Published in final edited form as:

J Am Soc Mass Spectrom. 2012 May ; 23(5): 923–934. doi:10.1007/s13361-012-0350-x.

Partial *de novo* sequencing and unusual CID fragmentation of a 7 kDa, disulfide-bridged toxin

Katalin F. Medzihradsky^{*,1} and Christopher J. Bohlen²

¹Mass Spectrometry Facility, Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94158-2517 USA

²Department of Physiology, University of California San Francisco, San Francisco, CA 94158-2517 USA

Abstract

A 7 kDa toxin isolated from the venom of the Texas coral snake (*Micrurus tener tener*) was subjected to collision-induced dissociation (CID) and electron-transfer dissociation (ETD) analyses both before and after reduction at low pH. Manual and automated approaches to *de novo* sequencing are compared in detail. Manual *de novo* sequencing utilizing the combination of high accuracy CID and ETD data and an acid-related cleavage yielded the N-terminal half of the sequence from the reduced species. The intact polypeptide, containing 3 disulfide bridges produced a series of unusual fragments in ion trap CID experiments: abundant internal amino acid losses were detected, and also one of the disulfide-linkage positions could be determined from fragments formed by the cleavage of two bonds. In addition, internal and c-type fragments were also observed.

Keywords

CID; ETD; *de novo* sequencing; fragmentation; disulfide-bridge; high mass accuracy; peak-picking

Introduction

Mass spectrometry has become a significant player in peptide sequence determination approximately a quarter of a century ago [1, 2]. While the function of this technique has dramatically shifted towards high-throughput peptide identification utilizing the ever-expanding protein and genomic databases with automated search engines, mass spectrometry's role in *de novo* sequencing remains important. Quite a few automated search engines can compensate for a series of "problems" in identifying protein fragments such as non-specific cleavages, misidentification of the monoisotopic mass, amino acid substitutions, and unexpected covalent modifications, but the combination thereof may still prevent correct identification of a peptide. In addition, even fully-sequenced genomes yield

The correspondence should be addressed to: Katalin F. Medzihradsky, University of California San Francisco, 600 16th Street Genentech Hall, suite N472A, Box 2240, San Francisco, CA 94158-2517, Phone number: (415)-476-5160, Fax number: (415)-502-1655, folk1@cgl.ucsf.edu.

incomplete protein databases because translations are usually predicted based on similarities with other species; species-specific sequences may not overlap sufficiently with known proteins for accurate prediction, and thus may be overlooked. Last but not least, we have very limited genomic (and therefore proteomic) information for a vast number of species.

Neuroscience frequently utilizes toxins as biochemical and pharmacological tools that can be used to manipulate key physiological receptors [3]. These toxins can be isolated from a wide variety of species, including sea snails, insects, fishes, centipedes, spiders, scorpions, lizards, and snakes. They represent a bewildering variety of structures, and in most instances genomic information is lacking. Furthermore, unusual covalent modifications frequently occur. Thus, *de novo* sequencing is typically required for the characterization of these interesting and valuable polypeptides.

Most *de novo* sequencing programs were developed for (and work most reliably for) tryptic peptides [4-6]. Presently, the PEAKS *de novo* sequencing software is used most frequently. Most available software cannot interpret/identify “atypical” fragmentation patterns very successfully. A new approach, composition-based sequencing (CBS), was developed to utilize the high mass accuracy of the precursor ions and some fragment masses for *de novo* sequencing. However, since the paradigm involves the determination of the amino acid composition of the entire peptide from a few *per se* unambiguous fragment ions, the method can only be used for relatively short sequences [7].

Collisional activation of non-tryptic sequences may represent a significant challenge because of the unpredictable fragmentation pattern (most tryptic peptides feature a single basic residue at the C-terminus, which leads to preferential charge retention and abundant y ions). Thus, the CBS method could be more successful for such peptides. Another potential solution would be to apply two complementary activation methods: electron-capture or electron-transfer dissociation (ECD & ETD) and collision-induced dissociation (CID) for the analysis of each molecule. While the primary cleavage site in CID is the peptide bond, yielding **b** and **y** fragments, the bond between the amino group and the α carbon will fragment in ECD/ETD experiments producing mostly **z**. and **c** ions [8]. The combination of CID and ECD data has been reported for improved peptide identification [9] as well as for *de novo* sequencing [10]. Identification of complementary fragment ions in corresponding CID and ECD/ETD spectra provides a strong framework for sequence determination, and high-accuracy mass measurements make these identifications more reliable. The mass accuracy afforded may be sufficient to apply the CBS paradigm for a single ion series and obtain unambiguous sequence assignments. Indeed, that is what we demonstrate in this paper.

In this study we describe a polypeptide, MitTx α purified from the venom of the Texas coral snake (*Micrurus tener tener*), that proved to be part of a heteromeric complex functioning as a selective agonist for acid-sensing ion channels [11]. MALDI mass measurement indicated a single, 7 kDa component in the fraction of interest. Thus, we set out to use different MS/MS techniques with high accuracy mass measurement to obtain sufficient sequence information either to identify the polypeptide (if it has been included in the public protein databases) or to provide suitable information for successful cloning experiments.

We will present how the combination of high accuracy ion trap CID and ETD data permitted sequencing of the N-terminal half of the molecule. We will show that *de novo* sequencing is still an area where human intervention is important for success. We will also show some unusual CID fragmentation for the intact polypeptide.

Experimental

Reduction/alkylation attempts—To aliquots of toxin solution in 25 mM ammonium bicarbonate buffer, pH~8, 15 nmoles of DTT was added, and the mixture was incubated at 56°C for 30 min. Then 32 nmoles of iodoacetamide or iodoacetic acid was added, and the mixture was incubated at room temperature for 30 min.

Successful reduction of the polypeptide—Toxin solution was incubated with 125 nmoles of TCEP in 0.1% formic acid, at 37°C, for 24h.

An aliquot of the mixture was analyzed on a NanoACQUITY (Waters)-LTQ-Orbitrap Velos LC/MS system in LC/MS mode in order to assess the success of reduction. For all LC/MS experiments solvents A and B were 0.1% formic acid in water or acetonitrile, respectively. Gradient elution from 5% to 40 % B in 35 min was used to fractionate the components.

Peptide fragmentation analyses

MS/MS and MS³ analyses of the intact polypeptide were performed with an infused solution (~400 nL/min) of the polypeptide, using an LTQ-Orbitrap XL with ETD or CID activation in the linear trap. Fragments were measured in the Orbitrap. The isolation window was 10 Th for both MS2 and MS3 experiments, and 2 microscans were acquired. ETD activation time was 30 msec. The CID activation energy was set at 35% and 40%, for the ETD->CID and the CID experiments, respectively. AGC targets were set at 10⁴ for the linear trap and 2x10⁵ for the Orbitrap, which was operated at a resolution of 30000.

MS/MS analysis of the short peptide was performed using an LTQ-Orbitrap Velos mass spectrometer – an aliquot of the TCEP-containing mixture was diluted with an equal amount of acetonitrile, and infused at a flow rate of ~400 nL/min. CID experiments on both the doubly and singly charged precursor ions were performed in the linear trap at a 35% normalized collision energy, while the fragments were measured in the Orbitrap. Higher-energy collision-dissociation (HCD) collision energy was also set to 35%. The precursor ion window was 2 Th, resolution was set to 30000, and the AGC setting was as in the LC/MS/MS experiment below.

MS/MS analysis of the long peptide was performed using an LTQ-Orbitrap Velos mass spectrometer – the analysis was performed in LC/MS mode. Precursor as well as MS/MS fragment masses were measured in the Orbitrap, at a resolution of 30000 and 15000, respectively. The isolation width was set to 7 Th, the minimum peak intensity to trigger CID or ETD analysis was set to 2000. The AGC target for MS/MS experiment was set to 10⁴ and 8x10⁴ for the linear trap and the Orbitrap, respectively. Single microscans were acquired for CID at 35% normalized collision energy and for ETD at 37.5 msec activation time. The fluoranthene AGC target was 10⁶.

Data processing

Scans representing the same MS/MS experiments were merged using the Xcalibur (v2.1.0 build 1139) software.

Peak lists were generated using Xtract as well as manually. Xtract is a feature of Xcalibur, which converts the ions observed into a list of singly charged masses. The resolution afforded and the signal-to-noise ratio of the ions to be deconvoluted can be specified. These parameters are listed with the peak lists presented.

De novo sequencing was performed manually as presented below.

MS-product and MS-comp of Protein Prospector v5.8.0 (www.prospector.ucsf.edu) was used to display instrument-specific fragmentation and potential amino acid compositions, respectively. PEAKS Studio 5.3 build 20110719 was also tested for *de novo* sequencing [6].

Spectra were also evaluated using in-house software FAVA, that performs peak-picking using the sequence determined [12].

Results and discussion

A toxin-containing fraction, named MitTx α from the venom of the Texas coral snake (*Micrurus tener tener*) [11] was subjected to mass spectrometry analysis *via* infusion in an LTQ-Orbitrap. The polypeptide's monoisotopic mass was determined from ions at m/z 888.3915(8+), 1015.1630(7+), 1184.1883(6+), and 1420.8202(5+) as 7100.0830 (MH⁺). The toxin molecule seemed to be both sufficiently small and sufficiently charged for ETD analysis. Unfortunately, the ETD analysis of m/z 1015(7+) (the 8+ ion was not abundant enough) yielded no information besides the charge reduced ions (Data not shown).

Considering that sometimes fragmentation occurs, but the newly formed ions are kept together by electrostatic interactions or hydrogen-bridges, an MS3 experiment was performed: first m/z 1015(7+) was subjected to ETD activation, and then the most abundant charge-reduced ion m/z 1422(5+) was subjected to CID analysis (Figure 1). Only a few abundant fragments were detected. The mass differences observed in the first ion cluster were 31.9719 Da (1204-1172) & (1238-1206), and 33.9881 Da (1206-1172) & (1238-1204). The theoretical mass for S is 31.9720 and for H₂S is 33.9877, and so the observed mass differences are the signature of an intermolecular disulfide linkage and correspond to cleavages across the disulfide bridge [13]. Since the other abundant fragment at m/z 1474.89(4+) corresponds to the 'missing' part of the molecule, these data suggested the presence of a disulfide-bound heterodimer.

Disulfide-bridges in toxins are rather common. Thus, we proceeded with reduction/ alkylation of the molecule following a general protocol with DTT and iodoacetamide. This reaction, attempted multiple times, led to complete sample loss. First we suspected solubility problems as a result of carbamidomethylation, and also experimented with iodoacetic acid with the same negative results. Oxidation of the disulfides was an option that we discarded, since the introduction of negatively charged groups (i.e. cysteic acids) would be unlikely to improve the charge density and ETD efficiency.

We decided to keep the peptide in an acidic solution (0.1% formic acid in water) resembling that which was used for its purification and performed the reduction at low pH with TCEP. Extended reaction time, elevated temperature and larger than usual reagent excess was used in order to achieve complete reduction. The results were somewhat surprising. The major component observed by mass spectrometry demonstrated a molecular weight equal to that of the full-length polypeptide with a 6 Da mass increase, indicating 3 reduced disulfide bridges. Its ions were detected at m/z 889.1488(8+), 1016.0285(7+), 1185.1955(6+) and 1422.0243(5+), yielding a monoisotopic mass of 7106.1281 (MH^+). The (7+) ion of the full length toxin before and after the reduction is shown in Figure 2. In addition, the reduction yielded two individual peptides, one with $MH^+ = 5900.6184$ (m/z 738.4600(8+), 843.8096(7+) and 984.2748(6+); Figure 2, lower panel), and the other with $MH^+ = 1224.5352$ (also at m/z 612.7715(2+)), which is 18 Da larger than expected based on the MS3 fragmentation.

The smaller peptide ($MH^+ = 1224.5352$) was analyzed first from the infused reaction mixture. Sequencing this peptide was relatively straightforward; we had good quality CID and HCD data for both the doubly and singly charged precursor ions, with the latter ones being more informative for *de novo* sequencing (Figure 3). We started with the CID data, since it was expected that both N-terminal and C-terminal ions would be present, while **b**-type fragments frequently do not survive the multiple collisions occurring during HCD activation (Nomenclature: [14]). The very abundant ion triplet at m/z 1063, 1091 and 1109 was readily identified as **a**_{n-1}, **b**_{n-1} and **b**_{n-1}+H₂O (where n= the number of residues in the peptide) revealing that the C-terminal residue must be Asp (1224-1109=115) and also suggesting that there must be a basic residue somewhere in the sequence, because such rearrangement has been reported for peptides with preferential charge retention at the N-terminus [15, 16]. The other **a-b** fragment pairs could be easily identified downstream at m/z 934 & 962, 771 & 799, 668 & 696, 521 & 549, identifying Glu, Tyr, Cys, and Phe, respectively. Interestingly and unexpectedly, internal ions were also detected in the CID spectrum, m/z 582 and 554, that were separated by 28 Da as well, but luckily ions belonging to the **b** ion series also featured an ammonia loss and were therefore distinguishable. Thus, the sequence could be tentatively determined at this point to be ...FCYED. However, the clues fizzled out here, and we had to turn to the more complex HCD data to proceed (Figure 3, lower panel). From there the next **a-b** pair was identified at m/z 353 & 381 (the latter was also detected in the CID data). The 168 Da gap between this **b** fragment and the previously identified **b** at m/z 549 could only correspond to a Pro-Ala combination. Since Pro residues usually produce an abundant **y** fragment, it was easy to determine that these amino acids are indeed present and that they appear in this order. Hence, m/z 478 (and the highly abundant m/z 461 ion the represents an ammonia loss) could be unambiguously identified as the "missing" **b**-fragment, while the **y** fragment formed *via* cleavage at the N-terminal side of Pro is at m/z 844, and our working sequence is now ...PAFCYED. Considering internal fragments from this sequence helps to assign numerous ions in the low mass region, such as m/z 169 & 141 (PA & PA-28), 316 & 288 (PAF & PAF-28), 267 & 239 (CY & CY-28) etc. (for complete list see Table 1). At the same time we believe that there must be a basic amino acid in the sequence, and from low mass ions at m/z 112 and 115, one suspects an Arg [17]. Indeed, there is an **a-b** pair at m/z 197 and 225, indicating that the next residue towards the

N-terminus is an Arg. However, according to the 'MS-Comp' feature of Protein Prospector, there is no amino acid combination that would yield a **b** fragment at this mass, even if we permit a 50 ppm mass measurement error, which is much higher than our instrument actually affords. However, if a common N-terminal modification, the cyclization of Gln residue is considered, then the fragment is within 2 ppm for **b**₂ of <Gln[Ile/Leu] (N-terminal acetylation was also considered and checked).

Thus, the final sequence of the short peptide is <Gln-[Ile/Leu]-Arg-Pro-Ala-Phe-Cys-Tyr-Glu-Asp. Fragments calculated by 'MS-Product' of Protein Prospector for ESI-Q-TOF instrument selection (which also corresponds to HCD fragmentation) and fragments observed are listed in Table 1, along with the errors in mass measurement.

We tested the most popular sequencing program, PEAKS, *de novo* with this spectrum. The raw data were loaded, the precursor ion *m/z* value was manually corrected, and the spectrum was processed by the program. For *de novo* sequencing, no enzyme was specified, and the mass accuracies considered were 5 ppm and 0.02 Da for the precursor and fragments, respectively (relative mass accuracy cannot be specified for fragments). After a series of failed attempts we realized that instrument selection was critical, as the software had to recognize the data as 'high energy CID' fragmentation in order to determine sequences (even though this definition is not correct). From this experience we arrived at the same conclusion as the 2011 ABRF iPRG study: being highly familiar with the software produces much better results [18]. When ion trap CID was considered, the abundant internal fragments proved sufficient to trip the algorithm. Q-TOF, Q-FT, and FTMS instrument selection yielded the correct sequence (obviously only if the software was informed about the possibility of blocked terminus, i.e. N-acetylation or pyroglutamate formation were permitted). However, less than 10% confidence was attached when the first 2 instruments were specified. With the FTMS instrument selection, the following sequence was determined with 90% confidence: [224.1]RPA[413.12]E[115]. The software considers **y** fragments twice as significant as **b** ions, and considers immonium ions and internal fragments only during the 3rd step while reevaluating the 10000 best candidates [6]. The first premise is not necessarily true for non-tryptic sequences, as illustrated with the abundant N-terminal ions of this peptide, and we believe, this fact definitely contributed to the low confidence of the sequence determination. At the same time we cannot tell which internal fragments were identified and used by the algorithm: The fragment labeling option included the internal fragments as default, however these ions were not assigned by the software. Strangely only *m/z* 112 and 169 were indicated, both incorrectly as PAFCY-28. The confidence difference between the assignments is also puzzling, since data processing with each instrument-type yielded 93 identical masses (judging from the mass accuracy charts accompanying each assignment).

Next, we turned to analyzing the larger ($MH^+ = 5900.6184$) peptide that was observed after reduction of the toxin. While the short peptide gave good quality data upon direct infusion of the reaction mixture, the longer sequence had to be analyzed by LC/MS/MS analysis. The precursor ion selection was restricted, alternating CID and ETD data were acquired from *m/z* 738(8+). Both activation steps were performed in the linear trap, and the fragments were measured in the Orbitrap.

CID spectra and ETD spectra between 38.5 and 39 min (representing the apex of the eluting peak), were merged starting around 50% XIC intensity and were used for partial *de novo* sequencing (spectra are shown with Supplementary Tables 1 and 2). Monoisotopic masses and charge states were determined manually and inserted into an Excel workbook, using separate sheets for CID and ETD data (Supplementary Tables 1 and 2, Table 2 represents a combined, concise version). We compared the manual peak-picking with that of the deconvolution program, Xtract supplied by the manufacturer. Because of the overlapping multiply-charged peaks and weak or absent monoisotopic ions, neither solution was entirely satisfactory (Supplementary Tables 3-5). We found that the Xtract program tends to overlook or misassign some singly-charged ions (see Supplementary Table 3), but otherwise the greatest difficulties are caused by overlapping fragment ions. In general, for large multiply-charged ions both “eyeballing” and modeling based on the “average” peptide composition work equally well, but not completely reliably. The manually-generated peak list may not be as complete as the software-generated one, but it contained all of the important singly-charged ions and was definitely more reliable for *de novo* sequencing. This was especially true for the information-rich CID data. In the Excel workbook the singly-charged accurate masses were calculated from the monoisotopic masses and charges using the formula $m/z * z - (z-1) * 1.007825$. Then, the fragment masses were sorted. Since members of the ion series usually cannot be identified *per se*, a series of other masses were calculated using the observed masses as reference points. In the CID Table (Table 2, Supplementary Table 1) the nominal masses of the corresponding complementary ions (**y** fragment for **b** and vice versa) were calculated using the formula $b_i + y_{n-i} = MH^+ + 1$. In the ETD Table (Table 2, Supplementary Table 2) two columns were created: with masses 17.0265 Da (NH₃) lower and 16.0187 Da (NH₂) higher than the fragments detected, representing the corresponding potential **b** and **y** fragments, respectively. With all this information in the Excel worksheets, the manual interpretation began. Determining the termini was relatively easy. ETD-fragment m/z 197 was identified as **z**₂, consisting of a His and a Gly, while m/z 212 is **c**₂ corresponding to ProPro. These assignments are unambiguous because these were the only potential structures within 5 ppm (according to the MS comp feature of Protein Prospector). Based on the ETD data, a series of **b** fragments could be identified in the CID spectrum, and so the sequence was built from the N-terminus. First, 2 Phe residues were added to the sequence, but these were followed by a 256.155 gap that could be filled by either an {Ala, Gly, Lys} or a {Gln, Lys} combination. Obviously an Ala, Gly combination would yield the very same ‘**c**’ fragment as the Gln. We discarded this option for the lack of any supporting ions in the CID. Based on this information, the **c** fragment at m/z 634 was considered, which, due to the afforded mass accuracy, identifies a Gln in position-5 and, consequently, a Lys in position-6.

Interestingly, this ‘**c**’ fragment was also detected in the CID spectrum, as will be discussed later. Assuming that m/z 865 in the ETD peak list was a **c** fragment, the next residue was identified as Cys, and the following **c**-fragments, originally assigned based on the corresponding **b**-ions in CID, determined that Gly, Ala, and Phe are the next three residues. Thus, our working sequence at this point is ProProPhePheGlnLysCysGlyAlaPhe.

The CID data did not help for the next two residues, but **c** ions detected in the ETD spectrum determined them to be Val and Asp. The next 3 residues were assigned as SerTyrTyr in a similar manner. This step also identified *m/z* 1621 as a potential **y** fragment because it was present in both datasets and does not belong to the N-terminal series. By the same logic, the CID fragment at *m/z* 1735 must be also a **y** fragment. Following the clues provided by the ETD data, considering amino acid mass differences, and taking advantage of the mass accuracy, we could confidently determine the N-terminal sequence as PPFFQKCGAFVDSYYFNRS.

