Published in final edited form as:
Mass Spectrom Rev. 2015 ; 34(1): 43–63.

LESSONS IN *DE NOVO* PEPTIDE SEQUENCING BY TANDEM MASS SPECTROMETRY

Katalin F. Medzihradszky[1,2] and **Robert J. Chalkley**[1]

[1]Mass Spectrometry Facility, Department of Pharmaceutical Chemistry, School of Pharmacy, University of California San Francisco, 600 16th Street, Genentech Hall N474A, San Francisco, CA 94158-2517

[2]Laboratory of Proteomics Research, Institute of Biochemistry, HAS Biological Research Centre, Szeged, Hungary

Abstract

Mass spectrometry has become the method of choice for the qualitative and quantitative characterization of protein mixtures isolated from all kinds of living organisms. The raw data in these studies are MS/MS spectra, usually of peptides produced by proteolytic digestion of a protein. These spectra are “translated” into peptide sequences, normally with the help of various search engines. Data acquisition and interpretation have both been automated, and most researchers look only at the summary of the identifications without ever viewing the underlying raw data used for assignments. Automated analysis of data is essential due to the volume produced. However, being familiar with the finer intricacies of peptide fragmentation processes, and experiencing the difficulties of manual data interpretation allow a researcher to be able to more critically evaluate key results, particularly because there are many known rules of peptide fragmentation that are not incorporated into search engine scoring. Since the most commonly used MS/MS activation method is collision-induced dissociation (CID), in this article we present a brief review of the history of peptide CID analysis. Next, we provide a detailed tutorial on how to determine peptide sequences from CID data. Although the focus of the tutorial is *de novo* sequencing, the lessons learned and resources supplied are useful for data interpretation in general.

Keywords

peptides; fragmentation; *de novo* sequencing; CID; HCD

I. INTRODUCTION

A. Present/future role of *de novo* sequencing

With the ever-increasing number of complete genomes published, one might think that there is now less need for *de novo* protein sequence determination from mass spectrometry fragmentation data. However, each species features slightly different sequences due to single nucleotide/residue variants and splice-variants. The increased sensitivity of instrumentation

Correspondence to: K.F. Medzihradszky, Mass Spectrometry Facility, UCSF, 600 16th Street, Genentech Hall, suite N472A, Box 2240, San Francisco, CA 94158-2517. folkl@cgl.ucsf.edu.

is also revealing a multitude of unpredicted post-translational and sample-handling modifications, which if not specified as possible during database searching, will not be identified. In addition, protein prediction from genomes is partly based on homology. Thus, really unique, species-specific sequences might stay undiscovered. For example, a BLAST search performed with the first 29 amino acids of a snake venom toxin could not find any similar sequence in the NCBI database. Even the full-length sequence (61 residues) produced only one remotely similar structure: a venom peptide from another snake [Bohlen et al., 2011]. Similarly, other relatively small, biologically active polypeptides, such as toxins and antibacterial agents, although coded in the genome, cannot be readily predicted. Thus, at minimum, determining sequence tags might be necessary.

Last, but not least, a generation of proteomics researchers has grown up relying heavily on automated data interpretation, and does not know enough about the fragmentation processes that underlie the results. This lack of hands-on experience prevents the critical evaluation of automated search results, and still frequently manifests itself in the acceptance and publishing of dubious or obviously incorrect assignments. The situation is more problematic for post-translational modification (PTM) analysis, especially when multiple different modifications are considered during a search, and permitted on a single peptide. Some common-sense rules have been suggested when someone should get suspicious about the automated sequence assignments and look for an alternative interpretation [Stevens et al., 2008; Chalkley, 2013]. The varied experience of researchers is one of the reasons why proteomics journals request assigned spectra and raw data to be deposited for single-peptide-based protein identifications and especially when reporting post-translational modifications.

B. Historical overview of *de novo* sequencing

Enkephalins are frequently used as mass spectrometry standards, or convenient small peptides to study [Sztaray et al., 2011]. Most researchers are not aware that these structures were deciphered using mass spectrometry [Hughes et al., 1975]. At that time, peptide structural elucidation using mass spectrometry was no small feat; extensive derivatization was required to make even small peptides volatile enough to be detected in a mass spectrometer. Mass spectrometry was an ‘exotic’ analytical technique for protein chemists, because peptides, just like most other biologically interesting compounds, decomposed rather than ionized when the available ionization techniques (electron impact, chemical ionization) were applied. The analysis of earlier ‘off-limits’ biomolecules became possible with the advent of soft ionization techniques, Fast Atom Bombardment (FAB) [Barber et al., 1981] and Liquid Secondary Ion MS (LSIMS) [Benninghoven & Sichtermann, 1978]. The rules and nomenclature of peptide fragmentation were established with these now-obsolete techniques. The first nomenclature was proposed by Roepstorff and Fohlman [Roepstorff & Fohlman, 1984], but the recently accepted terminology was established by Biemann [Biemann, 1990]. This time period was the prime time of peptide *de novo* sequencing, and most analysis was done manually. There were two ‘competing’ schools that were perhaps equally successful. Magnetic 4-sector instruments used monoisotopic precursor-ion selection and high-energy CID, whereas triple-quadrupole instruments yielded low-energy CID spectra with lower resolution and mass accuracy, but with higher sensitivity. The use of high-energy CID for peptide sequence determination was pioneered by Biemann [Scoble, Martin & Biemann, 1987; Biemann & Scoble, 1987; Johnson & Biemann, 1987]. Johnson's thesis research resulted in an impressive list of fragmentation rules [Johnson, 1988; Johnson, Martin & Biemann, 1988]. Burlingame was the other leading figure in *de novo* sequencing using high-energy CID [Medzihradszky et al., 1992; Wen et al., 1992]. Hunt and his team were the most prominent users of low-energy CID data [Hunt et al., 1986; Hunt, Zhu & Shabanowitz, 1989]. They also introduced efficient chemical derivatization methods in order to aid spectral interpretation [Krishnamurthy et al., 1989].

The Nobel prize-worthy new ionization techniques of matrix-assisted laser desorption/ ionization (MALDI) and electrospray ionization (ESI) revolutionized biological mass spectrometry. MALDI-Time-of-Flight (TOF) MS is the easier-to-use technique, and the monomolecular decomposition of peptides, triggered by the ionization (i.e., post-source decay (PSD)) may yield spectra suitable for *de novo* sequence determination [Medzihradszky, 2005]. However, this ionization combined with collisional-activation and more sophisticated instrumentation (i.e., MS-MS analysis on MALDI-QTOF instruments [Baldwin et al., 2001]; MALDI-TOF-TOF instruments [Medzihradszky et al., 2000]; or MALDI-FTICR instruments [Spengler, 2004]) produces much more informative results.

Nevertheless, ESI coupled with on-line fractionation methods is the most widely used peptide analytical technique. The method can be used on low- and high-end instrumentation, and MS/MS analysis is usually performed on-line in a computer-controlled, completely automated fashion. With the availability of such a powerful technique for protein analysis, mass spectrometry laboratories engaged in such research appeared all over the world. The sheer amount of data generated by ESI-LC/MS/MS prompted a revolution in data interpretation; a series of different search engines and automated *de novo* sequencing programs were created (discussed below), and for larger datasets statistical analyses were developed to assess the reliability of reported assignments. Manual *de novo* sequencing is still performed, but is much less common [Perlson et al., 2004]. In addition, two new MS/MS activation methods have been developed: electron-capture and electron-transfer dissociation, ECD [Zubarev et al., 2000] and ETD [Syka et al., 2004], respectively, where a radical ion is formed and undergoes fragmentation to yield almost exclusively peptide backbone fragmentation *via* cleavages between the amino groups and the alpha carbons. The spectra produced by these activation methods are complementary to CID data. Thus, the newest trend in *de novo* sequencing is the combination of different MS/MS activation techniques. High resolution CID and ECD spectra provided sufficient information for the *de novo* sequencing of a bacterial protein [Branca et al., 2007]. Since the bacterial kingdom is huge and divergent, the sequenced genomes may not reflect this biodiversity in the foreseeable future. Thus, bacterial proteins could be the primary targets for *de novo* sequence determination (the case study presented in this manuscript is a bacterial enzyme). Other common targets are toxins from the venom of obscure species and medium sized polypeptides, where the fragments are measured with high mass accuracy. Most of the time, data interpretation is performed by a combination of automated sequencing and manual evaluation [Jia et al., 2012; Medzihradszky & Bohlen, 2012; Samgina et al., 2008].

In the sea of data produced by high-throughput proteomics, new peptide fragmentation rules were discovered [Chalkley, Brinkworth & Burlingame, 2006; Godugu et al., 2010; Simón-

Manso et al., 2011; Kilpatrick et al., 2012; Kelstrup et al., 2011; Medzihradszky & Trinidad, 2012]. However, these results rarely get incorporated into the software used for data interpretation. Furthermore, most of the search engines and *de novo* sequencing programs still do not permit relative mass accuracy for fragment ions, even though its benefit is obvious, and there are numerous mass spectrometers that afford mass accuracy within a few ppm. Accurately measured fragments limit the amino acid combinations that need to be considered within the peptide [Spengler, 2004; Schlosser & Lehmann, 2002]. Accurate assignments of small N- and C-terminal ions, such as $\mathbf{y_1}$, $\mathbf{y_2}$, and $\mathbf{b_2}$ in CID or terminal $\mathbf{z}^{\cdot}$ and **c** fragments in ECD/ETD may provide *per se* ‘unidirectional’ or ‘bidirectional’ sequence information [Schlosser & Lehmann, 2002; Medzihradszky & Bohlen, 2012]. It also has been stipulated that with adequate mass accuracy $\mathbf{z}^{\cdot}$ fragments could be differentiated from other fragment ion types due to their unique chemical composition [Hubler at al., 2008].

From the beginning of mass spectrometry-based sequencing there were attempts to use chemical derivatization to influence/control peptide fragmentation [Roth et al., 1998] or to distinguish one ion series from the other [Shevchenko et al., 2000; Muenchbach et al., 2000; Gu et al., 2003; Gao et al., 2012]. One of the simplest and most convenient derivatizations is the incorporation of ^{18}O into the C-terminal residue during or after digestion with a protease, which produces characteristic doublet peaks for all C-terminally derived fragment ions [Schnölzer, Jedrzejewski & Lehmann, 1996]. However, detailed discussion of the advantages and limitations of such derivatizations is beyond the scope of this article. Reviews covering this subject have been published elsewhere [Roth et al., 1998; Standing, 2003; Seidler et al., 2010].

II. *DE NOVO* SEQUENCING 101

A. Amino acids and peptide fragments

Table 1 provides some basic information on the 20 amino acids directly encoded by the universal genetic code. Figure 1 and Tables 2 and 3 convey some basic mass spectrometry information, such as fragment ion structures, occurrence, and how to calculate their masses. Some fragments provide information about the amino acid composition of a peptide. These ions are the immonium ions ($^{+}NH_2{=}CHR$) (Table 4), and the fragments that are formed as a result of full or partial side-chain losses from the precursor ion. The immonium ions are labeled with the one-letter code of the amino acid residue, whereas the side-chain loss fragments are usually assigned with the masses lost.

The major sequence ions are formed *via* peptide backbone cleavages. Fragmentation might occur at each bond (i.e., between the alpha carbon and the carbonyl group; at the peptide bond; between the amino group and the alpha carbon). In these processes **a**, **b**, and **c** fragments are formed when the charge is retained at the N-terminus, while **x**, **y**, and **z** ions are produced with C-terminal charge retention. Some of these fragments may be ‘odd-electron’ radical ions, which are formed in ECD/ETD processes and in high energy CID (Table 2). The fragments are numbered from their respective termini, so a fragment that contains three amino acid residues will be a ‘$\mathbf{n}$’$_3$ ion. As discussed below, the sequence ions might be accompanied by satellite ions formed *via* the loss of small neutral molecules. An extreme manifestation of neutral loss is the **v** fragment formation from **y** ions *via* elimination of the side-chain of the N-terminal amino acid [Johnson, Martin & Biemann, 1988]. This fragmentation can be seen only in high-energy CID. Additional high-energy CID-related satellite ions are the **d** and **w** fragments, which are formed from radical **a+1**$^{\cdot}$ and **z**$^{\cdot}$ ions produced in the charge-remote processes of high-energy CID [Johnson, 1988; Johnson, Martin & Biemann, 1988]. These satellite ions might be observed in ECD/ETD spectra as well (Table 2). Last, but not least, when two peptide bonds are cleaved the short amino acid string produced is an internal fragment, which might also produce satellite ions as discussed below.

