Peptide de Novo Sequencing Using 157 nm Photodissociation in a Tandem Time-of-Flight Mass Spectrometer

Liangyi Zhang and James P. Reilly*

Department of Chemistry, Indiana University, 800 East Kirkwood Avenue, Bloomington, Indiana 47405

It has previously been shown that photodissociation of tryptic peptide ions with 157 nm light in a matrix-assisted laser desorption/ionization (MALDI) tandem time-of-flight (TOF) mass spectrometer generates an abundance of x-type ions. A peptide de novo sequencing algorithm has now been developed to interpret these data. By combination of photodissociation and postsource decay (PSD) spectra, the algorithm identifies x-type ions and derives peptide sequences. The confidence of amino acid assignments is evaluated by observing complementary y-, v-, and w-type ions that provide additional constraints to sequence identification. In the analysis of 31 tryptic peptides from 4 model proteins, the algorithm identified 322 (or 90.7%) of the 355 amino acids and made only 3 incorrect assignments. The other 30 amino acids were not identified because specific needed x-type ions were not detected. Based on the observation of v- and w-type ions, 45 of 50 detected leucine and isoleucine residues were successfully distinguished and there was only one mistake. The remaining four residues were not distinguished because the corresponding v- and w-type ions were not detected. These de novo sequencing results translated into successful identification of proteins through homology searches. To evaluate the robustness of the present sequencing approach, a collection of 266 tryptic peptides from 23 model proteins were analyzed and then sequenced. A total of 167 peptides yielded sequence tags of 5 or more residues. In 5 peptides, 1 or 2 residues were incorrectly assigned.

Mass spectrometry (MS) is widely used to investigate biological systems following recent advances in both instrumentation and bioinformatics.[1−5] A number of MS-based techniques have been developed to characterize protein constituents in biological samples.[6−8] The two most common protein-identification methods involve tandem mass spectrometry (MS/MS)[9,10] and MALDI peptide mass mapping.[11−13] In either case, proteolytic peptides are analyzed by mass spectrometry and proteins are assigned by comparing mass spectrometric data with predicted peptide and fragment masses derived from a protein sequence database. Even though these methods have been successfully applied in numerous experiments, they have several fundamental limitations. Experimental data do not lead to correct protein identifications when there are database errors, genetic mutations, and modifications that occur post-translationally or during sample handling. In addition, some peptide fragmentation spectra contain limited sequence information. As a result, only about 10−20% of spectra typically lead to peptide identifications, although some high-quality experiments do yield as high as 50% identifications.[14,15] Furthermore, organisms without sequenced genomes cannot be studied by database-matching techniques. Finally, protein databases continue to grow in size, so the time it takes to search against them increases exponentially. In light of these limitations, methods that can identify peptides without protein databases are desirable.

De novo sequencing methods have been developed to derive peptide sequences from tandem mass spectra without reference to a database.[16] Typically de novo sequencing algorithms identify amino acids using mass differentials between consecutive peaks in tandem mass spectra. Several of these algorithms have been developed to interpret low-energy collisionally induced dissociation (CID) spectra including Sherenga, Lutefisk, PEAKS, DACSIM, EigenMS, PepNovo, NovoHMM, and MSNovo.[1,17−23] Most pro-

* To whom the correspondence should be addressed. E-mail: reilly@indiana.edu.

(1) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327–342.
(2) Domon, B.; Aebersold, R. *Science* **2006**, *312*, 212–217.
(3) Forner, F.; Foster, L. J.; Toppo, S. *Curr. Bioinf.* **2007**, *2*, 63–93.
(4) Frank, A. M.; Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A.; Pevzner, P. A. *J. Proteome Res.* **2007**, *6*, 114–123.
(5) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. *Nat. Methods* **2007**, *4*, 787–797.
(6) Yates, J. R. *J. Mass Spectrom.* **1998**, *33*, 1–19.
(7) Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001**, *101*, 269–295.
(8) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198–207.
(9) Hunt, D. F.; Yates, J. R.; Shabanowitz, J.; Winston, S.; Huauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233–6237.
(10) Biemann, K.; Schoble, H. A. *Science* **1987**, *237*, 992–998.
(11) Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5011–5015.
(12) Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3*, 327–332.
(13) Yates, J. R.; Speicher, S.; Griffin, P. R.; Hunkapillar, T. *Anal. Biochem.* **1993**, *214*, 397–408.
(14) Cox, J.; Hubner, N. C.; Mann, M. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 1813–1820.
(15) Deutsch, E. W.; Lam, H.; Aebersold, R. *Physiol. Genomics* **2008**, *33*, 18–25.
(16) Standing, K. G. *Curr. Opin. Struct. Biol.* **2003**, *13*, 595–601.
(17) Taylor, J. A.; Johnson, R. S. *Anal. Chem.* **2001**, *73*, 2594–2604.
(18) Zheng, Z. Q. *Anal. Chem.* **2004**, *76*, 6374–6383.
(19) Bern, M.; Goldberg, D. *J. Comput. Biol.* **2006**, *13*, 364–378.
(20) Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77*, 964–973.
(21) Fischer, B.; Roth, V.; Roos, F.; Grossmann, J.; Baginsky, S.; Widmayer, P.; Gruissem, W.; Buhmann, J. M. *Anal. Chem.* **2005**, *77*, 7265–7273.
(22) Ma, B.; Zhang, K. Z.; Hendrie, C.; Liang, C. Z.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337–2342.
(23) Mo, L. J.; Dutta, D.; Wan, Y. H.; Chen, T. *Anal. Chem.* **2007**, *79*, 4870–4878.

10.1021/ac902050y © 2010 American Chemical Society
Published on Web 01/08/2010

ceed through two major steps. First, a pool of sequence candidates is generated based on peak mass differentials. The most commonly used method for sequence generation employs the principles of graph theory in which a spectrum graph is created and paths that connect a number of peaks represent peptide sequences.[23] Since different series of fragment ions cannot be differentiated by most conventional algorithms,[1,17–23] a large number of sequence candidates are normally created. Second, a scoring algorithm is designed to rank the derived sequence candidates and then the top ranking sequence is considered to be the correct sequence. Since most scoring approaches aim to identify the best sequence from a pool of candidate sequences, the generated scores usually represent some measure of the confidence level of a derived sequence.[23] De novo sequencing not only directly identifies peptides from organisms without a database; it also provides an alternative approach for protein identification by sequence homology comparison when a database is available. As a result, de novo sequencing provides a nonredundant way to confirm protein identifications derived by database searching algorithms.[24] Sequence-tag based searching also combines de novo sequencing and database searching algorithms. In this approach, de novo sequencing is first used to generate short peptide sequences from fragmentation data. Along with a few fragment ion masses, these short sequences are then matched against protein databases for peptide identification. Since sequence tags significantly constrain the peptide pool during searches, this hybrid approach is much more efficient than traditional database searching algorithms.[25,26]

Despite its advantages, de novo sequencing has not become a routine protocol to interpret proteomic data. One reason is that it requires a series of fragment ions that extend through each peptide sequence. Unfortunately, since low-energy CID of peptides induces preferential backbone cleavages,[27] most CID spectra are dominated by a limited number of peaks that correspond to cleavage of weak bonds. As a result, only portions of peptide sequences can be derived. A second complication for de novo sequencing is that most tandem mass spectra contain multiple series of ions. For example, both b- and y-type ions appear in low-energy CID spectra and until a peptide is actually identified, there is no simple way to distinguish them. Confusion of these two ion series leads to incorrect sequences when the spacing between two peaks happens to match the mass of an amino acid. According to a recent assessment of several de novo sequencing software packages, only 66% or fewer of the amino acids in peptide sequences are correctly identified in the analysis of tryptic peptides from model proteins using low-energy CID or TOF-TOF CID.[28] All of the remaining residues were incorrectly assigned. Several peptide derivatization methods have been developed to distinguish a different fragment ion series.[29–32] Typically, samples are divided into two fractions. Peptides in one fraction are labeled at their N- or C-terminus while peptides in the other fraction are unmodified. As a result, only fragment ions that carry the label are shifted in mass between the two spectra and N- and C-terminal fragments can thus be distinguished. Unfortunately, splitting of the sample causes a loss in sensitivity. In another derivatization strategy, Keough and co-workers attached a negatively charged sulfonate group to peptide N-termini to enhance peptide ion fragmentation and suppress N-terminal fragments.[33] Although activated sulfonated peptides yield dominant y-type ions with negligible b-type ions, precursor intensities in the positive matrix-assisted laser desorption ionization (MALDI) ion mode tend to be low.[33] In an alternative approach, Zubarev and co-workers recorded complementary CID and electron-capture dissociation (ECD) MS/MS spectra with an LTQ-FT mass spectrometer. Peptides of interest were first fragmented by low-energy CID in the linear ion trap and then dissociated by ECD in the Fourier transform ion cyclotron resonance (FTICR) mass spectrometer.[34,35] In contrast with CID, ECD spectra are dominated by c- and $z\bullet$ type ions. Since the c-type ions are 17 Da heavier than the b-type ions while the $z\bullet$ type ions are 16 Da lighter than y-type ions, they provide so-called "golden complementary pairs".[34] On the basis of these distinctive mass spacings, b- and y-type ions in CID spectra can be distinguished. Combination of two complementary sets of data also increases sequence coverage of fragment ions. However, this technique is not applicable to MALDI instruments because ECD requires multiply charged precursor ions. A third issue for de novo sequencing is that amino acid assignments are not usually checked except by comparison with a protein sequence database. A better approach would be to employ multiple series of fragment ions. For example, observation of b- and y-type ion pairs adds confidence to amino acid assignments. However, low-energy CID rarely produces both b- and y-type ions at every amino acid and the two ion types are difficult to distinguish.

Differentiation of isobaric amino acids is another challenge for peptide de novo sequencing. Leucine and isoleucine are not distinguished by most de novo sequencing algorithms. This is not a trivial matter since these two amino acids account for 16.4% of all amino acids and their abundance is even higher in transmembrane proteins.[36] Differentiation of leucine and isoleucine requires side-chain fragmentation, which is induced by high-energy CID in a sector instrument or high-energy ECD in a FTICR mass spectrometer[36–38] but not by low-energy CID. Likewise, there are two other pairs of amino acids with similar masses: lysine (128.0950 u) and glutamine (128.0589 u) as well as phenylalanine (147.0684 u) and oxidized methionine (147.0354 u). Although these residues are readily distinguished by high-resolution FTICR mass spectrometers and usually in TOF instruments, they are

[24] Xu, C.; Ma, B. *Drug Discovery Today* **2006**, *11*, 595–600.

[25] Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390–4399.

[26] Bern, M.; Cai, Y. H.; Goldberg, D. *Anal. Chem.* **2007**, *79*, 1393–1400.

[27] Huang, Y.; Triscari, J. M.; Tseng, G. C.; Pasa-Tolic, L.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. *Anal. Chem.* **2005**, *77*, 5800–5813.

[28] Bringans, S.; Kendrick, T. S.; Lui, J.; Lipscombe, R. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 3450–3454.

[29] Munchbach, M.; Quadroni, M.; Miotto, G.; James, P. *Anal. Chem.* **2000**, *72*, 4047–4057.

[30] Gagney, G.; Emili, A. *Nat. Biotechnol.* **2002**, *20*, 163–170.

[31] Brancia, F. L.; Montgomery, H.; Tanaka, K.; Kumashiro, S. *Anal. Chem.* **2004**, *76*, 2748–2755.

[32] Beardsley, R. L.; Sharon, L. A.; Reilly, J. P. *Anal. Chem.* **2005**, *77*, 6300–6309.

[33] Keough, T.; Youngquist, R. S.; Lacey, M. P. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 7131–7136.

[34] Nielsen, M. L.; Savitski, M. M.; Zubarev, R. A. *Mol. Cell. Proteomics* **2005**, *4*, 835–845.

[35] Savitski, M. M.; Nielsen, M. L.; Kjeldsen, F.; Zubarev, R. A. *J. Proteome Res.* **2005**, *4*, 2348–2354.

[36] Kjeldsen, F.; Haselmann, K. F.; Sorensen, E. S.; Zubarev, R. A. *Anal. Chem.* **2003**, *75*, 1267–1274.

[37] Johnson, R. S.; Martin, S. A.; Biemann, K.; Stults, J. T.; Watson, J. T. *Anal. Chem.* **1987**, *59*, 2621–2625.

[38] Paizs, B.; Suhai, S. *Mass Spectrom. Rev.* **2005**, *24*, 508–548.

hardly separated by ion traps. Lysine can be guanidinated to distinguish it from glutamine.[32] However, differentiation of phenylalanine and oxidized methionine remains a challenge for low-resolution mass spectrometers.[28]

In our previous work, we have demonstrated that 157 nm photodissociation of singly charged peptides yields abundant high-energy fragments.[39−42] For peptides with a C-terminal arginine, a series of x-type ions along with numerous v- and w-type ions are the primary products. Since the x-type ions extend through the peptide backbone, they can be used to derive peptide sequences directly. On the basis of these unique characteristics, a de novo sequencing algorithm is developed in this work to automatically derive peptide sequences from photodissociation data. The algorithm combines photodissociation and PSD data to identify x/y ion pairs and derive peptide sequences. Observation of y-, v-, and w-type ions provides constraints to assess the confidence of each residue assignment in a computed sequence. Tryptic peptides from four model proteins were guanidinated, fragmented by photodissociation and PSD, and then sequenced. With reference to the protein sequences, the accuracy and amino acid coverage of the sequencing results were evaluated. Differentiation of isobaric leucine and isoleucine was also investigated. These derived sequences were subsequently matched against the SwissProt protein sequence database to investigate the capability for protein identification. To further test the robustness of this algorithm, a total of 266 tryptic peptides from 23 model proteins were analyzed. The derived sequences along with their confidence scores and rankings were imported into a database to evaluate sequencing accuracy and completeness as well as the scoring system.

EXPERIMENTAL SECTION

Materials. Human hemoglobin, horse myoglobin, horse cytochrome c, and bovine ubiquitin were purchased from Sigma (St. Louis, MO). Acetonitrile (ACN) and trifluoroacetic acid (TFA) were obtained from EMD Chemicals, Inc. (Gibbstown, NJ). α-Cyano-4-hydroxycinnamic acid (CHCA) were bought from Sigma (St. Louis, MO). *O*-Methylisourea was purchased from Acros Organics (NJ). Trypsin was obtained from Sigma (St. Louis, MO). Ammonium bicarbonate (ABC) was purchased from Sigma (St. Louis, MO).

Tryptic Digestion. Tryptic peptides from human hemoglobin, horse myoglobin, horse cytochrome c, and bovine ubiquitin were generated using bovine trypsin. Each protein was prepared in 25 mM ammonium biocarbonate to make a concentration of 100 μM. Tryptic digestion was performed by mixing 100 μL of each protein solution with 5 μg of lyophilized trypsin. The digestion was allowed to incubate at 37 °C overnight before being stored at −20 °C.

To build a library of peptides, a total of 23 proteins listed in Table S1 of the Supporting Information were digested with trypsin. For the 20 proteins with a limited number of disulfide bonds, tryptic digestion was performed following the procedure described above. For three proteins that contain many disulfide bonds, the employed digestion procedure was adopted from previous work by Reilly and co-workers,[43] in which disulfide bonds were first chemically reduced and cysteine residues were then alkylated before enzymatic digestion.

Peptide Guanidination. Tryptic peptides were guanidinated using *O*-methylisourea.[44] Guanidination reagent solution was made by dissolving 0.05 g of *O*-methylisourea in 51 μL of water. For each derivatization, 5 μL of peptide solution was mixed with 5.5 μL of ammonium hydroxide (7 N) and 1.5 μL of the guanidination reagent. The pH of reaction solutions was about 10.6. The reaction was incubated at 65 °C for 5−10 min before being terminated by adding 15 μL of 10% TFA (v/v). Reaction mixtures were dried by a speed vac before being stored at −20 °C.

Sample Preparation for MALDI Analysis. Guanidinated peptides were resuspended in water to make 10 μM solutions and were desalted by homemade microextraction zip-tip columns packed with C18-derivatized silica gel (Grace Vydac, Hesperia, CA) before MALDI analysis. In a typical experiment, 10 g/L CHCA in 49.95% ACN and 49.95% H_2O with 0.1% TFA was prepared as the matrix solution. Aliquots (1 μL) of peptide solutions were loaded onto zip-tip columns. After being washed by 0.1% TFA in water, peptides were released into 2 μL of matrix solution. MALDI spots were made by depositing 0.5 μL aliquots of the peptide−matrix mixture onto a plate. For each peptide, two MALDI spots were created.

Mass Spectrometry. Photodissociation and PSD were performed on an ABI 4700 TOF−TOF mass spectrometer (Applied Biosystems, Framingham, MA) as described previously.[42,45] In brief, photodissociation was implemented using an F_2 laser (CompexPro F_2, Coherent Lambda Physik, Germany). The laser was attached to the collision cell through a feed-through in the TOF−TOF main chamber. A computer program was developed to coordinate the photodissociation laser with the mass spectrometer. Peptide masses were first measured in the MS mode. Photodissociation timings were then calculated by the computer program. In the MS/MS mode, peptide ions of interest were isolated by a timed ion gate. When the ion packet arrived at the photodissociation spot, a programmable delay generator (BNC model 555, Berkeley Nucleonics Corporation, San Rafael, CA) was then used to trigger the laser based on the calculated timing. A 10 mJ, 10 ns pulse of light (19 mm high by 6 mm wide) was typically produced. The precursor ions as well as the PSD fragments were photoexcited. The resulting photofragments along with the remaining precursor ions and PSD fragments were then reaccelerated into the reflectron-TOF for mass analysis. Peptide PSD spectra were obtained with the photodissociation laser switched off. Currently this apparatus runs at 50 Hz because this is the maximum repetition rate of the F_2 laser.

All tryptic peptide mixtures from model proteins were first analyzed without fragmentation to obtain a list of precursor ion masses. The 10 most intense peaks in each mass spectrum were isolated for photodissociation and PSD. Each photodissociation

[39] Thompson, M. S.; Cui, W.; Reilly, J. P. *Angew. Chem., Int. Ed.* **2004**, *43*, 4791–4794.

[40] Cui, W.; Thompson, M. S.; Reilly, J. P. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1384–1398.

[41] Kim, T. Y.; Thompson, M. S.; Reilly, J. P. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 1657–1665.

[42] Zhang, L.; Reilly, J. P. *Anal. Chem.* **2009**, *81*, 7829–7838.

[43] Beardsley, R. L.; Sharon, L. A.; Reilly, J. P. *Anal. Chem.* **2005**, *77*, 6300–6309.

[44] Beardsley, R. L.; Karty, J. A.; Reilly, J. P. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 2147–2153.

[45] Zhang, L.; Reilly, J. P. *J. Proteome Res.* **2009**, *8*, 734–742.

spectrum was recorded by averaging 2000 MALDI shots at 50 Hz (about 40 s). A total of 2000 shot PSD spectra were recorded at 200 Hz with the same MALDI laser intensity that was used in the photodissociation experiments. All spectra were processed in Data Explorer version 3.0 (Applied Biosystems, Framingham, MA) and then plotted by Origin version 7.0 (OriginLab, Northampton, MA). Peptide fragment masses were predicted using Protein Prospector (http://prospector.ucsf.edu).

Preliminary Data Processing. All photodissociation and PSD spectra were smoothed and corrected to zero baseline using Data Explorer (Applied Biosystems, Framingham, MA). All monoisotopic peaks with a signal-to-noise ratio (S/N) greater than 15 were detected by the ABI 4000 Explorer software (Applied Biosystems, Framingham, MA). Within every 200 Da mass window, a maximum of 10 peaks were selected for output based on their intensities. In total, a maximum of 100 peaks were exported from each spectrum by the ABI 4000 Explorer software. Along with the precursor mass, a list of mass-to-charge ratios (m/z) and peak intensities of the selected fragments was stored in an ASCII file. Photodissociation and PSD peak lists were stored in two separate files. These text files were processed using in-house software programmed with Visual Basic (Microsoft, Seattle, WA) as follows. First, PSD fragments in photodissociation spectra were identified by alignment with PSD spectra. Peaks that occur in both were labeled as PSD fragments; otherwise, they were considered to be photofragments. Second, x-type ions are 25.98 Da heavier than the corresponding y-type ions. Thus any photofragments that were 25.98 Da heavier than PSD fragments were tentatively labeled as x-type ions, and the PSD fragments were labeled as y-type ions. These cooperative x/y pairs offer confirmatory information, and they are used to derive sequences as discussed below. Only y-type ions that appear in PSD spectra are used to generate x/y pairs. Third, although x_1 ions were sometimes not observed, y_1 ions were almost always detected. Therefore, if a 175.12 or 189.13 Da y_1 ion (corresponding to arginine or guanidinated lysine) was observed, the C-terminal residue was established.

De Novo Sequencing Algorithm. The algorithm uses x-type ions to derive peptide sequences. The computation routine includes two major steps: deriving tentative sequences and checking for errors. The algorithm first constructs a rough sequence using the putative x/y ion pairs identified above. When mass spacings between two x-type ions match the mass of an amino acid, the corresponding amino acid is tentatively assigned; otherwise, a gap is produced. To bridge two gapped x/y ion pairs, the algorithm then looks for possible x-type ions based on peak mass spacings. This routine starts from the smaller x-type ion in a gapped x/y ion pair and measures the mass spacing between it and all peaks within 186.5 Da (which is just above the mass of the heaviest amino acid). When a spacing matches the mass of an amino acid, the peak is labeled as a tentative x-type ion and the corresponding amino acid is tentatively assigned. When multiple x-type ions are found, each amino acid assignment leads to an individual sequence and multiple tentative sequences are generated. These assignments are subsequently checked by the spacing to the next x- or y-type ion. If the spacing matches the mass of an amino acid, the assignment is considered to be correct; otherwise, it is false and is removed. If no x-type ions are found, the algorithm looks for possible y-type ions based on peak mass spacings from y-type ions in the gapped x/y ion pairs. When neither an x- nor a y-type ion is found, the size of the mass gap is simply listed. After tentative residue assignments are completed, sequences are examined by observation of v-, w-, and y-type ions that are then used to check these assignments. Confidence levels for amino acids are determined based on the occurrence of constraint ions that are terminated by these residues. When an *x/y* ion pair is observed, the assignment is rewarded with the highest confidence score of 1.0; otherwise, assignments are evaluated by how many of the x-, y-, v-, and w-type ions are detected in photodissociation spectra relative to the number that could be formed. The average score of all amino acid assignments is then used to assess the credibility of a sequence. All tentative sequences are further ranked based on their confidence scores.

For peptide mixture analysis, peak lists from all photodissociation spectra are stored in a single file along with precursor mass information. PSD data are stored in a similar fashion. For each peptide, the de novo sequencing algorithm imports photodissociation and PSD peak lists along with precursor mass information to derive peptide sequences. The top 10 ranked sequences are usually output.

Homology Sequence Searching. In order to evaluate the sequencing accuracy and test the method's capability for identifying proteins, de novo sequencing results were matched against a database using the MS-homology searching program publicly available from UCSF Mass Spectrometry Facility (http://prospector.ucsf.edu). Although this search engine accepts gapped sequences, gaps must be small corresponding to fewer than 50 possible amino acid combinations. Therefore before searches, all sequences were checked by an in-house computer program. Complete sequences and those with a small gap were unchanged. However for sequences with a large gap, only the longest consecutive amino acid assignments were output. Sequence matching was done assuming no enzyme specificity was employed during sequence matching. A mass tolerance of ±0.3 Da was applied for gap mass values. In order to test whether each derived peptide sequence identified a unique protein, the Swiss-Prot.2008.06.10 database with 389 046 protein entries was used as the search library. Exact sequence matching was initially attempted. Unmatched sequences were subjected to another search allowing for two incorrect amino acid assignments.

RESULTS AND DISCUSSION

Combination of Photodissociation and PSD Data. It has been shown that photodissociation with 157 nm light yields abundant high-energy fragments that are strikingly different from those low-energy fragments generated with postsource decay.[40,42] To compare the fragments produced by these two techniques, Figure 1 displays a typical pair of photodissociation and PSD spectra. The PSD spectrum of guanidinated tryptic peptide MFLSFPTTK* (Figure 1A) is dominated by y-type ions. Ions that have lost ammonia are labeled with an asterisk (*). The photodissociation spectrum of the same peptide is displayed in Figure 1B. In addition to abundant high-energy x-, v-, and w-type photofragments, most of the PSD fragments are also observed. This is because PSD fragments are not separated from precursor ions by the timed ion gate in the tandem-TOF mass spectrometer. Some of these fragments (y_1, a_2, b_2, and b_3) are further enhanced during photodissociation. Postsource decay spectra can be used

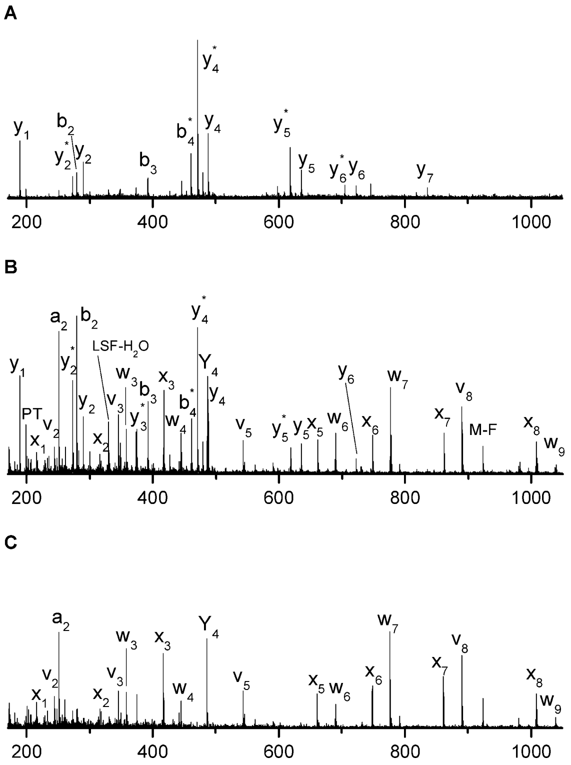


Figure 1. (A) PSD and (B) photodissociation spectra of peptide MFLSFPTTK after peptide guanidination. (C) Photodissociation spectrum of the same peptide after removal of PSD fragments. In all figures, an asterisk (*) denotes the loss of ammonia, and capital letters label internal fragment ions.

to identify peaks in a photodissociation spectrum that are not x-, v-, and w-type ions. When desired, PSD peaks can be removed from photodissociation spectra. Direct subtraction of spectra does not accomplish this since peaks in the two spectra differ in intensity. In this example, by magnifying the PSD spectrum in Figure 1A by 5.0 times, subtracting it from the photodissociation spectrum in Figure 1B, and not plotting negative peaks, the result shown in Figure 1C is obtained. It is evident that the photodissociation spectrum by itself contains primarily x-, v-, and w-type ions that are complementary to the b- and y-type fragments in the PSD spectrum. Since y-type ions are always 25.98 Da lighter than the corresponding x-type ions, x/y ion pairs can easily be identified by comparison of PSD and photodissociation spectra. The observed y-type ions can sometimes be used to identify complementary b- type ions since the sum of their masses equals ($M + 1$) Da where M is the precursor mass.

In addition to identifying x-type ions, y-type ions are used to assign amino acids when the corresponding x-type ions are missing. As noted above, y_1 ions identify peptide C-terminal residues since they tend to be abundant in spectra recorded with the ABI 4700 MALDI TOF−TOF mass spectrometer.[42] As a result, peptide C-terminal residues can still be unambiguously identified even when an x_1 peak is missing. y-type ions are very important for identifying proline residues. As demonstrated in previous work, x-type ions terminated by proline are not observed at all in photodissociation spectra; instead y-type ions at proline are always abundant.[42] For example in Figure 1C, an intense Y_4 ion appears while the x_4 ion is missing. Fortunately in most PSD spectra, y-type ions terminated by proline are abundant because of preferential cleavage of Xxx-Pro bonds.[27] Since the y-type ions are always 2 Da heavier than the corresponding Y-ions, one helps to confirm the other. Observation of such ion pairs along with the absence of corresponding x-type ions points to proline residues. Pro-Xxx bonds are not normally cleaved to produce y-type ions upon low-energy vibrational excitation.[27] For example in Figure 1A, the y_3 ion is missing. This information also helps to confirm proline assignments. The combination of two sets of sequence ions reduces the risk of incomplete sequence identifications when one of the sequence ions is low in intensity and is not detected by the peak picking software. This is rather important for x-type ions in the low- or high-mass region that often exhibit low intensities.[42] A good example is the x_2 ion Figure 1C. If it were not recognized, the de novo sequencing algorithm would still be able to identify the amino acid by using the y_2 ion in the PSD spectrum.

Error Checking Using v-, w-, and y-Type Ions. In addition to x-type ions, photodissociation spectra (i.e., Figure 1C) also contain many v- and w-type ions that result from amino acid side chain fragmentation. Since these fragments do not appear in low-energy PSD spectra (i.e., Figure 1A), they can easily be recognized by comparison of the two sets of data. It has been shown that side chain losses are widely observed in 157 nm photofragmentation spectra.[39,40,42,46] Since the observed masses depend on the residue side chain structures, they provide information about amino acid identities that complements the x- and y-ion mass differential data discussed above.

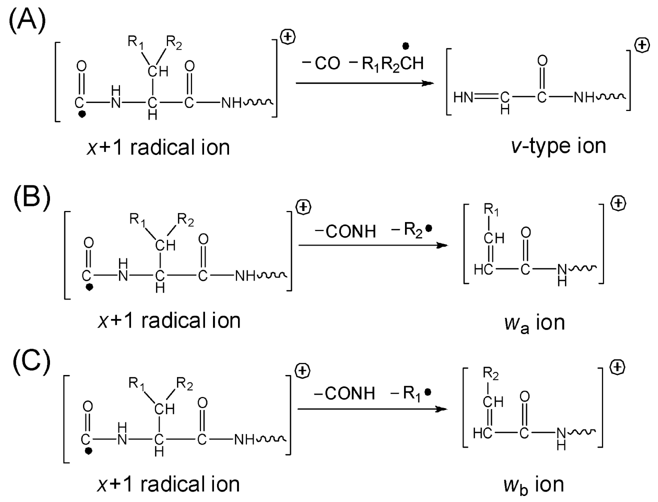
The production of v- or w-type fragments is residue-dependent. In Figure 1C, for example, abundant v_5 and v_8 ions are observed at phenylalanine while w_5 and w_6 are formed at serine and leucine, respectively. On the basis of more than 200 photodissociation spectra obtained previously[39,40,42,46] and in this experiment, the dependence of v- and w-type ions on amino acids is summarized in Table 1. For aromatic amino acids, only v-type ions are abundantly formed (Scheme 1A); w-type ions are not produced because cleavage of the side chain $C_\beta-C_\gamma$ bonds at aromatic residues is thermodynamically disfavored.[47] For most nonaromatic amino acids except valine, threonine, and isoleucine, only w-type ions appear in photodissociation peak lists while the v-type ions are usually low in intensity. This is because these nonaromatic amino acids do not contain any substituent on the β-carbon. Formation of v-type ions (Scheme 1A) is disfavored since it involves releasing an unstable secondary radical.[48] In addition, these v-type ions overlap with the ^{13}C component of the corresponding w-type ions. As a result, the v-type ions of most nonaromatic amino acids are not identified by the ABI Data Explorer peak picking software and are not exported in photodissociation peak lists. In contrast, isoleucine, threonine, and valine yield v- and w-type ions that are spaced by 13.02 Da. Production of v-type ions at these three residues is enhanced because it involves the release of a stable tertiary

[46] Reilly, J. P. *Mass Spectrom. Rev.* **2009**, *28*, 425–447.
[47] Zhang, L.; Cui, W.; Thompson, M. S.; Reilly, J. P. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1315–1321.
[48] Zhang, L.; Reilly, J. P. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1378–1390.

Table 1. Observation of v- and w-Type Ions at Different Amino Acids

		mass spacings from x-ion (Da)	
amino acids	monoisotopic masses (Da)	v-type ion	w-type ion
glycine (G)	57.02		
alanine (A)	71.04	42.01	
proline (P)	97.05		69.02
valine (V)	99.07	70.04	57.02
leucine (L)	113.08		85.05
isoleucine (I)	113.08	84.06	71.04
glutamine (Q)	128.06		100.03
glutamic acid (E)	129.04		101.01
serine (S)	87.03		59.00
threonine (T)	101.05	72.02	59.00
phenylalanine (F)	147.07	118.04	
tyrosine (Y)	163.06	134.04	
tryptophan (W)	186.08	157.05	
histidine (H)	137.06	108.03	
aspartic acid (D)	115.03		87.00
asparagine (N)	114.04		86.01
methionine (M)	131.04		103.01
cysteine (C)	103.01		74.98
lysine (K)	128.09		100.06
arginine (R)	156.10		

Scheme 1



radical.[48] The observed side chain fragments and their mass spacings from corresponding x-type ions are summarized in Table 1.

It is noteworthy that isoleucine and threonine have two different β-substitutents on their side chains and thus both can yield two different w-type ions (parts B and C of Scheme 1). However, only one of these is usually observed in photodissociation spectra. At isoleucine, loss of the ethyl group occurs while loss of the hydroxyl group is favored at threonine. The mass spacings of these two w-type ions from the corresponding x-type ions are listed in Table 1.

Observation of y-, w-, and v-type ions adds different degrees of confidence to amino acid assignments. y-type ions are the major fragments in PSD spectra, and they are always 25.98 Da lighter than the corresponding x-type ions. Detection of these ions in PSD spectra confirms most x-type ion assignments. As a result, amino acid assignments based on x/y ion pairs are given a confidence score of 1.0. It is noteworthy that y-type ions observed only in photodissociation spectra are not used to confirm the x-type ion assignments because photodissociation spectra contain many fragments and there is a non-negligible probability that some of these may happen to match the mass of a y-type ion. Nevertheless, observation of these y-type ions adds some confidence to amino acid assignments. Likewise, observation of v- and w-type ions does not confirm x-type ion assignments even though their mass spacings from the corresponding x-type ions are unique to each amino acid. This is because multiple amino acids lead to v- and w-type ions that have the same mass. For example, w-type ions terminated by all nonaromatic amino acids except valine, threonine, and isoleucine that are isobaric. Likewise, since v-type ions are formed by loss of a complete side chain, their masses are independent of the N-terminal residue. Nevertheless, observation of these ions adds some confidence to amino acid assignments that are not based on x/y ion pairs. In this algorithm, amino acid assignments based on x- or y-ions are checked by observation of all sequence ions including x-, y-, v-, and w-type fragments. Since the production of v- and w-type ions is residue-dependent, the number of expected sequence ions varies with amino acid. As a result, the ratio of the number of observed sequence ions over the expected number is arbitrarily used to define the confidence level of amino acid assignments based on x- or y-ions. Immonium ions are abundantly observed in photodissociation spectra,[42] and they could provide additional constraints on the presence of some amino acids. However, they are not included in the present algorithm.

The confidence of a proposed sequence is evaluated by the average score associated with all amino acid assignments. The overall scores not only reflect the credibility of a sequence but are also useful to rank sequence candidates when necessary. In addition, the algorithm displays all of the observed sequence-related ions for each amino acid assignment. These details facilitate manual checking of sequencing results, and they should be useful in future statistical evaluations of the residue-dependence of photofragmentation.

De Novo Sequencing. In order to demonstrate how the algorithm derives sequences and checks for errors, the step by step process as applied to the spectra in Figure 1 will be

Table 2. De Novo Sequencing of Tryptic Peptides from Four Model Proteins[a]

actual sequence	interpreted sequence	total residues	correct assignments	incorrect assignments	unassigned residues
		Hemoglobin, Human			
EFTPPVQAAYQK	EFTPPVQAAYQK	12	12	0	0
TYFPHFDLSHGSAQVK	[264.098]FPHFDLSHGSA[227.29]K	16	12	0	4
FFESFGDLSTPDAVMGNPK	[423.105]SFGDLSTPDAVMGNPK	19	16	0	3
VLGAFSDGLAHLDNLK	V[I|L]GAFSDGLAHLDN[I|L]K	16	16	0	0
VGAHAGEYGAEALER	VGAHAGEYGAEALER	15	15	0	0
MFLSFPTTK	MFLSFPTTK	9	9	0	0
LLVVYPWTQR	[I|L]LVVYPWTQR	10	10	0	0
VHLTPEEK	VHLTPEEK	8	8	0	0
SAVTALWGK	SAVTALWGK	9	9	0	0
VNVDEVGGEALGR	VNVDEVGGEALGR	13	13	0	0
		Myoglobin, Horse			
HGTVVLTALGGILK	HGTVVLTALGGILK	14	14	0	0
HGTVVLTALGGILKK	HGTVVLTALGGIL[170.22]K	15	14	0	1
LFTGHPETLEK	[I|L]FTGHPETLEK	11	11	0	0
HPGDFGADAQGAMTK	HPGDFGADAQGAMTK	15	15	0	0
VEADIAGHGQEVLIR	VEADIAGHGQEVL$\underline{L}$R	15	14	1	0
YLEFISDAIIHVLHSK	[405.145]FISDAIIHVLHSK	16	13	0	3
ALELFR	[184.184]ELFR	6	4	0	2
		Cytochrome c, Horse			
TGQAPGFTYTDANK	TGQAPGFTYTDANK	14	14	0	0
TGPNLHGLFGR	TGPNLHGLFGR	11	11	0	0
TGPNLHGLF	TGPN[567.36]	9	4	0	5
MIFAGIK	MIFAGIK	7	7	0	0
EDLIAYLK	EDLIAYLK	8	8	0	0
KYIPGTK	[170.151]YIPGTK	7	6	0	1
EETLMEYLENPK	EET$\underline{DE}$EYLENPK	12	10	2	0
		Ubiquitin, Bovine			
TLSDYNIQK	TLSDYNIQK	9	9	0	0
ESTLHLVLR	ESTLHLVLR	9	9	0	0
EGIPPDQQR	EGIPPDQQR	9	9	0	0
IQDKEGIPPDQQR	[526.263]EGIP[468.31]R	13	5	0	8
M*QIFVK [a]	[M*|F]QIFVK	6	6	0	0
MQIFVK	MQIFVK	6	6	0	0
TITLEVEPSDTIENVK	[315.145]LEVEPSDTIENVK	16	13	0	3
total residues		355	322	3	30
percentage			90.7	0.8	8.5

[a] The asterisk symbol (*) denotes oxidation of methionine. [X|Y] denotes a residue that can be either the X or Y residue. Incorrect amino acid assignments are underlined.

considered. First, peak lists from parts A and B of Figure 1 are compared, which identifies five x/y ion pairs: x_1/y_1, x_2/y_2, x_5/y_5, x_6/y_6, and x_7/y_7. The mass differentials between the x-type ions identify four amino acids and two gaps: [278.18 Da]LS[345.18 Da]TK*, where the asterisk (*) represents guanidinated lysine. Second, the algorithm tries to bridge the 345.18 Da gap by looking for possible x-type ions between the x_2 (316.16 Da) and x_5 (661.34 Da). To identify the x_3 fragment, the algorithm searches for peaks that are within 186.5 Da of x_2. The x_3 ion in Figure 1B is considered to be a candidate since its spacing from the x_2 ion matches that of threonine (101.05 Da). This assignment is further confirmed by observation of the v_3 and w_3 ions at masses consistent with the threonine residue. No other x_3 ion candidates are found. The algorithm subsequently looks for possible x_4 ions based on mass spacings from x_3, but none of the peaks in the spectrum matches. It then looks for possible y_4 ions based on mass spacings from the calculated y_3 ion at 391.26 Da ($x_3 - 25.98$ Da). The y_4 ion is found to be a candidate since its spacing from the putative y_3 ion matches that of proline (97.05 Da). This assignment is further confirmed by observation of a 2 Da lighter Y_4 ion and the absence of y_3 and x_4 ions from the spectrum. The algorithm then calculates the value that the x_4 ion should be ($y_4 + 25.98$ or 514.25 Da) and considers if this is consistent with the assignment of the following amino acid. The fifth residue is then determined by the mass spacing between x_5 and the calculated x_4 mass. The mass differential of 147.09 Da suggests that it can be either a phenylalanine or an oxidized methionine. An abundant v_5 ion at a mass appropriate for phenylalanine supports this assignment. (Differentiating these two amino acids is further discussed in the following section.) The gap of 345.18 Da is thus interpreted as FPT. Likewise, detection of an x_8 ion at 1008.52 Da suggests that the eighth residue from the peptide C-terminus is either a phenylalanine or an oxidized methionine since it is shifted from the x_7 ion by 147.08 Da. An abundant v_8 ion at a mass appropriate for phenylalanine supports this assignment. