using a combination of and comparison between theoretical and experimental spectra. (51). In other words, this method incorporates structural information such as intensity, ion series contiguity, and spectral signal-to-noise ratios in addition to *m/z* information, and the extended match score reflects the quality of a match. By analyzing a testing set of spectra with known sequences, Phenyx calculates the probability of observing the above extended match information when the match is correct or if the match is purely random; the ratio of these two probabilities is the Phenyx score. When attempting to identify a peptide sequence from an unknown spectrum, similar extended match information can be generated against candidate sequences in a given database to determine the ratio score. Evaluation of the score will enable true matches to be distinguished from false.

While these algorithms are popular and successful in proteomics studies worldwide, they are not without limitations. Because every spectrum is assigned to a sequence candidate, a variety of studies have shown that in a typical MS/MS run, over 80% of the peptide identifications by SEQUEST are false and filters are necessary to eliminate those low confidence matches; programs have been developed, such as DTASelect, Peptide Prophet, and Protein Prophet, that remove these low confidence matches (52–54). However, scoring cut-off filters may also require that some correctly identified spectra are discarded in order to remove a majority of the false positive identifications. Though many proteins can still be identified using current algorithms, and the use of multiple algorithms can be combined to increase protein identification confidence as demonstrated by Searle et al. (15), these algorithms are still far from optimally meeting the rapid identification demands of the proteomics experiments that generate large volumes of peptide fragmentation spectra.

One common characteristic for all of these widely used algorithms is that they mainly utilize the mass-to-charge ratio information from a mass spectrum while ignoring the intensity component beyond the intensity threshold (12, 14). This is generally a result of insufficient knowledge of the peptide dissociation process, as we mentioned previously, though some efforts have been made recently to include intensity into peptide identifications algorithms (46, 47, 55–58). As discussed previously, reproducible intensity patterns have been identified for several residues, such as the study by Breci and coworkers on the enhanced cleavages N-terminal to proline (37). The integration of intensity is emphasized in certain algorithms not because it is more critical than *m/z*, but because it can provide additional correlating information that can assist with the peptide identification. Studies have shown that the incorporation of intensity can reduce peptide fragmentation identification error by 50–96% (33). Clearly, the use of intensity to improve peptide identification rates is an attractive prospect. Indeed, while this chapter has placed strong emphasis on the relevance of fragment ion intensity to proteomic strategies, the importance of *m/z* values cannot be minimized. Because a wide variety of MS platforms are being applied to proteomics, it is of utmost importance that proteome researchers be aware of the mass resolution and mass accuracy performance characteristics of the mass analyzer being used. Such information is essential for the appropriate setting of precursor and fragment ion mass tolerances, and the specification of average versus monoisotopic masses at the database search stage.

Different from the popular algorithms mentioned above, algorithms incorporating intensity do not work under the assumption that the all amino acid pairs and peptide patterns dissociate non-selectively to generate peaks without discrimination in intensity. Though the appearance of a given spectrum is difficult to predict, results have shown that given the same experimental conditions mass spectra are reproducible (33, 37, 46, 57). Schutz and colleagues assessed this reproducibility by using an ion trap dataset produced by the same instrument and parameters via three different methods: correlation between the intensities of two spectra as a measure of their similarity, normalized dot product of both the peak intensities from pairs of spectra, and the square root of the intensities (59). They found that MS/MS spectra, especially of peptides with low charge states, exhibit reproducible fragmentation intensities and patterns, which enables the prediction of peak intensity. Newer algorithms that incorporate complex intensity models that are based on either probability or chemical properties will be discussed below.

7. Probability Based Algorithms

Elias and coworkers used a probabilistic decision tree – specifically, a treelike feather extracting graph, which requires the members of each branch to have similar properties – to model the probability of observing certain peak intensities in a mass spectrum from 27,266 high quality spectra (33). The most confident true matches from SEQUEST were selected and decision trees were generated using 63 different attributes, including b ion length, y ion length, fraction of basic residues, and peptide length. Each node of the tree represents a chemical property that can separate the intensity into different bins, and the likelihood that a certain fragment ion peak will have a certain intensity that can be calculated from the distribution of the sizes of the resulting branches. With the input of a predicted ion from a candidate sequence, the likelihood of yielding the measured intensity in the experimental spectrum can be obtained from the decision tree. For both correctly matched and mismatched peptides, the decision trees are made and compared to serve as a guideline as to whether an identification is correct or incorrect. More than a 50% decrease in peptide identification error rate was achieved when using this method in conjunction with SEQUEST.

Another intensity based algorithm is Narasimhan's Multinomial Algorithm for Spectral Profile-based Intensity Comparison (MASPIC) scorer (60). Though based on a popular random match assumption that the correct match should have the least likelihood to be achieved randomly by chance only, MASPIC considered the possibility of random intensity matches as an alternative to using *m/z* only. This method divides the whole experimental spectrum into +1, +2, and +3 zones according to the charge of the fragment. In each zone, peaks are binned into classes with descending intensity, where lower intensity classes have more peak members. This process converts the experimental spectrum into a probability profile along the *m/z* axis. It is more likely to randomly match a predicted peak from a candidate sequence into the lower intensity class because this class has more members, thus decreasing the importance of a match with decreasing intensity. When all predicted peaks from a candidate sequence are compared with this probability profile, the number of matched and unmatched peaks for each class is counted, and further calculations are performed to give a probability of matching.

8. Chemical Property Based Algorithms

Zhang reported a kinetic model for prediction of low-energy CID spectra from sequence in 2004, with a general idea to abandon the traditional statistics model used by intensity prediction efforts and mimic the peptide dissociation process based on kinetics and the mobile proton model (57). The key assumption is that the intensity of a fragment ion is determined by the rate of the dissociation pathway generating this fragment; if the rate constants for all fragment ion pathways are known, then the relative intensity of each fragment can be predicted. Collision energy, proton density, fragmentation rate, ion cooling rate, activation energy, and gas-phase basicity are considered and incorporated into the rate calculation of eleven different backbone cleavage pathways as well as side-chain cleavages and neutral losses. Based on this iterative calculation model, Zhang developed an algorithm called MassAnalyzer, which uses a Sim score to evaluate the similarity of a simulated and experimental spectrum (57).

The kinetic model is mainly used to confirm the results from popular algorithms rather than to provide independent protein identification. This is due to various limitations, including variability between spectra acquired on different instruments under different experimental conditions and the large number of parameters that must be considered, as mentioned above. The Resing group later used this model as one part of the Manual Analysis Emulator (MAE), a program intended to improve the validation of tandem mass spectra (61). Another part of this MAE program takes into account the proportion of the ion current (PIC), which represents the percentage of intensities in an experimental spectrum that can be derived from the peptide sequence. A higher PIC score means that the program was using the most intense peaks for peptide identification as opposed to noise and low abundance peaks. With the incorporation of these two intensity-related scores, MAE yielded a better discrimination between true and false matches of SEQUEST and Mascot results.

Clearly, peptide searching algorithms utilize a variety of spectral and chemical information to assign peptide sequences to spectra. Selecting a single algorithm over another will likely lead to different sets of peptide and protein assignments based on the criteria that an algorithm uses. As briefly mentioned earlier, the use of multiple search algorithms has been shown to improve confidence of a peptide identification. Programs such as Scaffold, available from Proteome Software, provide an interface for direct comparison of MS/MS data analyzed using a variety of algorithms (15). As new algorithms are developed, it is important to understand what spectral characteristics allow the algorithm to more accurately match certain spectra to peptide sequences while the matches for other spectra with different characteristics are poor. Programs such as Scaffold will allow algorithms to be more readily compared.

9. Prospectus

We can imagine a time in the future when our fundamental knowledge and computational capabilities are sufficiently advanced to rapidly and accurately predict theoretical MS/MS spectra for any given peptide sequence. This will ultimately require that different protonation motifs, their relative probabilities of existence, their relative propensities for interconversion, and their overall contribution to dissociation kinetics all be taken into account. This would be a significant advance, as theoretical sequences could be generated to match a measured accurate mass and the corresponding synthetic tandem mass spectra could be generated and compared to the experimental spectrum. This should, in principle, allow peptide sequence identification to be obtained even in the absence of protein level information and even in the absence of genomic information. In approaching this goal, it will be necessary to continue systematic investigation of peptide structure and gas-phase unimolecular ion chemistry of protonated peptides and to incorporate the forthcoming insights into the next generation of proteomic search algorithms.

References

1. Aebersold, R., and Goodlett, D. R. (2001) Mass spectrometry in proteomics. *Chemical Reviews* **101**, 269–96.
2. Resing, K. A., and Ahn, N. G. (2005) Proteomics strategies for protein identification. *FEBS Letters* **579**, 885–89.
3. Griffin, T. J., and Aebersold, R. (2001) Advances in proteome analysis by mass spectrometry. *Journal of Biological Chemistry* **276**, 45497–500.
4. Mo, W., and Karger, B. L. (2002) Analytical aspects of mass spectrometry and proteomics. *Current Opinion in Chemical Biology* **6**, 666–75.
5. Smith, R. D. (2002) Trends in mass spectrometry instrumentation for proteomics. *Trends in Biotechnology* **20**, s3–s7.
6. Boutilier, K., Ross, M., Podtelejnikov, A. V., Orsi, C., Taylor, R., Taylor, P., and Figeys, D. (2005) Comparison of different search engines using validated MS/MS test datasets. *Analytica Chimica Acta* **534**, 11–20.
7. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winton, S., and Hauer, C. R. (1986) Protein

sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 6233–37.

8. Wysocki, V. H., Resing, K. A., Zhang, Q., and Cheng, G. (2005) Mass spectrometry of peptides and proteins. *Methods* **35**, 211–22.
9. Yates, J. R., 3rd, Speicher, S., Griffin, P. R., and Hunkapiller, T. (1993) Peptide mass maps: a highly informative approach to protein identification. *Analytical Biochemistry* **214**, 397–408.
10. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71.
11. 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**, 2871–82.
12. Eng, J. K., McCormack, A. L., and Yates, J. R. (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**, 976–89.
13. 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**, 958–64.
14. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–67.
15. Searle, B. C., Turner, M., and Nesvizhskii, A. I. (2008) Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *Journal of Proteome Research* 7, 245–53.
16. Liska, A. J., and Shevchenko, A. (2003) Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics* **3**, 19–28.
17. Mann, M., and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. *Nature Biotechnology* **21**, 255–61.
18. Fricker, L. D., Lim, J., Pan, H., and Che, F. (2006) Peptidomics: identification and quantification of endogenous peptides in neuroendocrine tissues. *Mass Spectrometry Reviews* **25**, 327–44.
19. Paizs, B., and Suhai, S. (2005) Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews* **24**, 508–48.
20. Wysocki, V. H., Tsaprailis, G., Smith, L. L., and Breci, L. A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry* **35**, 1399–406.
21. 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**, 1942–51.
22. Savitski, M. M., Kjeldsen, F., Nielsen, M. L., and Zubarev, R. A. (2006) Complementary sequence preferences of electron-capture dissociation and vibrational excitation in fragmentation of polypeptide polycations. *Angewandte Chemie International Edition* **45**, 5301–03.
23. Molina, H., Matthiesen, R., Kandasamy, K., and Pandey, A. (2008) Comprehensive comparison of collision induced dissociation and electron transfer dissociation. *Analytical Chemistry* **80**, 4825–35.
24. Harrison, A. G., and Yalcin, T. (1997) Proton mobility in protonated amino acids and peptides. *International Journal of Mass Spectrometry* **165–166**, 339–47.
25. Cox, K. A., Gaskell, S. J., Morris, M., and Whiting, A. (1996) Role of the site of protonation in the low-energy decompositions of gas-phase peptide ions. *Journal of the American Society for Mass Spectrometry* 7, 522–31.
26. 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**, 137–54.
27. Tsaprailis, G., Nair, H., Somogyi, A., Wysocki, V. H., Zhong, W., Futrell, J. H., Summerfield, S. G., and Gaskell, S. J. (1999) Influence of secondary structure on the fragmentation of protonated peptides. *Journal of the American Chemical Society* **121**, 5142–54.
28. Dongre, A. R., Jones, J. L., Somogyi, A., and Wysocki, V. H. (1996) Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: evidence for the mobile proton model. *Journal of the American Chemical Society* **118**, 8365–74.
29. Gu, C., Somogyi, A., Wysocki, V. H., and Medzihradszky, K. F. (1999) Fragmentation of protonated oligopeptides XLDVLQ (X=L, H, K or R) by surface induced dissociation: additional evidence for the 'mobile proton' model. *Analytica Chimica Acta* **397**, 247–56.
30. Gu, C., Tsaprailis, G., Breci, L., and Wysocki, V. H. (2000) Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in

fixed-charge derivatives of Asp-containing peptides. *Analytical Chemistry* **72**, 5804–13.

31. Rozman, M. (2007) Aspartic acid side chain effect-experimental and theoretical insight. *Journal of the American Society for Mass Spectrometry* **18**, 121–27.
32. Tsaprailis, G., Nair, H., Zhong, W., Kuppannan, K., Futrell, J. H., and Wysocki, V. H. (2004) A mechanistic investigation of the enhanced cleavage at histidine in the gas-phase dissociation of protonated peptides. *Analytical Chemistry* **76**, 2083–94.
33. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., and Gygi, S. P. (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology* **22**, 214–19.
34. Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Jonscher, K. R., Pierce, K. G., Old, W. M., Cheung, H. T., Russell, S., Wattawa, J. L., Goehle, G. R., Knight, R. D., and Ahn, N. G. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Analytical Chemistry* **76**, 3556–68.
35. Tabb, D. L., Smith, L. L., Breci, L. A., Wysocki, V. H., Lin, D., and Yates, J. R. (2003) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Analytical Chemistry* **75**, 1155–63.
36. Kapp, E. A., Schuetz, F., Reid, G. E., Eddes, J. S., Moritz, R. L., O'Hair, R. A., Speed, T. P., and Simpson, R. J. (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Analytical Chemistry* **75**, 6251–64.
37. Breci, L. A., Tabb, D. L., Yates, J. R., and Wysocki, V. H. (2003) Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Analytical Chemistry* **75**, 1963–71.
38. Huang, Y., Wysocki, V. H., Tabb, D. L., and Yates, J. R. (2002) The influence of histidine on cleavage C-terminal to acidic residues in doubly protonated tryptic peptides. *International Journal of Mass Spectrometry* **219**, 233–44.
39. Barton, S. J., and Whittaker, J. C. (2008) Review of factors that influence the abundance of ions produced in a tandem mass spectro-meter and statistical methods for discovering these factors. *Mass Spectrometry Reviews* **28(1)**, 117–187.
40. Huang, Y., Triscari, J. M., Pasa-Tolic, L., Anderson, G. A., Lipton, M. S., Smith, R. D., and Wysocki, V. H. (2004) Dissociation behavior of doubly-charged tryptic peptides: correlation of gas-phase cleavage abundance with ramachandran plots. *Journal of the American Chemical Society* **126**, 3034–35.
41. Huang, Y., Triscari, J. M., Tseng, G. C., Pasa-Tolic, L., Lipton, M. S., Smith, R. D., and Wysocki, V. H. (2005) Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Analytical Chemistry* **77**, 5800–13.
42. Tabb, D. L., Huang, Y., Wysocki, V. H., and Yates, J. R. (2004) Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry* **76**, 1243–48.
43. Vaisar, T., and Urban, J. (1996) Probing the proline effect in CID of protonated peptides. *Journal of Mass Spectrometry* **31**, 1185–87.
44. Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., and Pevzner, P. A. (2008) Clustering millions of tandem mass spectra. *Journal of Proteome Research* **7**, 113–22.
45. Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y. F., Conrads, T. P., Veenstra, T. D., and Udseth, H. R. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2**, 513–23.
46. Zhang, Z. (2005) Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Analytical Chemistry* **77**, 6364–73.
47. Huang, Y., Tseng, G. C., Yuan, S., Pasa-Tolic, L., Lipton, M. S., Smith, R. D., and Wysocki, V. H. (2008) A data-mining scheme for identifying peptide structural motifs responsible for different MS/MS fragmentation intensity patterns. *Journal of Proteome Research* **7**, 70–79.
48. Barton, S. J., Richardson, S., Perkins, D. N., Bellahn, I., Bryant, T. N., and Whittaker, J. C. (2007) Using statistical models to identify factors that have a role in defining the abudnace of ions produced by tandem MS. *Analytical Chemistry* **79**, 5601–07.
49. Craig, R., and Beavis, R. C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrometry* **17**, 2310–16.
50. Geer, L. Y., Bai, D. L., Kowalak, J. A., Chi, A., Xu, M., Shabanowitz, J., Markey, S. P., Hunt, D. F., and Bryant, S. H. (2008), National Library of Medicine, NIH, Bethesda, MD. University of Virginia, Charlottesville, VA, National Institute of Mental Health, NIH, Bethesda, MD.
51. Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: towards high-throughput tandem mass spec-

trometry data identification. *Proteomics* **3**, 1454–63.

52. Tabb, D. L., McDonald, W. H., and Yates, J. R. (2002) DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *Journal of Proteome Research* **1**, 21–26.
53. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifiactions made by MS/MS and database search. *Analytical Chemistry* **74**, 5383–92.
54. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry* **75**, 4646–58.
55. Gibbons, F. D., Elias, J. E., Gygi, S. P., and Roth, F. P. (2004) SILVER helps assign peptides to tandem mass spectra using intensity-based scoring. *Journal of the American Society for Mass Spectrometry* **15**, 910–12.
56. Havilio, M., Haddad, Y., and Smilansky, Z. (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Analytical Chemistry* **75**, 435–44.
57. Zhang, Z. (2004) Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry* **76**, 3908–22.
58. Zhou, C., Bowler, L., and Feng, J. (2008) A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC Bioinformatics* 9325.
59. Schütz, F., Kapp, E. A., Simpson, R. J., and Speed, T. P. (2003) Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochemical Society Transactions* **31**, 1479–83.
60. Narasimhan, C., Tabb, D. L., VerBerkmoes, N. C., Thompson, M. R., Hettich, R. L., and Uberbacher, E. C. (2005) MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Analytical Chemistry* 77, 7581–93.
61. Sun, S., Meyer-Arendt, K., Eichelberger, B., Brown, R., Yen, C.-Y., Old, W. M., Pierce, K., Cios, K. J., Ahn, N. G., and Resing, K. A. (2007) Improved validation of peptide MS/MS assignments using spectral intensity prediction. *Molecular and Cellular Proteomics* **6**, 1–17.
62. Barton, Sheila J. Whittaker, John C. (2009) Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. *Mass Spectrometry Reviews*, **28**(1), 177-187.
63. Zhou Cong, Bowler Lucas D., Feng Jianfeng (2008) A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC bioinformatics*, 9 325.