B. Rules

The most commonly used activation method in peptide analysis is collisional activation. The fragmentation mechanisms of peptides have been extensively studied, and a quite comprehensive review has been published [Paizs & Suhai, 2005]. At the same time, even die-hard theoretical experts agree that one does not have to know all of the fragmentation pathways in order to successfully interpret data. However, some basic knowledge about fragmentation is a must. Perhaps the most important point is to clear the confusion about low- and high-energy CID analysis. High-energy CID refers to collision energies in the keV range, as performed by 4-sector or MALDI-TOF-TOF instruments. High-energy CID permits satellite ion generation from radical ions *via* carbon-carbon cleavages (i.e., **d** and **w** fragment formation), and thus, offers the potential to differentiate between isomeric Leu and Ile residues. Recently, most tandem instruments (except *some* MALDI TOF-TOF mass spectrometers as mentioned above) use low-energy collisional activation, where the collision energy is in the tens of eV range, even when it is called “higher-energy C-trap dissociation”; i.e., HCD. The collision energy applied during HCD might be the same as in ion trap CID experiments - the difference lies in energy being imparted into fragment ions. Activation conditions in beam-type collision cells (quadrupoles in QQQ or Qq-TOF geometry mass spectrometers and HCD in Orbitraps) accelerate all ions across the chamber, permitting multiple collisions. Thus, fragments might fragment further to create products of two bond cleavages. As **b**-type fragments are structurally less stable than **y** ions, they have a lower rate of survival than **y** ions upon these additional activation steps [Lau et al., 2009]. As a result, large **b** ions generally fragment further to form smaller **b** ions, internal, and immonium ions. Certain **y** fragments also will yield internal ions [Ballard & Gaskell, 1991]. These extra fragment ion types might aid or complicate data interpretation.

Ion trap CID is performed by resonance-activation of the precursor *m/z* only [Jonscher & Yates, 1997]. Once a bond is cleaved, internal energy is released, and if the *m/z* of the fragment ion differs from the precursor *m/z*, then the product will not be further activated, so should not fragment further. Thus, ion trap CID spectra feature more **b** ions than collision-cell-derived data, but no internal or immonium ions. At the same time, the low mass range (up to 1/3rd of the *m/z* of the precursor) cannot be trapped under normal circumstances [Jonscher & Yates, 1997]. Thus, valuable information about the termini might be lost. Rules that govern peptide fragmentation are most comprehensively summed up in the ‘mobile proton’ model developed by Gaskell and Wysocki [Cox et al., 1996; Dongré et al., 1996; Wysocki et al., 2000; Paizs & Suhai, 2005]. In low-energy CID the fragmentation is usually controlled by charge-directed processes [Burlet et al., 1992; Tang, Thibault & Boyd, 1993]. Peptides feature many potential protonation sites, but basic residues preferentially retain the proton. In singly-charged precursor ions this preferential charge retention will seriously limit the protonation of other sites, especially when Arg (that displays the highest proton affinity) is present, and thus, charge-remote fragmentation may occur. Charge-remote fragmentation is a common phenomenon in high-energy CID, and satellite ions **d**, **v**, and **w** are formed this way [Johnson, Martin, & Biemann, 1988; Alexander, Thibault, & Boyd, 1989]. In low-energy CID, the aspartic acid effect (discussed later), and the fragmentation-promoting effect of cysteic acid [Burlet, Yang & Gaskell, 1992] also represent charge-remote fragmentation processes. In multiply-charged peptides there may be more charges than basic residues. In this situation, one proton may be firmly anchored at a basic residue, while the additional charge(s) can migrate along the peptide and promote fragmentation. Due to the C-terminal Arg or Lys and, thus, preferential charge retention at this site, tryptic peptides normally display abundant **y** fragments. Data acquired from 2+ precursor ions are the easiest to interpret because other basic residues are usually not present, so the second proton is mobile and produces fragmentation all along the peptide sequence. The more abundant fragment ions, especially above the precursor mass, are usually **y** ions, especially in collision-cell-derived CID spectra. There are also sequence-dependent cleavage preferences that have been characterized from statistical analysis of CID spectra [Kapp et al., 2003; Huang et al., 2005]. Unusually abundant ions might indicate a Pro in the sequence: cleavage at its N-terminus usually leads to an intense **y** ion formation, and the corresponding **b** ion is frequently also abundant. Conversely, the **y** fragment formed *via* cleavage at the C-terminal side of Pro is usually weak in intensity or missing, as is the corresponding **b** ion. Gly residues in general also tend to yield abundant **y** fragments, while the cleavage at their C-terminus tends to be suppressed, resulting in missing or weak **y** and **b** fragments, just like for Pro residues. Asp promotes fragmentation by donating its own proton to the peptide bond to drive the cleavage rather than relying on the mobile proton: cleavage C-terminal to Asp is a favored fragmentation step in low-energy CID, and yields the most abundant fragment in spectra where there is no mobile proton, such as in MALDI CID spectra [Wattenberg et al., 2002]. Interestingly, this side-chain-promoted cleavage might yield very abundant **b** fragments in cases of preferential charge retention at the N-terminus [Cotter et al., 2005].

Since **y** and the corresponding **b** ions are formed when a peptide bond is cleaved, there is a simple mathematical relationship between them: $b_i+y_{n-i}=MH^++1$. As mentioned earlier, ion trap CID spectra frequently display extensive **b** fragment series, whereas in collision-cell-derived spectra these ions usually peter out, unless there is a basic residue at or close to the N-terminus. What is common between the two techniques is that i) $\mathbf{b_1}$ is usually not detected (even when that mass is within the detection range of the ion trap), and ii) the $\mathbf{a_2}$−$\mathbf{b_2}$ pair is usually present and abundant. The cyclic **b**-fragment ion structure is formed by the nucleophilic attack of the carbonyl group from the neighboring, N-terminal amino acid residue [Yalcin et al., 1995; Harrison, 2009]. This carbonyl group is absent for an N-terminal residue in an N-terminally unmodified peptide. However, when the N-terminus is acylated, the $\mathbf{b_1}$ fragment can be detected [Medzihradszky, 2005]. For example, acetylation of protein N-termini is common in eukaryotes, and according to the rules of biological processing Met, Gly, Ala, Ser, and Thr may be acetylated [Bradshaw, Brickey, & Walker, 1998] and, thus, might yield $\mathbf{b_1}$ ions. Peptides derivatized on primary amino groups with isobaric tagging reagents such as iTRAQ® (Isobaric tag for relative and absolute quantitation, Sciex) and TMT™ (Tandem Mass Tag, Thermo Scientific) can also form $\mathbf{b_1}$ ions. Although the presence of the $\mathbf{a_2}$−$\mathbf{b_2}$ fragment pair is nearly universal (provided they are in the mass range of ions that can be trapped), ion trap CID spectra do not feature abundant larger mass **a** ions, while **b**-type fragments in the collision-cell-derived CID data frequently can be identified by their “satellite” **a** fragments, at a mass lower by 28 Da. The ‘**b**-type’ fragment designation here also encompasses **b**-type internal ions, which also might be accompanied by **a**type ions.

Major sequence ions (i.e., **y** and **b** fragments) frequently display abundant neutral losses. Both series might lose ammonia or water, depending on their amino acid composition (see Table 2). However, in a given spectrum it might happen that only one of the ion series features these neutral losses abundantly and consistently and, thus, the presence of such satellite ions might aid the identification of fragments that belong to the same ion series.

In peptides with preferential charge retention at the N-terminus a rearrangement reaction might take place that leads to loss of the C-terminal residue and formation of a $\mathbf{b_{n-1}+H_2O}$ fragment. Sometimes, the penultimate amino acid also might be eliminated this way [Thorne, & Gaskell, 1989; Thorne, Ballard, & Gaskell, 1990]. This fragmentation is very characteristic, and can be observed in all CID experiments; recognizing its occurrence should help during *de novo* sequencing [Medzihradszky & Bohlen, 2012]. As far as we know, among the search engines, only Protein Prospector (prospector.ucsf.edu) and Spectrum Mill (http://spectrummill.mit.edu/) consider and report this fragment.

The mass difference between members of the same ion series correspond to amino acid residues. However, one cannot differentiate between isomeric Leu and Ile. There is also an isobaric amino acid pair: Gln and Lys differ by 36 mmu. Thus, if the fragment masses are measured with sufficient accuracy, then these residues can be distinguished. However, the fragments that contain these residues must be measured accurately enough; whereas 36 mmu corresponds to an approximately 120 ppm mass difference at *m/z* 300, the same absolute mass deviation is only a 25 ppm difference at *m/z* 1500. To complicate the matter further, this mass with the very same elemental composition corresponds not only to Gln, but also to an AlaGly combination. Unfortunately, one cannot rely on the fragmentation of every peptide bond, and similar isomeric and isobaric pairs exist. For example: AlaAsn and GlyGln; SerGlu and ThrAsp are isomeric; while Phe and Met(O); Trp, GlyGlu and SerVal; or Arg and GlyVal represent isobaric combinations. Similarly, fragment ions from different ion series might yield identical nominal masses. For example, to see a $\mathbf{b_3}$ ion composed of TGS (246.1084) separate from a $\mathbf{y_2}$ fragment from AR (246.1561) one would need more than 10000 resolution, even though each mass can be measured quite accurately at lower resolution if the other ion is not present (at this *m/z* value, a mass accuracy within 100 ppm is sufficient to distinguish between these compositions). When internal ions are formed, their presence might lead to the overlapping of a series of different fragments requiring high mass resolution in addition to good mass accuracy in order to separately detect these ions.

Low-energy CID described here produces the above-mentioned ions. ECD and ETD spectra also display some **y** fragments and a few **b** ions [Chalkley et al., 2010], but are dominated by $\mathbf{z}^{\bullet}$ and **c** ions and also feature **z**+1 and **c**−$1^{\bullet}$ fragments, formed *via* a hydrogen shift [Bakken, Helgaker & Uggerud, 2004]. In certain instances, **c** ions might be detected in collision-cell CID experiments: usually, an abundant $\mathbf{c_1}$ fragment is observed whenever Gln is in the second position from the N-terminus [Lee & Lee, 2004]. Based on the proposed mechanism, Ser, His, Lys and Arg also might trigger **c**-ion formation from any N-terminally adjacent amino acid [Farruggia, O’Hair & Reid, 2001]. However, **c** fragments have only been reported for residues N-terminally adjacent to Gln and Lys [Medzihradszky & Bohlen, 2012].

Formation of the additional fragments, **d**, **x**, **v**, and **w** shown in Figure 1 require high collision energy and/or radical processes. Thus, they can be observed in high-energy CID and in some ECD/ETD experiments.

While recently the alternative activation techniques of ECD and ETD have become popular for proteomic analysis, our understanding of the rules that govern ECD and ETD fragmentation processes is much more limited than for CID, as is our experience with ECD and ETD data alone for *de novo* sequence determination. The combination of data acquired from different activation methods could be advantageous, as mentioned earlier. For example, radical-ion fragmentation yields **c** and $\mathbf{z}^{\bullet}$ ions, 17 Da higher or 16 Da lower than their N-terminal and C-terminal equivalents in CID; i.e., **b** and **y** fragments, respectively. Thus, based on these mass differences, corresponding members of the different ion series can be identified. While Pro residues (imino acids) cannot form a $\mathbf{z}^{\bullet}$ fragment in ECD/ETD, and for the very same reason their N-terminal neighbors cannot produce **c** ions, CID might supply that missing information, because the N-terminal side of Pro is a favored fragmentation site. High resolution, high mass accuracy data produce more reliable results in the CID/ETD combination approach, as in all *de novo* sequencing efforts. At the same time, ETD spectra acquired in an ion trap are of higher sensitivity, and might contain more fragment ions. In addition, the best CID spectra are usually derived from 2+ precursors, whereas efficient ETD fragmentation requires higher charge-state precursors at lower *m/z* values (i.e., relatively high charge density) [Good et al., 2007]. Thus, mostly larger polypeptides from different toxins have been sequenced utilizing both CID and ETD data. For such large molecules the interpretation process can be rather complicated [Medzihradszky & Bohlen, 2012].

However, most researchers have to rely on a single technique. Better understanding of CID data might lay the groundwork for the successful application of more sophisticated sequencing/data interpretation workflows. In this article we will discuss CID-based *de novo* sequencing in depth.

C. Resources

Predicting fragmentation or assigning spectra—Free tools such as Protein Prospector’s (www.prospector.ucsf.edu) “**MS-Product**” program calculate the predicted masses of MS/MS fragments, and one can upload an associated peak list on which to annotate the specified sequence (one can even compare different sequence assignments to the same spectrum). A wide variety of covalent modifications can be specified, as named modifications or simply as defined mass additions. User-defined amino acids are also permitted, and the computer will provide a comprehensive list of instrument-specific fragments. MS-Product is "conservative" in that it counts the number of basic residues in the fragment in order to decide how many charges on that fragment should be permitted. Also, as a default, it will list only single neutral losses from sequence or internal fragments. Although ammonia loss can be observed from **y** fragments in general, **MS-Product** follows a rule of reporting 17 Da losses only when there is Lys, Arg, Asn, or Gln in the sequence; and water loss is listed when the fragments contain Ser, Thr, Glu, or Asp. Most of these experience-based rules have been confirmed by mechanistic studies, as well as statistical analysis of CID data [Paizs & Suhai, 2005; Sun et al., 2008]. Though there are exceptions to these rules, this conservative listing and labeling generally leads to a greater reduction in spurious peak labeling than missing of valid labels, and "creative" peak-assignment is sometimes a problem [Stevens et al., 2008]. Unfortunately, **MS-Product** does not provide any information about the probability whether any of the fragments will be detected and with what intensity. Statistical analyses of fragmentation spectra have provided some information that can be used to predict the appearance of ion trap CID fragmentation spectra [Kapp et al., 2003; Huang et al., 2005] and software has been developed to model ion trap CID and ETD spectra appearance [Zhang, 2004, 2010, 2011].

Mass-based composition prediction—Protein Prospector also has a program, **MS-Comp**, that will match measured masses, and ion types selected with amino acid compositions within a specified mass accuracy. Due to the highly similar elemental composition of peptides in general, unambiguous assignments can only be achieved for small, accurately measured fragments. For example, two low-mass fragments, *m/z* 197.0797 and 212.1393, identified both termini of a peptide as $\mathbf{z_2}$ for {HisGly} and $\mathbf{c_2}$ for ProPro, respectively, because these were the only potential ETD fragments within the 5 ppm mass accuracy at which these ions were measured [Medzihradszky & Bohlen, 2012]. This program is especially useful for determining the potential composition of **b**-type fragments, sequence, or internal ions in the low mass region. The program can be used also for higher mass fragments when the identity of the ion series is known and most of the residues within the ion have been determined, meaning only a small gap has to be filled.

Finding spectral families—Within a dataset there are often spectra of related peptides, whether they derive from unmodified and modified versions of the same sequence, or due to the presence of a missed enzyme cleavage site in one version. Finding these families of related spectra can be useful for *de novo* sequencing, because comparison of spectra often allows identification of which ions are part of the same series based on whether they are systematically shifted in one spectrum. A program such as Protein Prospector's **MS-Filter** can be used to filter peak lists to include only those spectra that contain a fragment deemed diagnostic (with the assumption that it will remain unchanged). Hence, this tool might help to identify spectral families.

Homology considerations—Once a reliable sequence tag has been determined, performing a database search is a logical next step in order to try to find a protein that contains the sequence interpreted or something homologous. **BLASTp** is a biologist's tool of choice for sequence comparisons. However, **BLASTp** does not work well for short sequence tags of 5–8 residues, which is the typical length of tag one confidently can decipher from CID data. **MS-BLAST** is slightly more effective, and one can test multiple data interpretation versions at the same time (http://dove.embl-heidelberg.de/Blast2/msblast.html). However, **MS-Homology** in Protein Prospector offers a more powerful option that is tailored toward the strengths and weaknesses of mass spectrometry data. **MS-Homology** allows indication of uncertainty about the order of residues: {FT} indicates these two residues in either order; it allows alternatives at a single site: [L|I] means either Leu or Ile; and it allows the inclusion of mass gaps: [248.02] means the program would consider any amino acid combination that leads to this mass. One can specify the number of permitted amino acid substitutions, and one can also search with multiple sequences at the same time. Obviously, if the same protein is reported for multiple unknowns, then this result provides further confidence in the homologous protein assignment. For scoring results, as a default it uses the BLOSUM62 matrix [Henikoff & Henikoff, 1992], which weights results toward more conservative amino acid substitutions (e.g., replacement of a small neutral Gly with an Ala is more likely than with a bulky, charged Arg).

Tutorials—An excellent tutorial on *de novo* sequencing with an emphasis on software was published recently as part of a set of tutorial articles commissioned by the education committee of HUPO [Ma & Johnson 2012]. There is also an earlier review that could be useful for the interpretation of covalently modified peptide spectra [Medzihradszky 2005].

D. Practical advice

1. Always work from the raw data – merged spectra might provide better quality data; however, they might also mask the fact that there were two isoforms present (for example, phosphopeptides modified at different positions). Similarly, data processing might alter the data and eliminate information, especially about weak or overlapping peaks. For example, de-isotoping will simplify the peak list, but might also eliminate fragments that happen to be ~1 Da higher in mass than another ion.

2. Careful examination of the precursor ion region in the survey MS spectrum is also warranted. In single-protein digests, coeluting overlapping isotope clusters usually do not cause trouble, but they do occur [Medzihradszky, 2005]. In complex mixtures one practically always works with mixture spectra.

3. When there are spectra from multiple charge states, use all of them: different charge states frequently yield slightly different information, and the combination thereof might lead to complete sequence determination that perhaps would not have been possible from one charge state alone.

4. Keep an eye out for spectral families – one can easily recognize CID data that represent related peptides. Based on mass differences between the precursor ions as well as retention time differences, one can guess the relationship between the different molecules.

Missed or non-specific cleavages, post-translational modifications, chemical side-reactions, as well as adduct formation or in-source fragmentation might be the reasons for the presence of such related molecules.

5. Once a sufficiently long sequence (≥ 5 residues) has been determined reliably, perform an MS-Homology search. Obviously, if there is a template in the form of a homologous sequence, then the remaining job is much easier, and identifying only 2 residues might point to the correct sequence stretch and, thus, could speed up the sequencing process. In addition, even short, non-unique sequences can be located within the protein.

6. Final sequences must be “verified”. Try to assign all abundant fragment ions. First of all, instrument-specific fragments listed by MS-Product should be considered. Then, “unusual” fragment ions (i.e., those unique to certain sequences or covalent modifications) should be considered. A significant number of such obscure fragmentation rules have been reported. Most of these rules are ignored by almost all proteomics data-interpretation software, as well as by the people who evaluate the data. Some fragmentation rules have been described above, and some additional references have been provided. In addition, we recommend to search for publications that report on the specific fragmentation of covalently modified sequences; or on fragmentation rules of the ionization, MS/MS activation, or analyzer that were used to generate the data under investigation.

7. In the final list of interpreted sequences, try to indicate which are thought to be more or less reliable than others.

III. A *DE NOVO* SEQUENCING STORY

As an example of the advantages and limitations of manual and automated sequencing, the sequence determination of the alpha subunit of sulfocatechol 3,4-dioxygenase from *Novosphingobium resinovorum* (*Sphingomonas subarctica*) (NCBI # 56787886) is presented. The isolated protein, which consists of alpha and beta subunits, was purified by SDS-PAGE, two bands were excised, and each was analyzed using in-gel digestion and LC/MS/MS on a QSTAR Pulsar hybrid tandem mass spectrometer of QqTOF geometry, which afforded ~ 200 ppm mass accuracy for both precursor and fragment ions. Peaklists (mgf file, Supplement 1) were generated using a script supplied by AB Sciex, and the CID data were subjected to an automated database search by Mascot [Perkins et al., 1999], partly to eliminate known potential contaminants such as human keratins, and partly to probe for sequence identity with proteins already listed in the database. Most of the peptides were subsequently sequenced *de novo* from the CID data. These data are obviously old, but the lessons learned are still valid (the raw data file is Supplement 2).

The database search identified a few peptides from protocatechuate 3,4-dioxygenase type II, beta subunit pcaH-II, NCBI # 11037226 (see Figure 2) in *both* digests. This result suggested i) the intended protein was isolated; ii) it shows similarity to at least one database entry; iii) if both subunits were isolated, then they were not completely separated by the SDS-PAGE.

High quality uninterpreted spectra were manually selected, printed, and *de novo* sequencing was attempted. The definition of a good quality spectrum is somewhat abstract, although some general rules have been formulated [Nesvizhskii et al., 2006]. One usually selects spectra of abundant precursor ions whose monoisotopic masses and charge states can be determined unambiguously; and the spectra feature abundant fragment ions over a wide mass range, especially above the precursor *m/z*.

Complete sequence determination is first illustrated for a relatively short peptide. The analysis approach presented here should work well for all kinds of CID data, but one always has to bear in mind the fragmentation differences introduced by different ionization and activation methods. Resolution and mass accuracy also make a huge difference.

The spectrum of a peptide with a precursor ion of *m/z* 557.8(2+) is shown in Figure 3. From the immonium ions one can ascertain the presence of Ser, Val, and Phe residues, due to the immonium ions at *m/z* 60, 72 and 120, respectively (Table 4). The unusually abundant *m/z* 87 indicates Asn, while the weak *m/z* 84 and 101 suggest the presence of Lys and/or Gln. The presence of Lys is confirmed at once, by an abundant $\mathbf{y_1}$ fragment at *m/z* 147. Since the high mass region contains fewer peaks, it is easier to identify a sequence tag from this region. There are a series of ion-pairs separated by *m/z* 17: *m/z* 1027-1010; 970-953; 856-838, 709-692; 622-605 and 475-458. Most tryptic peptides display abundant high mass **y** ions, and collision-cell-derived CID spectra contain few high mass **b** ions, so one can guess that these peaks are **y** fragments. From the *m/z* differences of 57, 114, 147, 87 and 147 one can interpret the sequence as Gly-Asn-Phe-Ser-Phe. Remember, the **y** ions extend from the C-terminus, so reading the mass differences from high to low mass reports the peptide sequence. The next **y** fragment is most likely at *m/z* 347, *m/z* 128 below the lowest mass **y** fragment identified so far, indicating the presence of Gln or Lys. These amino acids are isobaric, and we do not have the mass accuracy to confidently tell them apart at this moment. The lower mass region is a bit “crowded”. However, based on our assignments of **y** ions, one can label some of the low-mass peaks as corresponding **b** ions using the equation: $b_1+y_{n-1}=MH^{+}+1$. Since the last **y** fragment detected was *m/z* 1027, one has to account for *m/z* 88 as a **b**-fragment. Once the “excess” proton is discounted, *m/z* 87 is left that corresponds to a Ser residue. As previously described, the $\mathbf{b_1}$ fragment cannot be formed. However, $\mathbf{a_2}$ and $\mathbf{b_2}$ are usually abundant, and a few additional **b** fragments might be detected. Since our working sequence now is Ser-Gly-Asn-Phe-Ser-Phe-Gln/Lys, one can expect **b** fragments at *m/z* 145, 259, 406, and perhaps 493. Indeed, *m/z* 145 ($\mathbf{b_2}$) and 117 ($\mathbf{a_2}$) are quite abundant, and confirm our original sequence assignment. The $\mathbf{b_3}$ fragment (*m/z* 259) is unusually abundant, but there will be an explanation for this high signal later. The last **b** fragment detected is $\mathbf{b_4}$, at *m/z* 406. Multiple collisions further fragment the larger **b** ions, as discussed earlier. Some of the abundant low-mass ions can be assigned to predicted internal fragments based on our already deciphered sequence. The fragment at *m/z* 172 corresponds to GN; and *m/z* 262 is NF. Since internal ions also might undergo neutral losses, *m/z* 234 could be assigned as a CO loss (i.e. **a** type ion) from the previously assigned *m/z* 262 fragment. This possibility is interesting, since this mass also might represent a $\mathbf{y_2}$ fragment for a SerLys C-terminus, and a Leu/Ile residue (isomeric, both with a 113 Da residue mass) would complete the sequence. The calculated mass values are 234.1237 and

234.1448 for the internal **a** fragment of NF-CO and for the potential $\mathbf{y_2}$, respectively. The mass accuracy afforded does not help us here, because the measured mass was 234.1387; i.e., displays +64 ppm and -26 ppm mass deviations, respectively. However, according to the immonium ion region, there is no Ile/Leu in the peptide, but a Val should be somewhere. If we assume the presence of a Val, and subtract its residue weight from the *m/z* 200 difference between $\mathbf{y_1}$ and *m/z* 347 (the lowest mass **y** fragment assigned so far), then the identity of the missing residue is revealed, as Thr (200-99 (Val)=101). The ion at *m/z* 246 clearly indicates that Val is the penultimate amino acid. Thus, our almost complete sequence is Ser-Gly-Asn-Phe-Ser-Phe-Gln/Lys-Thr-Val-Lys.

The identity of the 7th residue cannot be decided with the assumption that trypsin would have cleaved there if it were Lys. Current high mass accuracy instrumentation could distinguish between these residues from the $\mathbf{y_4}$ ion: a 0.036 mass difference corresponds to ~75 ppm at this *m/z* value. However, with this ten-year old QSTAR data, observation of a lower mass fragment for which 0.036 would represent a larger ppm difference would increase confidence in the ability to distinguish between Gln and Lys. **MS-Product** in Protein Prospector, with the appropriate instrument selection, will list all the potential Gln/Lys-containing fragments. The fragment at *m/z* 259, which might correspond to FQ/K-$NH_{3,}$ would be a good candidate, except that mass also represents the overlapping $\mathbf{b_3}$ ion. However, there is a fragment at *m/z* 212.110 that represents Q/KT-water, where the mass error is +33 ppm vs. −137 ppm for Gln and Lys, respectively. Thus, the final sequence is SGNFSFQTVK.

Once one has deciphered a long-enough sequence stretch, multiple options are available, as mentioned earlier. Using this sequence and permitting 3 amino acid substitutions, from among other microorganism proteins in the database, **MS-Homology** listed the following entry NCBI# 11037227, protocatechuate 3,4-dioxygenase type II alpha subunit PcaG-II [*Agrobacterium tumefaciens*] (Figure 4).

With this match, the presence of the alpha subunit was confirmed. A homologous protein sequence might aid data interpretation in multiple ways: i) by providing a reference to fill in gaps in a sequence to be determined; ii) by indicating the relative sequence position of the tryptic peptides. For interpretation of further spectra in this project, the sequences manually determined were compared to the template sequence: identity/similarity was used to position the peptides within the full sequence, to fill in gaps, and to assign terminal amino acid residues (Table 5).

The peptide SGNFSFQTVKPGR is a good example for the benefit of spectra from multiple charge states. Precursor ions at *m/z* 475.58(3+) and *m/z* 712.86(2+) both produced good-quality CID data (Figure 5), but the determination of the C-terminal sequence was not straightforward from the 3+ precursor ion spectrum because there are a large number of ions in the low-mass region and $\mathbf{y_3}$ is isobaric with two potential internal ions (QTV & TVK). In contrast, in the CID spectrum of the 2+ precursor ion there are no abundant internal ions, so the assignment of low-mass fragments is much simpler. In general, as mentioned earlier, spectra of 2+ precursor ions are easier to decipher, due to mainly containing singly-charged fragment ions. However, data from the higher charge states might provide confirmatory/

complementary information. In addition, sometimes one of the charge states might produce a mixture spectrum because of the coelution and overlap with other components. In this particular case, these data are also good examples for the identification of spectral families, and the use of that information. From a few low-mass ions, such as *m/z* 117, 145, 259, and 262, one might suspect that there is a connection between this peptide and the one in Figure 3. Although the presence of these masses might be coincidental, if the sequences are really related, then the modification must be at the C-terminus (the mass difference is too high for any common side-reaction) and it is worthwhile to investigate. The mass difference between the two peptides is 310 Da. The fragment at *m/z* 175 clearly indicates a C-terminal Arg. The remaining mass difference of 154 Da represents a single potential amino acid composition: {GlyPro}. The fact that the C-terminal extension of this tryptic peptide is identical to that of the homologous protein is additional confirmation that the sequence belongs to the alpha subunit. The presence of the shorter peptide proves that trypsin may cleave N-terminal of Pro residues.

While making the connection between these two peptides discussed above either depends on noticing a handful of low-mass ions in common or would require partial sequencing, some spectral families can be recognized immediately. The CID data of *m/z* 684.7, 690.0 and 703.7, all 3+ ions, indicated that these peptides must be related (Figure 6), because practically all the masses up to ~*m/z* 550 are identical. The precursor mass differences relative to the smallest of these three peptides indicated oxidation (+16 Da) and carbamidomethylation (+57 Da). Even without complete interpretation of the data, one can conclude that both modifications probably occur on a Met residue; i.e., that a sulfoxide was formed in the ‘middle’ peptide, and that the third spectrum indicates that a carbamidomethylation side-reaction had taken place, instead of a desired Cys alkylation. This side-reaction might occur at amino termini, and on the side-chains of Lys, His, Asp, and Met residues, and might be mistaken for an additional Gly residue. The oxidized peptide displays +16 Da shifted ion series (*m/z* +8 differences in the doubly-charged ions can be noticed) as well as 64 Da (CH_3SOH) losses from the precursor ion and some fragments; a well documented characteristic loss observed for methionine sulfoxide-containing sequences [Lagerwerf et al., 1996]. For example, the *m/z* 936 peak is *m/z* 8 higher than the *m/z* 928 in the unmodified spectrum, and *m/z* 32 higher than *m/z* 904. Interestingly, other mono-oxidized thioether-bonds in peptides (e.g., carbamidomethyl-Cys-sulfoxide) also display a similar RSOH loss [Chowdhury et al., 2007]. The carbamidomethyl derivative behaves more interestingly. The base peak in the spectrum is a triply-charged ion at *m/z* 668.8 that indicates a 105 Da loss from the precursor (an analogous structure to the 64 Da-loss ion from the methionine sulfoxide). At the same time, the fragment ions that are changed display a *m/z* −48 shift in comparison to the unmodified sequence. This mass difference is explained by the neutral loss of CH_3-S-CH_2-CO-NH_2 from the carbamidomethyl Met-sulfonium ether. The N-terminal sequence easily can be determined as VPTADGVM[Q/K]APH[I/L] from the MH^+ value (2052.1 Da) and the abundant singly-, and doubly-charged **y** fragments, that are 977.1(2+) = 1953.2; 928.5(2+) = 1856; 878(2+) = 1755; 842.5(2+) = 1684; 785(2+) = 1569; 756.5(2+) = 1512; 706.9(2+) = 1412.8 that was also detected singly-charged; 1281.8; 1153.8; 1082.7; 985.6; 848.5; 735.5 in the top, unmodified peptide spectrum. Similarly, assigning the 3 C-terminal residues is simple because the *m/z* 147 fragment identifies a Lys as the last amino acid, the abundant *m/z* 204 signal indicates a Gly next to it and an abundant *m/z* 351 ion points to Phe as a probable extension. At this point, our working sequence is VPTADGVM[Q/K]APH[I/L] …FGK. Since there is at least one basic amino acid in the middle of the sequence (His), one could look for large **b** fragments displaying the appropriate mass shifts in the modified peptides. (The in-sequence basic amino acid increases the likelihood of survival of higher mass **b** ions.) However, possession of a homologous protein sequence offers a simpler solution. The sequence determined up to this point seems to fit to our alpha template as the C-terminal extension of the earlier characterized tryptic peptide, and suggests ALSI as the remaining residues. These residues together yield exactly the residue-mass combination uninterpreted. The **y** fragments based on this assumed sequence, at *m/z* 464, 551, and 664, were detected, as were the corresponding doubly-charged $\mathbf{b_{14}}$, $\mathbf{b_{15}}$ and $\mathbf{b_{16}}$ ions at *m/z* 694.9, 751.5, and 794.9, respectively. CID data of the modified peptides also display some of these **b** ion fragments shifted, but identification of these few diagnostic fragments among much more abundant multiply-charged ions would have been much harder than completion of the sequence from the homologous template. The identity of the 128 Da residue was also assigned based on the sequence homology.

Figure 7 displays an example for partial sequence determination. The precursor ion was *m/z* 656.8(2+). In this spectrum there are high mass ion pairs, separated by *m/z* 18 each, that probably represent a sequence ion series. These are identified as N-terminal **b** ions and the corresponding water-loss ions, because for *m/z* 1094, 947, and 800 the corresponding **a** fragments were also detected (*m/z* 1066, 919 and 772, respectively). With the mass of the last **b** ion detected at *m/z* 1223, the corresponding $\mathbf{y_1}$ fragment is determined ($\mathbf{y_1}$+$\mathbf{b_{n-1}}$=MH$^+$+1) as *m/z* 90, and indeed this ion is present in the spectrum, identifying an Ala as the C-terminal residue. The rest of the C-terminal sequence can be read from the **b** fragments: *m/z* 1223, 1094, 947, 800 and 701 as Glu, Phe, Phe, Val (going from the C- to the N-terminus). Corresponding **y** ions were detected at *m/z* 219 ($\mathbf{y_2}$), 366 ($\mathbf{y_3}$), and 513 ($\mathbf{y_4}$), along with abundant internal fragments. One can add one more residue to this string: a Thr, based on the presence of *m/z* 600 and 582, completing the C-terminal sequence tag as … TVFFEA. Without the template reference sequence, this tag might not be sufficient. However, due to the high degree of similarity between the two sequences in this position one can claim that the C-terminus of the alpha subunit has been found. Once the full sequence is known, LQGDGETTVFFEA (Figure 8), practically all the ions in the spectrum can be annotated. There is a $\mathbf{c_1}$ ion at *m/z* 131, as one would expect, because Gln is the second residue [Lee & Lee, 2004]. From the **b** fragment series only $\mathbf{b_3}$ and $\mathbf{b_5}$ are absent. These ions would have been formed *via* cleavages at the C-termini of Gly residues, which typically produce low intensity **b** fragments [Kapp et al., 2003; Huang et al., 2005]. No additional **y** fragments were detected; the rest of the fragments are internal ions and fragments formed *via* neutral loss(es) from these and the **b** ions. This peptide is a good illustration of the difficulties one faces when sequences without basic residues must be sequenced/identified.

The complete list of sequenced peptides is presented in Table 5. From this list it was concluded that this particular digest contained mostly the equivalent of the “alpha subunit

PcaG-II", and the relative sequence position of the tryptic pieces could be predicted based on the homology observed, as indicated in Figure 8. The lack of tryptic cleavage sites almost completely prevented obtaining sequence information about the N-terminal part of the protein. This illustrates that the use of only trypsin for most proteomic analyses is a significant handicap to comprehensive analysis. The tryptic peptide FAGAHPELR also illustrates that truly unique sequence stretches might be encountered in the sequencing process, when one cannot be sure whether the peptide indeed belongs to the protein of interest.

IV. *DE NOVO* SEQUENCING SOFTWARE

A. Brief historical overview

The first computer programs that aided data interpretation were developed in parallel with the establishment of the rules of fragmentation and with the first *de novo* sequencing studies [Johnson & Biemann, 1989; Hines et al., 1992]. One of the first MS/MS database search tools, PeptideSearch, used as input a three amino acid *de novo* sequenced tag, along with the masses of the uninterpreted N- and C-terminal regions on either side of this tag [Mann & Wilm, 1994]. Later, with the advent of nanospray LC/MS/MS analyses and high-throughput proteomics, the importance of *de novo* sequencing became even more apparent, and numerous *de novo* sequencing programs were developed, such as Lutefisk [Taylor & Johnson, 1997; 2001], SHERENGA [Dancík et al., 1999], PEAKS [Ma et al., 2003], PepNovo [Frank & Pevzner, 2005], EigenMS [Bern & Goldberg, 2006] and pNovo [Chi et al., 2010]. From the beginning, *de novo* sequencing was not only aimed to decipher novel sequences, but also to speed up database searches [Shevchenko, Wilm, & Mann, 1997; Taylor & Johnson, 1997; Bern, Cai & Goldberg, 2007; Zhang et al., 2012]. This approach is especially promoted by Pevzner and his colleagues [Wielsch et al., 2006; Waridel et al., 2007; Kim et al., 2009].

With different MS/MS methods available on the same instrument and, thus, reliably on the same precursor ions, complementary information can be obtained from CID and ECD/ETD. The practicality and benefit to combine these data for complete, reliable sequence determination has been eloquently pointed out by Zubarev [Zubarev, Zubarev & Savitski, 2008]. His group developed the first software performing *de novo* sequencing by combining these data [Savitski et al., 2005], and other groups followed suit [Datta & Bern, 2009; Chi et al., 2012].

B. Automated *de novo* sequencing with PEAKs

PEAKS is a commercially available search engine that also features a frequently used *de novo* sequencing program. In order to compare computer-generated results to manual sequencing data, a peaklist in the form of an unprocessed mgf file for the same data as summarized in Table 5 was uploaded to the program. The software processes the data, deconvolutes obviously multiply-charged ions, and removes some of the isotope peaks. The software requires instrument and MS/MS method selection. One must be aware that the methods assigned are not necessarily valid for the instrument selected. For example, for the QTOF instrument the program indicated that the MS mass measurement was performed in the quadrupole (rather than by TOF), and the CID selection corresponds to ion trap CID (rather than quadrupole CID), irrespective of the analyzer selected. Thus, the data were interrogated as CID as well as HCD data in order to determine whether proper activation selection would yield significantly different results. In addition, the peptides were sequenced as tryptic peptides or without enzyme specification; i.e., allowing for non-specific cleavages. The precursor mass accuracy was set at 200 ppm, and the fragment mass tolerance was set to 0.2 Da. The mass deviation permitted for fragments represents a much wider tolerance than desired for the low-mass fragments, which were measured within 200 ppm, but the software did not permit relative mass accuracy selection for the fragment ions. Carbamidomethylation of Cys residues was indicated as a fixed modification, while cyclization of N-terminal Gln, and oxidation of Met, Trp and His residues were considered as options. (Although oxidation of Trp residues is almost as common as Met sulfoxide formation, this option was linked with His oxidation, which is not such a common occurrence.) Supplement 3 lists the results for each sequencing attempt, while the second worksheet shows a filtered list in comparison with the database search and manual sequencing results (Supplement 4 features all these spectra, except the ones included in this article as Figures). Of the 27 spectra that were manually *de novo* interpreted, PEAKS correctly interpreted seven and for a further five the only errors were isobaric substitutions (e.g., oxidized Met instead of Phe). For a further six, more than half of the respective sequence was correctly interpreted; there were only nine examples in which the interpretations were not close. Although these numbers give the impression that software performance falls far short of what can be achieved manually, it is a slightly unfair comparison, because a homologous database sequence was used to complete missing parts in the manual process. Other human advantages were obvious. The software had to consider each spectrum independently, while manually one can combine data acquired from different charge states and from differently modified peptides, as described earlier. Obviously, peptides with missed cleavages also belong to this category if one considers a broader definition for modifications, so the N-terminally or C-terminally elongated sequences can be included in this category. Consideration of spectral families is an approach that can be implemented for computer programs as was described for ion trap CID data [Bandeira et al., 2007]; however, as far as we know it is not currently utilized by *de novo* sequencing software. Another obvious weakness for this software was the lack of an option to specify relative mass accuracy for fragments. As demonstrated above, isobaric amino acids, such as oxidized Met-Phe or Gln-Lys, or isobaric amino acid combinations (presented above) that are indicated as “interchangeable” in the PEAKS results could be distinguished from their mass differences if ppm mass accuracy could be specified. This issue is clearly something that should be addressed in a future software release.

The PEAKS program clearly had difficulties assigning fragmentation that is different from the “typical tryptic” paradigm. When the non-tryptic peptide ended with a basic His, and thus, produced similar fragmentation to a tryptic peptide, *de novo* sequence determination was successful. For the protein C-terminal peptide, though all the necessary information was there to identify at least the five C-terminal residues, the software did not correctly interpret any sequence. Similarly, the chymotryptic-type peptide at *m/z* 729.9(2+) received a more confident score for the completely incorrect tryptic assignment than for the peptide reported in the search with no enzyme specificity, where the program correctly identified the five C-terminal residues (Supplements 3 and 5, Figure 9). This spectrum is an interesting example, because the incorrect assignment fits better in terms of mass accuracy of both precursor and abundant fragment ions (Supplement 5). However, this interpretation ignores the fact that the $\mathbf{y_1}$ for Arg and the immonium ion for His are both missing, which would both be unusual absences. At the same time, there is an immonium ion that indicates the presence of Phe, and two quite abundant fragments at *m/z* 127 and 155 that do not fit the tryptic sequence at all.

There were no obvious assignment differences when selecting the different activation methods, which is somewhat of concern because the fragmentation rules are different for ion trap and collision-cell CID. Although we did not test any other software on this dataset, from the literature comparisons [Chi et al., 2012] it seems that other programs would deliver similar results. All these points emphasize that in a manual analysis one can make use of other types of information that might not be usable by automated software. Nevertheless, using a *de novo* sequencing program for tryptic sequences should speed up the data interpretation process.

V. CONCLUSIONS

We believe that learning more about peptide fragmentation rules, and how to manually interpret and annotate data, is an important first step for everyone who is engaged in proteomic research. It would be especially important for all those who develop tools that enable researchers to process the huge amount of data acquired in today’s high-throughput experiments to have a thorough understanding of the data they analyze. Mathematical tools cannot work optimally unless their developers comprehend the physical, chemical, and biological complexity of the data.

At the same time, we challenge the users of these programs to develop a better understanding of the data, as well as of the tools used to decipher them. A closer collaboration between the two sides would be desirable to improve the reliability of data published.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The *de novo* sequencing presented here was part of a collaborative project with Dr. Gabor Rakhely, Department of Biotechnology, University of Szeged, Hungary. Bioinformatics Solutions Inc. provided us with the opportunity to test extensively the PEAKS *de novo* software. Funding from the NIH, NIGMS 8P41GM103481 to the Bio-Organic Biomedical Mass Spectrometry Resource at UCSF (Director: A.L. Burlingame) is gratefully acknowledged.

References

Alexander AJ, Thibault P, Boyd RK. Collision-induced dissociation of peptide ions. 2. Remote charge-site fragmentation in a tandem, hybrid mass spectrometer. Rapid Commun Mass Spectrom. 1989; 3:30–34.

Bakken V, Helgaker T, Uggerud E. Models of fragmentations induced by electron attachment to protonated peptides. Eur J Mass Spectrom (Chichester, Eng). 2004; 10:625–638.

Baldwin MA, Medzihradszky KF, Lock CM, Fisher B, Settineri TA, Burlingame AL. Matrix-assisted laser desorption/ionization coupled with quadrupole/orthogonal acceleration time-of-flight mass spectrometry for protein discovery, identification, and structural analysis. Anal Chem. 2001; 73:1707–1720. [PubMed: 11338583]

Ballard KD, Gaskell SJ. Sequential mass spectrometry applied to the study of the formation of “internal” fragment ions of protonated peptides. Int J Mass Spectrom Ion Processes. 1991; 111:173–189.

Bandeira N, Tsur D, Frank A, Pevzner PA. Protein identification by spectral networks analysis. Proc Natl Acad Sci U S A. 2007; 104:6140–6145. [PubMed: 17404225]

Barber M, Bordoli RS, Sedgwick RD, Tyler AN. Fast atom bombardment of solids as an ion source in mass spectrometry. Nature. 1981; 293:270–275.

Bern M, Goldberg D. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. J Comput Biol. 2006; 13:364–378. [PubMed: 16597246]

Bern M, Cai Y, Goldberg D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. Anal Chem. 2007; 79:1393–1400. [PubMed: 17243770]

Benninghoven A, Sichtermann WK. Detection, identification and structural investigation of biologically important compounds by secondary ion mass spectrometry. Anal Chem. 1978; 50:1180–1184. [PubMed: 677465]

Biemann K, Scoble HA. Characterization by tandem mass spectrometry of structural modifications in proteins. Science. 1987; 237:992–998. [PubMed: 3303336]

Biemann K. Appendix 5. Nomenclature for peptide fragment ions (positive ions). Methods Enzymol. 1990; 193:886–887. [PubMed: 2074849]

Bohlen CJ, Chesler AT, Sharif-Naeini R, Medzihradszky KF, Zhou S, King D, Sánchez EE, Burlingame AL, Basbaum AI, Julius D. A heteromeric Texas coral snake toxin targets acid-sensing ion channels to produce pain. Nature. 2011; 479:410–414. [PubMed: 22094702]

Bradshaw RA, Brickey WW, Walker KW. N-terminal processing: the methionine aminopeptidase and N alpha-acetyl transferase families. Trends Biochem Sci. 1998; 23:263–267. [PubMed: 9697417]

Branca RM, Bodó G, Bagyinka C, Prokai L. De novo sequencing of a 21-kDa cytochrome c4 from Thiocapsa roseopersicina by nanoelectrospray ionization ion-trap and Fourier-transform ion-cyclotron resonance mass spectrometry. J Mass Spectrom. 2007; 42:1569–82. [PubMed: 18085548]

Burlet O, Yang CY, Gaskell SJ. Influence of cysteine to cysteic acid oxidation on the collision-activated decomposition of protonated peptides: evidence for intraionic interactions. J Am Soc Mass Spectrom. 1992; 3:337–344. [PubMed: 24243044]

Burlet O, Orkiszewski RS, Ballard KD, Gaskell SJ. Charge promotion of low-energy fragmentations of peptide ions. Rapid Commun Mass Spectrom. 1992; 6:658–662. [PubMed: 1467550]

Chalkley RJ, Brinkworth CS, Burlingame AL. Side-chain fragmentation of alkylated cysteine residues in electron capture dissociation mass spectrometry. J Am Soc Mass Spectrom. 2006; 17:1271–1274. [PubMed: 16809046]

Chalkley RJ, Medzihradszky KF, Lynn AJ, Baker PR, Burlingame AL. Statistical analysis of peptide electron transfer dissociation fragmentation mass spectrometry. Anal Chem. 2010; 82:579–584. [PubMed: 20028093]

Chalkley RJ. When target-decoy false discovery rate estimations are inaccurate and how to spot instances. J Proteome Res. 2013; 12:1062–1064. [PubMed: 23298186]

Chi H, Sun RX, Yang B, Song CQ, Wang LH, Liu C, Fu Y, Yuan ZF, Wang HP, He SM, Dong MQ. pNovo: de novo peptide sequencing and identification using HCD spectra. J Proteome Res. 2010; 9:2713–2724. [PubMed: 20329752]

Chi H, Chen H, He K, Wu L, Yang B, Sun RX, Liu J, Zeng WF, Song CQ, He SM, Dong MQ. pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. J Proteome Res. 2013; 12:615–625. [PubMed: 23272783]

Chowdhury SM, Munske GR, Ronald RC, Bruce JE. Evaluation of low energy CID and ECD fragmentation behavior of mono-oxidized thio-ether bonds in peptides. J Am Soc Mass Spectrom. 2007; 18:493–501. [PubMed: 17126025]

Cotter RJ, Iltchenko S, Wang D, Gundry R. Tandem time-of-flight (TOF/TOF) mass spectrometry and proteomics. J Mass Spectrom Soc Jpn. 2005; 53:7–17. [PubMed: 20717501]

Cox KA, Gaskell SJ, Morris M, Whiting AJ. Role of the site of protonation in the low-energy decompositions of gas-phase peptide ions. J Am Soc Mass Spectrom. 1996; 7:522–531. [PubMed: 24203424]

Dancík V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing *via* tandem mass spectrometry. J Comput Biol. 1999; 6:327–342. [PubMed: 10582570]

Datta R, Bern M. Spectrum fusion: using multiple mass spectra for de novo peptide sequencing. J Comput Bio. 2009; 16:1169–82. [PubMed: 19645594]

Dongré AR, Jones JL, Somogyi Á, Wysocki VH. Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model. J Am Chem Soc. 1996; 118:8365–8374.

Falick AM, Hines WM, Medzihradszky KF, Baldwin MA, Gibson BW. Low-mass ions produced from peptides by high energy collision-induced dissociation in tandem mass spectrometry. J Am Soc Mass Spectrom. 1993; 4:882–893. [PubMed: 24227532]

Farrugia JM, O'Hair RAJ, Reid GE. Do all b2 ions have oxazolone structures? Multistage mass spectrometry and ab initio studies on protonated N-acyl amino acid methyl ester model systems. Int J Mass Spectrom. 2001; 210:71–87.

Frank A, Pevzner P. Pepnovo: De novo peptide sequencing *via* probabilistic network modeling. Anal Chem. 2005; 77:964–973. [PubMed: 15858974]

Gao X, Wu H, Lee KC, Liu H, Zhao Y, Cai Z, Jiang Y. Stable isotope N-phosphorylation labeling for peptide de novo sequencing and protein quantification based on organic phosphorus chemistry. Anal Chem. 2012; 84:10236–10244. [PubMed: 23134482]

Godugu B, Neta P, Simón-Manso Y, Stein SE. Effect of N-terminal glutamic acid and glutamine on fragmentation of peptide ions. J Am Soc Mass Spectrom. 2010; 21:1169–1176. [PubMed: 20413325]

Good DM, Wirtala M, McAlister GC, Coon JJ. Performance characteristics of electron transfer dissociation mass spectrometry. Mol Cell Proteomics. 2007; 6:1942–1951. [PubMed: 17673454]

Gu S, Pan S, Bradbury EM, Chen X. Precise peptide sequencing and protein quantification in the human proteome through in vivo lysine-specific mass tagging. J Am Soc Mass Spectrom. 2003; 14:1–7. [PubMed: 12504328]

Harrison AG. To b or not to b: The ongoing saga of peptide b ions. Mass Spectrom Rev. 2009; 28:640–654. [PubMed: 19338048]

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992; 89:10915–10919. [PubMed: 1438297]

Hines WM, Falick AM, Burlingame AL, Gibson BW. Pattern-based algorithm for peptide sequencing from tandem high energy collision-induced dissociation mass spectra. J Am Soc Mass Spectrom. 1992; 3:326–336. [PubMed: 24243043]

Huang Y, Triscari JM, Tseng GC, Pasa-Tolic L, Lipton MS, Smith RD, Wysocki VH. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. Anal Chem. 2005; 77:5800–5813. [PubMed: 16159109]

Hubler SL, Jue A, Keith J, McAlister GC, Craciun G, Coon JJ. Valence parity renders z(*)-type ions chemically distinct. J Am Chem Soc. 2008; 130:6388–6394. [PubMed: 18444621]

Hughes J, Smith TW, Kosterlitz HW, Fothergill LA, Morgan BA, Morris HR. Identification of two related pentapeptides from the brain with potent opiate agonist activity. Nature. 1975; 258:577–580. [PubMed: 1207728]

Hunt DF, Yates III JR, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. Proc Natl Acad Sci USA. 1986; 83:6233–6237. [PubMed: 3462691]

Hunt DF, Zhu NZ, Shabanowitz J. Oligopeptide sequence analysis by collision-activated dissociation of multiply-charged ions. Rapid Commun Mass Spectrom. 1989; 3:122–124. [PubMed: 2520232]

Jia C, Hui L, Cao W, Lietz CB, Jiang X, Chen R, Catherman AD, Thomas PM, Ge Y, Kelleher NL, Li L. High-definition de novo sequencing of crustacean hyperglycemic hormone (CHH)-family neuropeptides. Mol Cell Proteomics. 2012; 11:1951–1964. [PubMed: 23028060]

Johnson, RS. PhD Thesis. MIT; Cambridge, MA: 1988.

Johnson RS, Biemann K. The primary structure of thioredoxin from Chromatium vinosum determined by high-performance tandem mass spectrometry. Biochemistry. 1987; 26:1209–1214. [PubMed: 3567166]

Johnson RS, Biemann K. Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. Biomed Environ Mass Spectrom. 1989; 18:945–957. [PubMed: 2620156]

Johnson RS, Martin SA, Biemann K. Collision-induced fragmentation of (M+H)+ ions of peptides. Side chain specific sequence ions. Int J Mass Spectrom Ion Processes. 1988; 86:137–154.