The CID data provided an additional non-adjacent sequence stretch, which was determined as HFFYGQCDV. Just as the **c** ions in the ETD spectrum provided information for the N-terminal amino acids, the first 6 residues were identified from consecutive **b** fragments. The last **b** fragment (*m/z* 3606) could then be tied to a series of **y** ions. The complementary **y** fragment at *m/z* 2295 was the first in this series, and *m/z* 1977 identified by ETD data was the last. (Table 2, Supplementary Tables 1 and 2). The **y** fragment identifying the Val residue (*m/z* 2077) was detected in 2 different charge states, and the ETD data excluded it from the **b**-series. The 218 Da gap between this mass and *m/z* 2295 may correspond to 3 different amino acid combinations: AlaPhe, SerMet and CysAsp, and the latter one was supported by a **y** fragment detected. Since the ETD spectrum provided a string of **c**-fragments, **c**₂-**c**₁₈, each measured within 5 ppm, one would expect that a computer program could have easily determined the N-terminal sequence. Thus, the raw data were loaded into the PEAKS software with the FTMS(ETD) instrument selection. We experimented with different mass accuracy windows from 0.005-0.1 Da, and in one trial also permitted methionine oxidation. All attempts yielded the correct N-terminal sequence, but unfortunately very little confidence was attached to it. The most confident assignment was: PPFFQKCGAFVDSYYFNRSCTCGWLMVTGPCHGRNFFYSDVFAGCKGSMTRV, where the amino acids in bold were assigned with a confidence higher than 60%, and those in both bold and italic indicate higher than 90% confidence level. The mass error permitted for this assignment was 0.1 Da. Notably, the underlined sequences are identical to one another, but in reverse order. As it happens, while *m/z* 359 could still be identified unambiguously as a **c** fragment (within 5 ppm), all the other N-terminal fragments could 'double' as **z** ions. Thus, it is clear that the supporting information from the CID data was essential to decide the position and correct order of the amino acids. In addition, when PEAKS software was used to search the CID data with 'ion trap CID' selected, it did not yield even a sequence tag. However, when the Orbitrap/Orbitrap combination was selected, then the HFFYG internal sequence was confidently identified. Interestingly, although this spectrum represented ion trap CID data, the number of peaks and masses seemed to be the same after peak-picking for both instrument selections (the same parameters were used for both searches).

Since the short peptide featured a C-terminal Asp, while the longer peptide had a Pro at its N-terminus, it was now easy to explain the 18 Da mass discrepancy between the expected and observed mass of the shorter peptide after reduction. Instead of a disulfide-bound heterodimer, the toxin consists of a single chain that experienced a double cleavage event in the MS3 experiment. It is possible that first the disulfide-bridge was broken by ETD [19],

and then CID activation induced the Asp-Pro peptide-bond cleavage. However, there is another fragmentation pathway one has to consider: a smaller peptide population that survived the ETD activation intact underwent a double cleavage upon collisional activation. The fragmentation-prone Asp-Pro bond could have been broken first, and then fragmentation could have occurred along the disulfide bridge. Only this second sequence of events explains the characteristic ion-triplets that were observed. The Asp-Pro bond is acid-sensitive and must have hydrolyzed during the extended reduction at low pH, producing two individual peptides.

Finding sequence similarity among known proteins frequently serves as confirmation of sequences determined *de novo*. A BLAST search was performed using the N-terminal sequence stretch (QLRPAFCYEDPPFFQKCGAFVDSYYFNRS), which did not yield any significant hits. Thus, the full sequence and the protein family to which our toxin belongs was determined by cloning the toxin-encoding cDNA using degenerate oligonucleotide-probes representing the 10 N-terminal residues of the longer peptide, as described earlier [11].

QIRPAFCYEDPPFFQKCGAFVDSYYFNRSRITCVHFFYGQCDVNQNHFTTMSECNRVCG, the final sequence, matched the experimentally determined molecular mass perfectly, after adjustment for the N-terminal pyroglutamic acid and three disulfide bridges. A BLAST search with the complete sequence categorized this polypeptide as a Kunitz-type protein, showing 40% sequence identity to its closest relative up to date, Vestiginin-3, of *Demansia vestigiata* (Black whip snake) and complete conservation of the six Cys residues [11].

Once the full sequence was determined, we reevaluated the CID and ETD data of the longer sequence. Manual inspection of the ETD data provided some additional information in the form of Cys-specific ‘w’-type fragment ions, which are formed from **z**. fragments *via* side-chain losses, a phenomenon reported for alkylated Cys-residues in ECD [20]. While the observation of these fragments was not unexpected, the detection of a series of potential **b** + 2H ions is most unusual and awaits explanation (Supplementary Table 2). We also used a sequence-based peak-picking software to test the information content of both the CID and ETD spectra [12]. Since the peak-picking of this software is based on the expected theoretical ion clusters calculated not from average but from real elemental composition, it can identify the fragment ions even from overlapping isotope clusters. While this approach did not make any difference for this particular ETD data (Supplementary Table 4), the CID spectrum contained substantially more information than either the manually prepared or Xtract-based peak lists revealed (Supplementary Table 3). Thus, as final confirmation a program utilizing the calculated masses and isotope distributions based on the sequence determined – as FAVA does – can reveal a wealth of supporting information still “hidden” in the spectrum, or may help to decide between different isobaric options.

While ETD analysis of the non-reduced peptide yielded very limited information even after CID activation of a charge-reduced ion (Figure 1) as discussed above, the polypeptide underwent significant fragmentation upon collisional activation that can be deciphered using the known sequence (Figure 4A and B, Supplementary Table 5). Because of the disulfide-bridges and the positions of the Cys-residues (residue 7 is the first Cys residue, and the last

is in the 58th position) only 6 N-terminal and 2 C-terminal residues are outside of the cross-linked structure. Thus, very few fragments can be formed *via* single bond cleavages. Indeed, 'regular' fragmentation was only detected from the N-terminal part. Interestingly, the position of one of the disulfide-bridges could be determined because of the favored fragmentation at the Asp-Pro linkage. Fragments **y₃-y₅** were detected with a 1203 Da shift, that indicates linkage between Cys-7 and Cys-58, since the mass increment corresponds to the N-terminal 10 amino acids linked to the C-terminal fragments. We assume that first the Asp-Pro bond most prone to fragmentation was cleaved, but this cleavage product has the very same molecular mass as the original precursor ion and thus was further activated (just like described above for the MS3 experiment). This could explain all the internal fragments as well as the internal amino acid losses from the intact molecular ion.

This CID spectrum also features two **c**-type internal fragments, formed in front of Gln and Lys residues. It has been published that abundant **c₁** ions can be observed in sequences in which the second residue is a Gln and for which the activation was performed in a collision cell [21]. The authors suggested that the **c** ions were formed from the **b** fragment corresponding to the next amino acid *via* the loss of a 6-membered ring. Such ring formation is also possible for Ser, Arg, His and Lys residues, based on thermodynamic calculations [22]. Thus, the observed **c** ion in front of the Gln and Lys residues can be produced *via* the same mechanism. The very same **c** fragments as well as **c₂₉** were also detected in the CID of the bigger reduced peptide (Supplementary Table 1). This is the first time that such ion formation is reported for Lys residues, for higher sequence positions, and in ion trap CID experiments.

Conclusions

Even the most popular *de novo* sequencing program has difficulties assigning sequences from unusual or incomplete fragmentation patterns. Sequence-specific fragments, such as in our case the C-terminal **a**, **b**, **b**+H₂O triplet, that significantly aid manual sequence determination frequently prove to be the undoing of the automated approach. In addition, reliable peak-picking/deconvolution from a complex spectrum is still a challenging task as illustrated with our high quality data. When peptide identification is the goal, these obstacles are easier to overcome than when a novel sequence has to be determined. Recently, numerous groups have started to advocate *de novo* sequencing instead of comparative database searching [23]. While the software available seems to function well for ion trap data, and for tryptic peptides, researchers aiming at developing such programs also should consider the wide-variety of sequence- or residue-specific fragmentation, which may provide important clues but remain not only underutilized, but rather completely ignored as insignificant. We would not recommend to incorporate such fragment ions into the search algorithm, but they could be (and should be) used to confirm the determined sequence and/or aid in the selection between similarly scoring alternatives. Furthermore, the recent surge in commercially available instruments that can measure precursor as well as fragment ions with a few ppm mass accuracy should force search engines to take advantage of this, since the results may not be reliable enough when only absolute mass accuracy can be specified.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank David Maltby for his technical assistance, and Shenheng Guan for his help with FAVA. KFM was supported by NIH grant NCRR P41RR001614 and the Howard Hughes Medical Institute (both support the National Bio-Organic Biomedical Mass Spectrometry Resource Center at UCSF, director: A.L. Burlingame). CJB was supported by a Ruth Kirschstein predoctoral fellowship (F31NS065597).

References

1. Hunt DF, Yates JR III, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci USA*. 1986; 83:6233–6237. [PubMed: 3462691]
2. 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]
3. Tipton, KF.; Dajas, F., editors. Neurotoxins in neurobiology: their actions and applications. Ellis Horwood Limited; Chichester: 1994.
4. 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]
5. 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]
6. 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]
7. 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]
8. Medzihradzky KF. Peptide sequence analysis. (Review). *Meth Enzymol*. 2005; 402:209–244. [PubMed: 16401511]
9. Nielsen ML, Savitski MM, Zubarev RA. Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry. *Mol Cell Proteomics*. 2005; 4:835–845. [PubMed: 15772112]
10. 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]
11. Bohlen CJ, Chesler AT, Sharif-Naein N, Medzihradzky KF, Zhou S, King D, Sánchez EE, Burlingame AL, Basbaum AI, Julius D. Heteromeric toxin from coral snake targets acid sensing ion channels to produce pain. *Nature*. 2011; 479:410–414. [PubMed: 22094702]
12. Guan S, Burlingame AL. Data processing algorithms for analysis of high resolution MSMS spectra of peptides with complex patterns of posttranslational modifications. *Mol Cell Proteomics*. 2010; 9:804–810. [PubMed: 19955076]
13. Bean MF, Carr SA. Characterization of disulfide bond position in proteins and sequence analysis of cystine-bridged peptides by tandem mass spectrometry. *Anal Biochem*. 1992; 201:216–226. [PubMed: 1632509]
14. Biemann K. Nomenclature for peptide fragment ions (positive-ions). *Meth Enzymol*. 1990; 193:886–887. [PubMed: 2074849]
15. 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]

16. 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.
17. Gehrig PM, Hunziker PE, Zahariev S, Pongor S. Fragmentation pathways of N^G -methylated and unmodified arginine residues in peptides studied by ESI-MS/MS and MALDI-MS. *J Am Soc Mass Spectrom.* 2004; 15:142–149. [PubMed: 14766281]
18. http://www.abrf.org/ResearchGroups/ProteomicsInformaticsResearch-Group/Studies/iPRG2011_poster_ABRF_final-2.pdf
19. Zubarev RA, Kruger NA, Fridrikson EK, Lewis MA, Horn DM, Carpenter BK, McLafferty FW. Electron Capture Dissociation of Gaseous Multiply-Charged Proteins is Favored at Disulfide Bonds and Other Sites of High Hydrogen Atom Affinity. *J Am Chem Soc.* 1999; 121:2857–2862.
20. 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]
21. Lee YJ, Lee YM. Formation of c_1 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]
22. Farrugia JM, O'Hair RAJ, Reid GE. Do all b_2 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.
23. 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]

Figure 1.

MS³ ETD/CID 1015(7+) $\rightarrow$ 1422(5+) spectrum of the intact molecule. Sample solution was infused, activation was performed in the linear trap, and fragments were measured in the Orbitrap. The fragments detected suggested the presence of a disulfide-bound heterodimer. The shorter component displayed the characteristic ion-triplet, as indicated. Fragments formed *via* cleavage across the disulfide bridge are labeled with asterisks.

Figure 2.

Results of the ESIMS analyses. The upper and middle panels show the (7+) ion for the intact toxin before and after reduction, respectively. The lower panel shows the (8+) ion of the bigger individual peptide formed during the low pH reduction experiment.

Figure 3.

CID (upper panel) and HCD (lower panel) of m/z 1224.53. Fragments in both analyses were measured in the Orbitrap. The peptide sequence was determined from these data as <QL/IRPAFCYED. In the CID spectrum the unexpected internal fragments are labeled, while some of the abundant fragments were assigned in the HCD spectrum. Annotation $^{\circ}$ indicates ammonia loss. Otherwise the Biemann nomenclature was applied [14]. For complete fragment assignment see Table 1. The base peak in the CID spectrum featured an intensity ~ 3000 , while in HCD ~ 6000 .

Figure 4A.

Figure 4B.

Figure 4.

Figure 4A. CID spectrum of the intact polypeptide. Precursor ion was 1015(7+) (Figure 2, upper panel). Sample solution was infused, CID activation was performed in the ion trap, and fragments were measured in the Orbitrap. ○ indicates ammonia loss. * indicates that the fragment contains a dehydroAla instead of the Cys. ** indicates that the Asp-Pro linkage was cleaved and the N-terminal peptide is linked to the C-terminal fragment with a disulfide-bridge, i.e. Cys-7 and Cys-58 are linked. ♦ indicates 'c' type internal fragment formation in front of Gln and Lys. For complete peak list and assignments see Supplementary Table 5.

Figure 4B. CID spectrum of the intact polypeptide. Precursor ion was 1015(7+) (Figure 2, upper panel). Sample solution was infused, CID activation was performed in the ion trap, and fragments were measured in the Orbitrap. Fragments 'S' and 'L' correspond to the short and long sequence when both the Asp-Pro linkage and the disulfide-bridge were broken. The subscript indicates that both sulfurs were retained on the fragment. Double bond cleavages and amino acid losses were observed as indicated from both the intact molecule and the larger fragment. For complete peak list and assignments see Supplementary Table 5.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

I cannot differentiate between isomeric amino acids Ile and Leu.															
<Gln[Ile/Leu]/ArgProAlaPheCysTyrGluAsp = <QL/IRPAFCYED															
calc mass	measured	ppm	ID	calc mass	measured	ppm	ID	calc mass	measured	ppm	ID	calc mass	measured	ppm	ID
237.1346	237.1339	-3	RP-NH3	391.1798			PAFC-28	558.2493	558.2476	-3	RPAFC-NH3	826.3076	826.3048	-3	y7-H2O
239.0849	239.0842	-3	CY-28	396.1224			CYE	568.3242			LRPAF-NH3	834.3967			LRPAFCY-NH3
242.1975			LR-28	408.1401			y3-H2O	575.2759			RPAFC	839.3869			RPAFCYE-28
245.0768	245.0761	-3	y2-H2O	410.2874			LRPA-28	582.2381	582.2363	-3	PAFCY	844.3182	844.3147	-4	y7

Table 2

Simplified peak list from the MS/MS analysis of the longer peptide. Comments indicate how the fragments were assigned. Residues linking up within the fragment series are indicated. Complementary y/b ions are listed for both CID and ETD data. Detailed explanation is in the text, and for complete lists, see Suppl. Tables 1 and 2.

deconv.	CID Comments	Residues ProPro	comp.f.	corr. CID fragments		deconv.	ETD Comment	Residues (HisGly) ProPro
				b	y			
342.1808	b, based on ETD	Phe↑	5559.6	180.0532	213.0984	197.0797	z2 for GH	
349.1504			5552.6	195.1128	228.1580	212.1393	c2 for PP	
461.2555	a, correct mass		5440.5	298.1672	331.2124	315.1937		
489.2495	b, based on ETD	Phe↑	5412.6	309.0639	342.1091	326.0904		
506.2758			5395.5	327.1700	360.2152	344.1965		
634.3351	c, based on ETD	Gln↑	5267.5	342.1809	375.2261	359.2074	c3, CID & MS comp	Phe↑
728.3748			5173.4	398.1492	431.1944	415.1757		
745.4025	b, based on ETD	Lys↑	5156.4	474.2378	507.2830	491.2643		
788.3290			5113.5	489.2485	522.2937	506.2750	e4, based on CID	Phe↑
848.4122			5053.4	539.2383	572.2835	556.2648		
905.4288	b, based on ETD		4996.4	617.3072	650.3524	634.3337	e5, MS-comp	Gln↑
976.4706	b, based on ETD	Ala↑	4925.3	668.2931	701.3383	685.3196		
1004.4040			4897.4	745.4035	778.4487	762.4300	e6, based on CID	Lys↑
1095.5432	a, correct mass		4806.3	848.4116	881.4568	865.4381	e7, based on CID	Cys↑
1105.5262	b-water		4796.3	890.4171	923.4623	907.4436		
1123.5374	b, based on ETD		4778.3	905.4335	938.4787	922.4600	e8, based on CID	Gly↑
1135.4384			4766.4	976.4692	1009.5145	993.4958	e9, based on CID	Ala↑
1337.6290	b, based on ETD		4564.2	1123.5376	1156.5829	1140.5642	c10, based on CID	Phe↑
1621.6580	y, based on ETD		4280.1	1218.4608	1251.5061	1235.4874		
1735.7065	y, based on ETD		4166.1	1222.6068	1255.6521	1239.6334	c11, from mass diff; acc	Val↑
1813.8824			4087.9	1337.6322	1370.6775	1354.6588	c12, based on CID	Asp↑
1841.8738			4059.9	1424.6648	1457.7101	1441.6914	c13, from mass diff; acc	Ser↑
1977.8071	y, from mass diff; acc.		3924.0	1425.6704	1458.7156	1442.6969		
				1466.5924	1499.6377	1483.6190		

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

	CID	Comments	Residues	comp.f.	corr. CID fragments		deconv.	ETD	Residues
deconv.					b	y		Comment	
2059.8556	y-ammonia			3841.9	1555.6426	1588.6879	1572.6692		
2076.8722	y, based on ETD		Val†	3824.9	1587.7274	1620.7727	1604.7540	c14, from mass diff; acc	Tyr†
2117.0048				3784.8	1604.6362	1637.6815	1621.6628	must be y ion	
2191.8922			Asp†	3709.9	1750.7906	1783.8359	1767.8172	c15, from mass diff; acc	Tyr†
2277.8844	y-ammonia			3623.9	1797.7370	1830.7823	1814.7636		
2294.9131	y, based on 'b'		Cys†	3606.9	1846.7344	1879.7797	1863.7610		
2310.9109				3590.9	1897.8578	1930.9031	1914.8844	c16, from mass diff; acc	Phe†
2387.9323				3513.9	1911.7830	1944.8283	1928.8096		
2404.9582	y-water			3496.8	1944.7648	1977.8101	1961.7914		
2422.9661	y, based on 'b'			3478.8	2011.9030	2044.9483	2028.9296	c17, from mass diff; acc	Asn†
2430.9634				3470.8	2151.9926	2185.0379	2169.0192		
2479.9888	y, based on 'b'			3421.8	2168.0048	2201.0501	2185.0314	c18, from mass diff; acc	Arg†
2625.0301	y-water			3276.8	2168.0031	2201.0484	2185.0297		
2626.0329	y-ammonia			3275.8	2174.8794	2207.9247	2191.9060		
2643.0533	y, based on 'b'			3258.7	2228.8854	2261.9307	2245.9120		
2728.2748				3173.5	2255.0190	2288.0643	2272.0456	c19, from mass diff; acc	Ser†
2790.1297	y, based on 'b'			3111.7	2390.9206	2423.9659	2407.9472		
2799.3484	a, b-a pair			3102.5	2525.2424	2558.2877	2542.2690		
2827.3435	b, b-a pair			3074.5	3345.3153	3378.3606	3362.3419		
2936.4127	a, b-a pair			2965.4	3701.5464	3734.5917	3718.5730		
2964.4032	b, b-a pair		His†	2937.4	3971.6732	4004.7184	3988.6997		
3014.4105				2887.4	3972.6891	4005.7344	3989.7157		
3083.4793	a, b-a pair			2818.3	4117.7216	4150.7668	4134.7481		
3111.4722	b, b-a pair		Phe†	2790.3	4281.8544	4314.8996	4298.8809		
3161.4809				2740.3	4444.8224	4477.8676	4461.8489		
3230.5462	a, b-a pair			2671.3					
3258.5382	b, b-a pair		Phe†	2643.3					
3393.6097	a, b-a pair			2508.2					

deconv.	CID	Comments	Residues	comp.f.	corr. CID fragments		deconv.	ETD	Residues
					b	y			
3421.6042		b, b-a pair	Tyr↑	2480.2					
3478.6262		b, from mass diff; acc.	Gly↑	2423.2					
3495.6442				2406.2					
3578.6767		a, b-a pair		2323.1					
3606.6845		b, b-a pair	Gln↑	2295.1					