Jonscher KR, Yates JR 3rd. The quadrupole ion trap mass spectrometer--a small solution to a big challenge. Anal Biochem. 1997; 244:1–15. [PubMed: 9025900]

Kapp EA, Schütz F, Reid GE, Eddes JS, Moritz RL, O'Hair RA, Speed TP, Simpson RJ. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. Anal Chem. 2003; 75:6251–6264. [PubMed: 14616009]

Kelstrup CD, Hekmat O, Francavilla C, Olsen JV. Pinpointing phosphorylation sites: Quantitative filtering and a novel site-specific x-ion fragment. J Proteome Res. 2011; 10:2937–2948. [PubMed: 21526838]

Kilpatrick LE, Neta P, Yang X, Simón-Manso Y, Liang Y, Stein SE. Formation of y + 10 and y + 11 ions in the collision-induced dissociation of peptide ions. J Am Soc Mass Spectrom. 2012; 23:655–663. [PubMed: 22161574]

Kim S, Gupta N, Bandeira N, Pevzner PA. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. Mol Cell Proteomics. 2009; 8:53–69. [PubMed: 18703573]

Krishnamurthy T, Szafraniec L, Hunt DF, Shabanowitz J, Yates JR 3rd, Hauer CR, Carmichael WW, Skulberg O, Codd GA, Missler S. Structural characterization of toxic cyclic peptides from blue-green algae by tandem mass spectrometry. Proc Natl Acad Sci U S A. 1989; 86:770–774. [PubMed: 2492662]

Lagerwerf FM, van de Weert M, Heerma W, Haverkamp J. Identification of oxidized methionine in peptides. Rapid Commun Mass Spectrom. 1996; 10:1905–1910. [PubMed: 9004526]

Lau KW, Hart SR, Lynch JA, Wong SC, Hubbard SJ, Gaskell SJ. Observations on the detection of b- and y-type ions in the collisionally activated decomposition spectra of protonated peptides. Rapid Commun Mass Spectrom. 2009; 23:1508–1514. [PubMed: 19370712]

Lee YJ, Lee YM. Formation of c1 fragment ions in collision-induced dissociation of glutamine-containing peptide ions: a tip for de novo sequencing. Rapid Commun Mass Spectrom. 2004; 18:2069–2076. [PubMed: 15378720]

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom. 2003; 17:2337–2342. [PubMed: 14558135]

Ma B, Johnson RS. *De Novo* Sequencing and Homology Searching Mol Cell Proteomics. 2012; 11:O111014902.

Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem. 1994; 66:4390–4399. [PubMed: 7847635]

Medzihradszky KF, Gibson BW, Kaur S, Yu ZH, Medzihradszky D, Burlingame AL, Bass NM. The primary structure of fatty-acid-binding protein from nurse shark liver. Structural and evolutionary relationship to the mammalian fatty-acid-binding protein family. Eur J Biochem. 1992; 203:327–339. [PubMed: 1735421]

Medzihradszky KF, Campbell JM, Baldwin MA, Falick AM, Juhasz P, Vestal ML, Burlingame AL. The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. Anal Chem. 2000; 72:552–558. [PubMed: 10695141]

Medzihradszky KF. Peptide sequence analysis. Methods Enzymol. 2005; 402:209–244. [PubMed: 16401511]

Medzihradszky KF, Trinidad JC. Unusual fragmentation of Pro-Ser/Thr-containing peptides detected in collision-induced dissociation spectra. J Am Soc Mass Spectrom. 2012; 23:602–607. [PubMed: 21952759]

Medzihradszky KF, Bohlen CJ. Partial de novo sequencing and unusual CID fragmentation of a 7 kDa, disulfide-bridged toxin. J Am Soc Mass Spectrom. 2012; 23:923–934. [PubMed: 22351294]

Muenchbach M, Quadroni M, Miotto G, James P. Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. Anal Chem. 2000; 72:4047–4057. [PubMed: 10994964]

Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics. 2006; 5:652–670. [PubMed: 16352522]

Paizs B, Suhai S. Fragmentation pathways of protonated peptides. Mass Spectrom Rev. 2005; 24:508–48. [PubMed: 15389847]

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20:3551–3567. [PubMed: 10612281]

Perlson E, Medzihradszky KF, Darula Z, Munno DW, Syed NI, Burlingame AL, Fainzilber M. Differential proteomics reveals multiple components in retrogradely transported axoplasm after nerve injury. Mol Cell Proteomics. 2004; 3:510–520. [PubMed: 14973157]

Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom. 1984; 11:601. [PubMed: 6525415]

Roth KD, Huang ZH, Sadagopan N, Watson JT. Charge derivatization of peptides for analysis by mass spectrometry. Mass Spectrom Rev. 1998; 17:255–274. [PubMed: 10224676]

Samgina TY, Artemenko KA, Gorshkov VA, Ogourtsov SV, Zubarev RA, Lebedev AT. De novo sequencing of peptides secreted by the skin glands of the Caucasian Green Frog Rana ridibunda. Rapid Commun Mass Spectrom. 2008; 22:3517–3525. [PubMed: 18855342]

Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA. Proteomics-grade de novo sequencing approach. J Proteome Res. 2005; 4:2348–2354. [PubMed: 16335984]

Schlosser A, Lehmann WD. Patchwork peptide sequencing: extraction of sequence information from accurate mass data of peptide tandem mass spectra recorded at high resolution. Proteomics. 2002; 2:524–533. [PubMed: 11987126]

Schnölzer M, Jedrzejewski P, Lehmann WD. Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry. Electrophoresis. 1996; 17:945–953. [PubMed: 8783021]

Scoble HA, Martin SA, Biemann K. Peptide sequencing by magnetic deflection tandem mass spectrometry. Biochem J. 1987; 245:621–622. [PubMed: 3663180]

Seidler J, Zinn N, Boehm ME, Lehmann WD. De novo sequencing of peptides by MS/MS. Proteomics. 2010; 10:634–649. [PubMed: 19953542]

Shevchenko A, Wilm M, Mann M. Peptide sequencing by mass spectrometry for homology searches and cloning of genes. J Protein Chem. 1997; 16:481–490. [PubMed: 9246632]

Shevchenko A, Chernushevich I, Wilm M, Mann M. De novo peptide sequencing by nanoelectrospray tandem mass spectrometry using triple quadrupole and quadrupole/time-of-flight instruments. Methods Mol Biol. 2000; 146:1–16. [PubMed: 10948493]

Simón-Manso Y, Neta P, Yang X, Stein SE. Loss of 45 Da from a2 ions and preferential loss of 48 Da from a2 ions containing methionine in peptide ion tandem mass spectra. J Am Soc Mass Spectrom. 2011; 22:280–289. [PubMed: 21472587]

Spengler B. De novo sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry. J Am Soc Mass Spectrom. 2004; 15:703–714. [PubMed: 15121200]

Standing KG. Peptide and protein de novo sequencing by mass spectrometry. Curr Opin Struct Biol. 2003; 13:595–601. [PubMed: 14568614]

Stevens SM Jr, Prokai-Tatrai K, Prokai L. Factors that contribute to the misidentification of tyrosine nitration by shotgun proteomics. Mol Cell Proteomics. 2008; 7:2442–2451. [PubMed: 18708664]

Sun S, Yu C, Qiao Y, Lin Y, Dong G, Liu C, Zhang J, Zhang Z, Cai J, Zhang H, Bu D. Deriving the probabilities of water loss and ammonia loss for amino acids from tandem mass spectra. J Proteome Res. 2008; 7:202–208. [PubMed: 18092745]

Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci USA. 2004; 101:9528–9533. [PubMed: 15210983]

Sztáray J, Memboeuf A, Drahos L, Vékey K. Leucine enkephalin--a mass spectrometry standard. Mass Spectrom Rev. 2011; 30:298–320. [PubMed: 20669325]

Tang XJ, Thibault P, Boyd RK. Fragmentation reactions of multiply-protonated peptides and implications for sequencing by tandem mass spectrometry with low-energy collision-induced dissociation. Anal Chem. 1993; 65:2824–2834. [PubMed: 7504416]

Taylor JA, Johnson RS. Sequence database searches *via* de novo peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom. 1997; 11:1067–1075. [PubMed: 9204580]

Taylor JA, Johnson RS. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. Anal Chem. 2001; 73:2594–2604. [PubMed: 11403305]

Thorne GC, Gaskell SJ. Elucidation of some fragmentations of small peptides using sequential mass spectrometry on a hybrid instrument. Rapid Commun Mass Spectrom. 1989; 3:217–221. [PubMed: 2520240]

Thorne GC, Ballard KD, Gaskell SJ. Metastable decomposition of peptide [M + H]+ ions *via* rearrangement involving loss of the C-terminal amino acid residue. J Am Soc Mass Spectrom. 1990; 1:249–257.

Yalcin T, Khouw C, Csizmadia IG, Peterson MR, Harrison AG. Why Are B Ions Stable Species in Peptide Spectra? J Am Soc Mass Spectrom. 1995; 6:1165–1174. [PubMed: 24214067]

Waridel P, Frank A, Thomas H, Surendranath V, Sunyaev S, Pevzner P, Shevchenko A. Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated de novo sequencing. Proteomics. 2007; 7:2318–2329. [PubMed: 17623296]

Wattenberg A, Organ AJ, Schneider K, Tyldesley R, Bordoli R, Bateman RH. Sequence dependent fragmentation of peptides generated by MALDI quadrupole time-of-flight (MALDI Q-TOF) mass spectrometry and its implications for protein identification. J Am Soc Mass Spectrom. 2002; 13:772–783. [PubMed: 12148802]

Wen DX, Livingston BD, Medzihradszky KF, Kelm S, Burlingame AL, Paulson JC. Primary structure of Gal beta 1,3(4)GlcNAc alpha 2,3-sialyltransferase determined by mass spectrometry sequence analysis and molecular cloning. Evidence for a protein motif in the sialyltransferase gene family. J Biol Chem. 1992; 267:21011–21019. [PubMed: 1400416]

Wielsch N, Thomas H, Surendranath V, Waridel P, Frank A, Pevzner P, Shevchenko A. Rapid validation of protein identifications with the borderline statistical confidence *via* de novo sequencing and MS BLAST searches. J Proteome Res. 2006; 5:2448–2456. [PubMed: 16944958]

Wysocki VH, Tsaprailis G, Smith LL, Breci LA. Mobile and localized protons: a framework for understanding peptide dissociation. J Mass Spectrom. 2000; 35:1399–1406. [PubMed: 11180630]

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics. 2012; 11:M111010587.

Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides. Anal Chem. 2004; 76:3908–3922. [PubMed: 15253624]

Zhang Z. Prediction of electron-transfer/capture dissociation spectra of peptides. Anal Chem. 2010; 82:1990–2005. [PubMed: 20148580]

Zhang Z. Prediction of collision-induced-dissociation spectra of peptides with post-translational or process-induced modifications. Anal Chem. 2011; 83:8642–8651. [PubMed: 21995278]

Zubarev RA, Horn DM, Fridriksson EK, Kelleher NL, Kruger NA, Lewis MA, Carpenter BK, McLafferty FW. Electron capture dissociation for structural characterization of multiply-charged protein cations. Anal Chem. 2000; 72:563–573. [PubMed: 10695143]

Zubarev RA, Zubarev AR, Savitski MM. Electron capture/transfer versus collisionally activated/induced dissociations: solo or duet? J Am Soc Mass Spectrom. 2008; 19:753–761. [PubMed: 18499036]

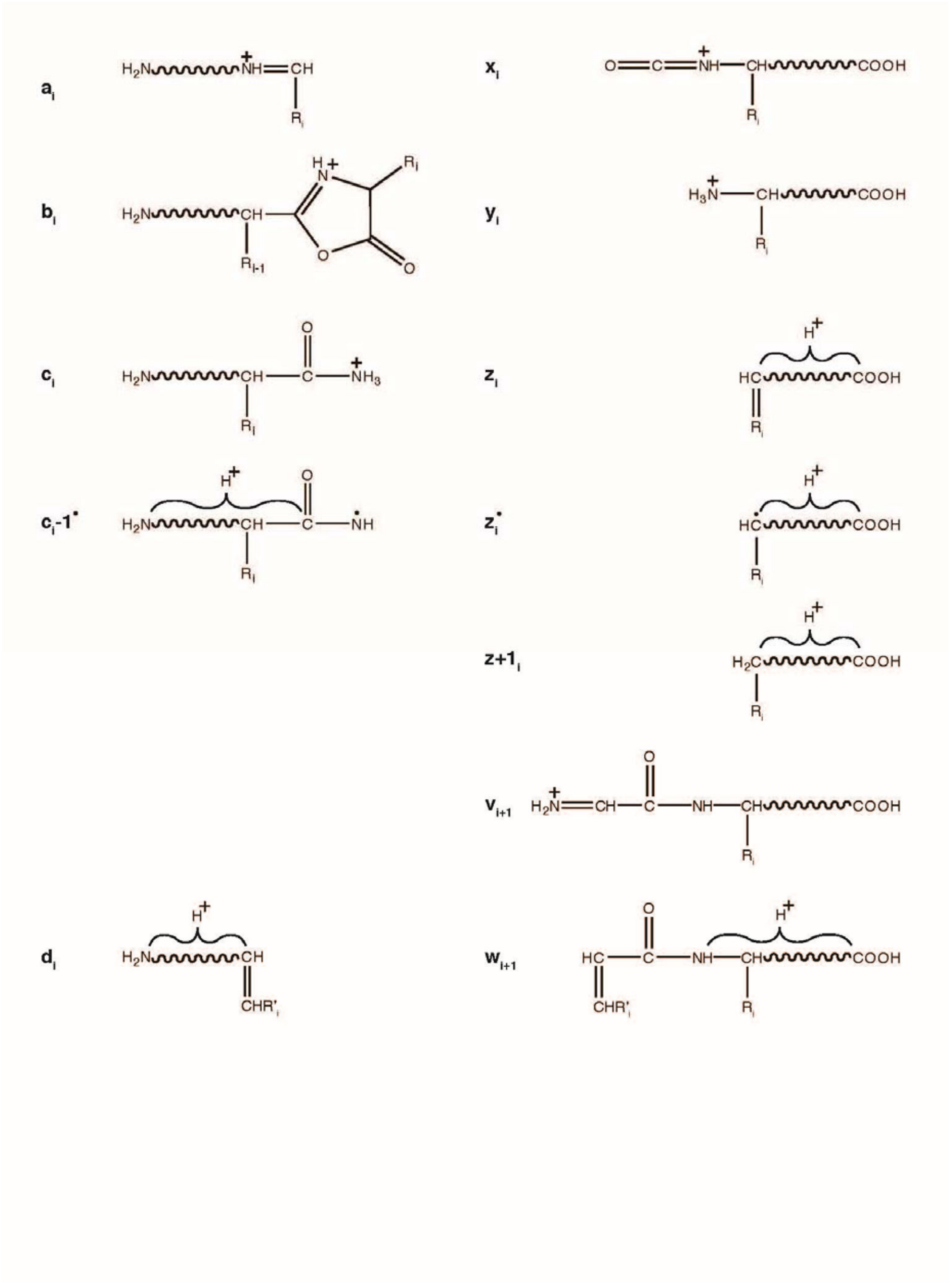

Figure 1.

Peptide fragment ions [Biemann, 1990]. The occurrence and mass calculations of these fragments are presented in Tables 2 and 3.

1 MALFLPGWPE VPAEYPSDSV GMHPPYDTPA YIFTRKRAPS RPLRYIPQTA 50

51 TELYGPVYGH ESVRPEDSDL TRQHDGEPIG ERIRITGRVI DEDGRGVPNA 100

101 LLEIWQANAA GRYIHK**LDQH LAPLDPNFSG AGR**TVTGSDG SYSFITIIPG 150

151 AYPVVGLHNV WRPRHIHISL FGPSFLSRLV TQLYFEGDPL LRYDSIYNAA 200

201 PDFSKRGMVA SLDLEATQSE WGLTYR**FDIV LR**GRNGNYFE ETHAH 245

Figure 2.

Sequence of the protocatechuate 3,4-dioxygenase type II beta subunit PcaH-II [*Agrobacterium tumefaciens*]. Two tryptic peptides (underlined) were identified in database searches of *both* of the digests.

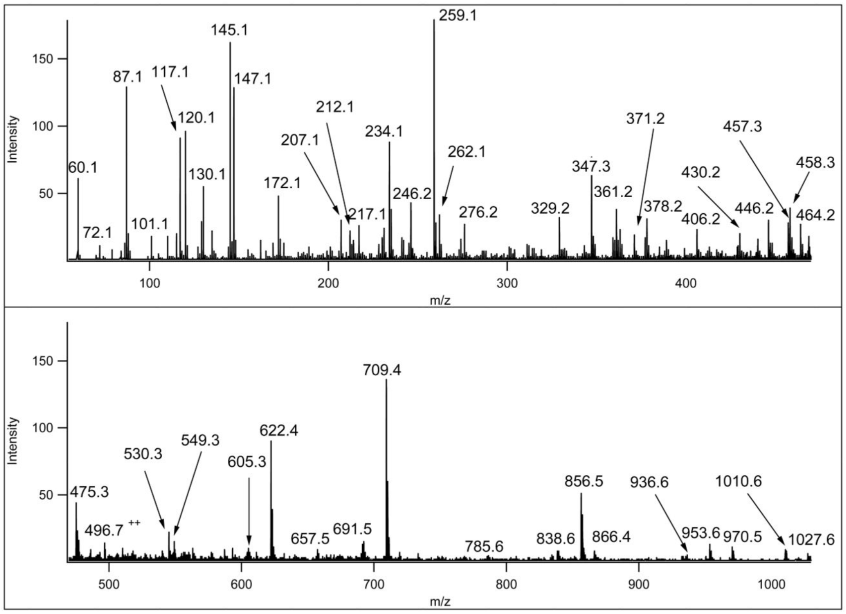

Figure 3.

Low-energy CID spectrum of *m/z* 557.8(2+). The corresponding sequence was determined from this spectrum as Ser-Gly-Asn-Phe-Ser-Phe-Gln-Thr-Val-Lys.

1 MRIEAPMTIT PSQTVGPFYA YCLTPDDYQT IPPIFGRNLA TDDAVDSAFQ 50

51 FRGRLIDGDD HAIPDGMIEL WQPDGNGNFV GAQINPRKSS FTGFGRTHCN 100

101 E**SGSFTFHTV K**PGRVPTSAG ILQAPHVALS IFGKGLNRRL YTRIYFADEV 150

151 SNDEDPILAL LSSDERATLI AEKIDDAAFH ITIRLQGQRE TVFFEV 196

Figure 4.

Sequence of protocatechuate 3,4-dioxygenase type II alpha subunit PcaG-II [*Agrobacterium tumefaciens*], NCBI# 11037227. The underlined sequence is homologous to the tryptic peptide sequenced from the spectrum in Figure 3, with S->N, T->S and H->Q substitutions in positions 3, 5, and 7, respectively.

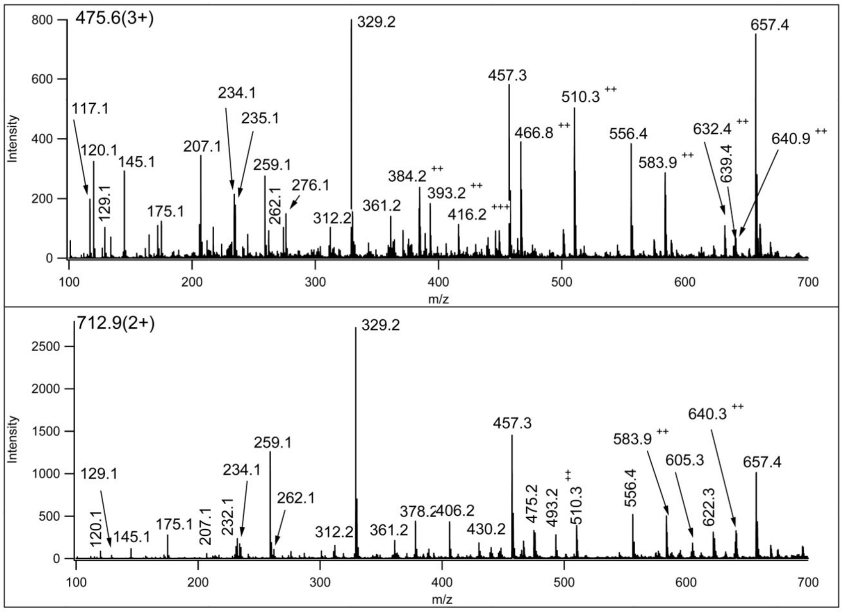

Figure 5.

Low-energy CID spectra of the tryptic peptide SGNFSFQTVKPGR, from the triply-charged precursor (upper panel) and the doubly-charged precursor (lower panel).

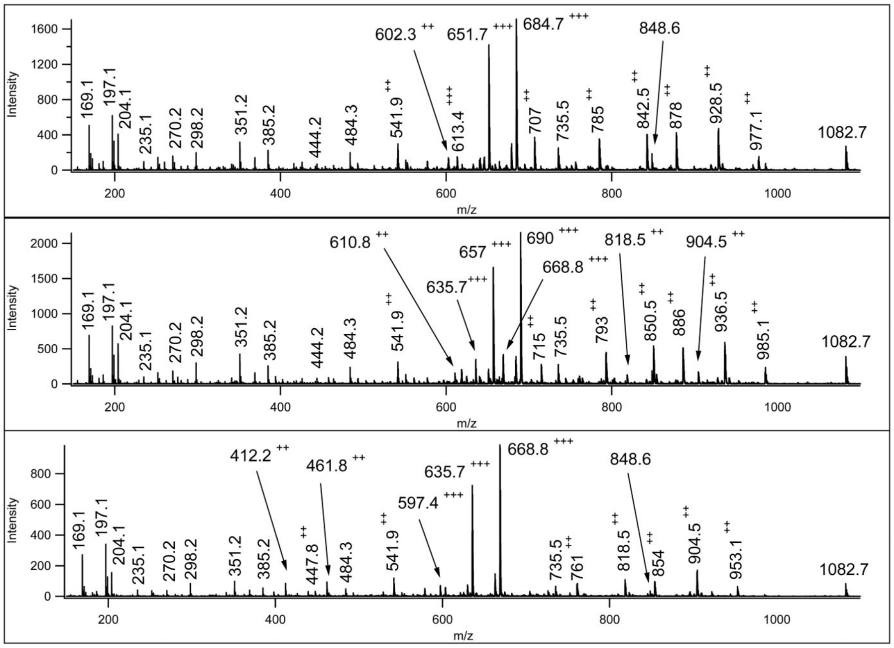

Figure 6.

Low-energy CID spectra of *m/z* 684.7(3+)(top), 690.0(3+) (middle) and 703.7(3+) (bottom). These spectra represent tryptic peptide, VPTADGVMQAPHLALSIFGK unmodified, with a Met-sulfoxide, and with a carbamidomethyl Met, respectively. The spectra are presented in reverse elution order.

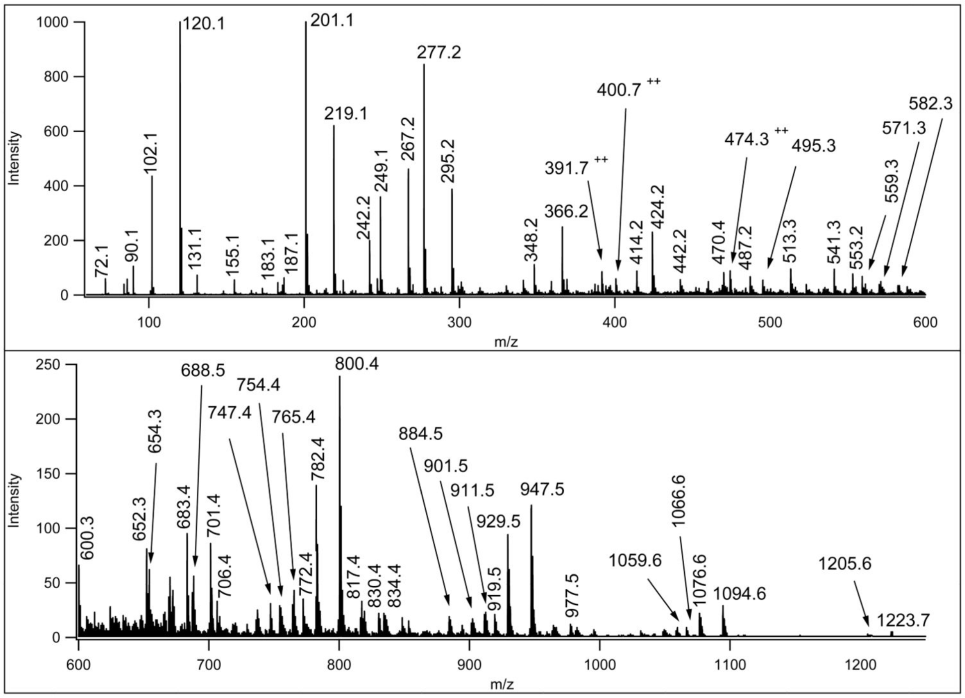

Figure 7.

Low-energy CID spectrum of *m/z* 656.8(2+), representing the C-terminal tryptic peptide of the protein.

1 XXXXLPTT**I**T PSQTVGPFYA YXXXXXXXXX XXXXXXXXXX XXXXXXXXXX 50

51 XXXXXXXXXX XXXXXXXXXX XXXXXXXXXX XXXXXXXXXX XXXXXXXXXX 100

101 XSGNFSFQTV KPGRVPTADG VM**Q**APHLA**L**S **I**FGKG**L**NRR**L** YTRXXXXXXX 150

151 XXXXXXXXXX XXXXXXVT**LI** ATSESPAAYR XXXXXXXXXX TVFFEA 196

1 MTTKTPLTIT PSQTVGPFYA YCLTPEDYGT LPPLFGAQLA TEDAEGERIT 50

51 IQGTITDGEG AMVPDALIEI WQPDGQGR**FA GAHPELR**NSA FKGFGRRHCD 100

101 KSGNFSFQTV KPGRVPTADG VMQAPHIALS IFGKGLNRRL YTRIYFADEA 150

151 SNAEDPVLSM LSEDERVTLI ATSESPAAYR LDIRLQGDGE TVFFEA 196

Figure 8.

The manually deciphered sequence of sulfocatechol 3,4-dioxygenase alpha-subunit of *Novosphingobium resinovorum* (*Sphingomonas subarctica*) (NCBI # 56787886) is shown in the upper panel. The genomic sequence was determined later utilizing this information, as presented in the lower panel. The correctly determined sequence of a tryptic peptide that did not show sufficient similarity to the “template” and thus its sequence position could not be predicted is printed in bold.

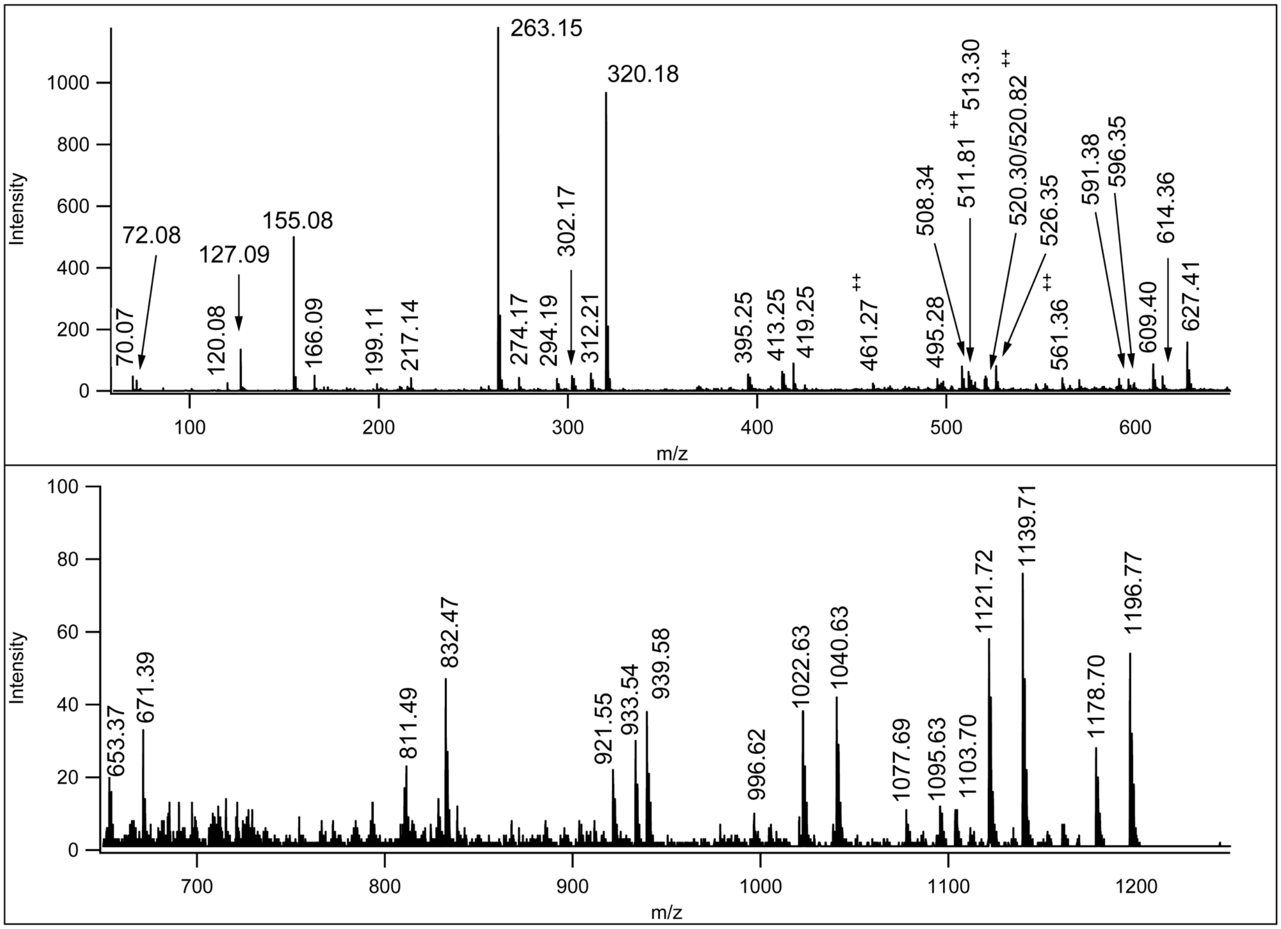

Figure 9.

Low-energy CID spectrum of a chymotryptic-type peptide; precursor at *m/z* 729.9(2+). A fragment assignment comparison for the different sequence solutions proposed by the PEAKS software (Supplement 3) from manual sequencing (Table 5), and for the correct sequence (Figure 8) is presented in Supplement 5.

Table 1

The 20 'standard' amino acids (directly encoded by the universal genetic code)

Name			Elemental composition	Residue mass
Full	**3 letter code**	**1 letter code**	**Neutral molecule**	**(Monoisotopic)**
Alanine	Ala	A	$C_3H_7NO_2$	71.0372
Arginine	Arg	R	$C_6H_{14}N_4O_2$	156.0981
Asparagine	Asn	N	$C_4H_8N_2O_3$	114.0429
Aspartic acid	Asp	D	$C_4H_8NO_4$	115.0269
Cysteine	Cys	C	$C_3H_7NO_2S$	103.0092
Glutamic acid	Glu	E	$C_5H_9NO_4$	129.0426
Glutamine	Gln	Q	$C_5H_{10}N_2O_3$	128.0586
Glycine	Gly	G	$C_2H_5NO_2$	57.0215
Histidine	His	H	$C_6H_9N_3O_2$	137.0589
Isoleucine	Ile	I	$C_6H_{13}NO_2$	113.0841
Leucine	Leu	L	$C_6H_{13}NO_2$	113.0841
Lysine	Lys	K	$C_6H_{14}N_2O_2$	128.0949
Methionine	Met	M	$C_5H_{11}NO_2S$	131.0405
Phenylalanine	Phe	F	$C_9H_{11}NO_2$	147.0684
Proline	Pro	P	$C_5H_9NO_2$	97.0528
Serine	Ser	S	$C_3H_7NO_3$	87.0320
Threonine	Thr	T	$C_4H_9NO_3$	101.0477
Tryptophan	Trp	W	$C_{11}H_{12}N_2O_2$	186.0793
Tyrosine	Tyr	Y	$C_9H_{11}NO_3$	163.0633
Valine	Val	V	$C_5H_{11}NO_2$	99.0684

Table 2

Fragment ions observed in different MS/MS experiments

	hiE CID	loE CID quad	loE CID trap	PSD	ECD/ETD
Immonium ions[1]	x	(x)		x	
side-chain losses[2]	x	(x)	(x)	(x)	x
a	x	x	(x)	x	
$a-NH_3$[3]	x	x	(x)	x	
$a-H_2O$[4]	x	x	(x)	x	
b	x	x	x	x	((x))
$b-NH_3$[3]	x	x	x	x	
$b-H_2O$[4]	x	x	x	x	
$b+H_2O$[5]	x	x	x	x	
$c-1^{\cdot}$[6]					x
c		((x))[7]			x
d	x				((x))[8]
v	x				
w	x				(x)[9]
x	x				
y	x	x	x	x	x
$y-NH_3$[3]	x	x	x	x	
$y-H_2O$[4]	x	x	x	x	
z (y−17)[3,10]	x				
$z+H/z^{\cdot}$[6,10]	x				x
z+2H/z+1[10]					x
internal b-type[11]	x	x		x	
internal a-type[11]	x	x		x	

(x) such fragments might be detected, but not reliably.

((x)) such fragments are detected rarely, sometimes only in special cases.

[1] this row also refers to related ions, listed in Table 4.

[2] neutral losses from the precursor ion, when the entire amino acid side-chains or parts of them are cleaved.

[3] as a rule, the fragment must contain Asn, Gln, Lys, or Arg.

[4] as a rule, the fragment must contain Asp, Glu, Ser, or Thr.

[5] usually, the ultimate residue may be eliminated, as discussed below.

[6] radical (odd-electron) ion.

[7] special cases of **c** fragment detection will be discussed below.

[8] in ECD [KF Medzihradszky, unpublished data]

[9] alkyl-Cys residues produce these fragments regularly in ECD and ETD; the other amino acids only in “hot” ECD.

[10] the Biemann-nomenclature defined **y**−17, as the **z**-fragment. However, recently the radical $\mathbf{z}^{\bullet}$ ion (**y**−16) is usually called **z** ion, while its counterpart formed *via* proton migration (**y**−15) is assigned as **z**+1 fragment.

[11] internal ions undergo neutral losses similarly to sequence ions, and they follow the same rules.

Table 3

Rules for the calculation of fragment ion masses

Fragment	Mass calculation	
	using residue weights[1]	from other fragments
Immonium ion	residue weight−26.9871	n.a.
$\mathbf{a_i}$	Σresidue weights −26.9871	$b_i-27.9949$
$\mathbf{b_i}$	Σresidue weights + 1.0078	$MH^{+}+1-y_{n-i}$
$\mathbf{c_i}$	Σresidue weights + 18.0344	$b_i+17.0265$
$\mathbf{d_i}$	Σprevious residue weights + 44.0500	$a_i-(R_i-15.0235)$
for Ile, Thr, Val	Σprevious residue weights + 58.0657	$a_i-28.0312$
2nd option for Ile	Σprevious residue weights + 72.0813	a_i −14.0156
2nd option for Thr	Σprevious residue weights + 60.0450	a_i −14.0156
$\mathbf{v_i}$	Σprevious residue weights + 74.0242	$x_{i-1}+29.0265$
$\mathbf{w_i}$	Σprevious residue weights + 73.0290	$x_{i-1}+28.0312$
for Ile, Thr, Val	Σprevious residue weights −26.03947	$x_{i-1}+42.0469$
2nd option for Ile	Σresidue weights −12.0238	x_{i-1} +56.0626
2nd option for Thr	Σresidue weights −28.0187	$x_{i-1}+44.0262$
$\mathbf{x_i}$	Σresidue weights + 44.9977	$y_i+25.9793$
$\mathbf{y_i}$	Σresidue weights + 19.0184	$MH^{+}+1-b_{n-i}$
$\mathbf{z_i}$	Σresidue weights + 2.0156	$y_i-17.0242$
Internal fragments		
b-type	Σresidue weights + 1.0078	
a-type	Σresidue weights − 26.9871	

[1] The mass of an electron was not subtracted

Table 4

Immonium and Related Ions Characteristic of the 20 Standard Amino Acids

Amino Acid	Immonium and related ion(s) masses		Comments
Ala	44		
Arg	129	59, 70, 73, 87, 100, 112	129, 73 usually weak
Asn	87	70	87 often weak, 70 weak
Asp	88		Usually weak
Cys	76		Usually weak
Gly	30		
Gln	101	84, 129	129 weak
Glu	102		Often weak if C-terminal
His	110	82, 121,123, 138, 166	110 very strong 82, 121, 123, 138 weak
Ile/Leu	86		
Lys	101	84, 112, 129	101 can be weak
Met	104	61	104 often weak
Phe	120	91	120 strong, 91 weak
Pro	70		Strong
Ser	60		
Thr	74		
Trp	159	130, 170, 171	Strong
Tyr	136	91, 107	136 strong, 107, 91 weak
Val	72		Fairly strong

Reprinted with kind permission from Springer Science+Business Media: Table 1. from "Low-mass ions produced from peptides by high energy collision-induced dissociation in tandem mass spectrometry." J Am Soc Mass Spectrom (1993) 4:882–89. Falick AM, Hines WM, Medzihradszky KF, Baldwin MA, Gibson BW.; © 1993 American Society for Mass spectrometry

Table 5

Manually interpreted CID data listed in elution order[1]

Precursor	Sequence	Comment[2]
308.19(2+)	GLNRR	alpha [135–139]
322.87(3+)	RAPTRPLR	beta [37–44]
354.71(2+)	RLYTR	alpha [139–143]
405.75(2+)	APTRPLR	beta [38–44]
427.21(3+)	VPTADGVM(CH_2CONH_2)QAPH	alpha [115–126]
499.26(2+)	FAGAHPELR	???
329.7(2+)	ILVTGR	mightbe beta [83–88]
527.8(2+)	(NH_2CO-CH_2)-FAGAHPELR	definitely side-reaction!
619.79(2+)	VPTADGVM(O)QAPH	alpha [115–126]
480.9(3+)	SGNYSFQTVKPGR	alpha isoform?
475.58(3+)	SGNFSFQTVKPGR	alpha [102–114].
552.95(3+)	YDTIYNTAPDLSKR	beta [193–206].
557.78(2+)	SGNFSFQTVK	alpha [102–111]
739.89(2+)	VTILATSES*PAAY*R	alpha [167–180]
528.03(4+)	VPTADGVM(CH_2CONH_2)QAPHLALSIFGK	alpha [115–134]
656.0(3+)	{SV}*EA*HPAYLTPDYVFTR	beta [19–35]
690.03(3+)	VPTADGVM(O)QAPHLALSIFGK	alpha [115–134]
684.70(3+)	VPTADGVMQAPHLALSIFGK	alpha [115–134]
729.90(2+)	*LPT*TIT*PSQ*TVGPF	alpha [5–18]
656.81(2+)	……TVFFEA	alpha C-terminus
928.48(2+)	*LPT*TIT*PSQ*TVGPFYAY	alpha [5–21]

[1] Complete peaklist in mgf format, and raw data were submitted as Supplements 1 and 2. All spectra were also included as Figures in the article or in Supplement 4.

[2] Sequence positions are given based on similarity to the homologous sequences presented in Figures 2 and 4.

[*] Dots indicate regions where the sequence could not be deciphered.

[*] Underlined amino acids were assigned based on homology, because CID data are identical for isomeric Leu and Ile. Isobaric Gln and Lys can be distinguished based on accurate mass measurements, but sometimes there were no sufficiently small fragments detected to allow differentiation.

[*] Amino acids in *italics* were included also based on homology considerations. Mass-wise, these residue combinations are perfect matches, but there was no proof for confident sequence assignment.

[*] When isomeric Ile/Leu cannot be assigned based on the homology, Leu is listed.

[*] {residues} indicate that their order could not be assigned

[*] Sequence in bold indicates the sequence identified using Mascot as present in the database.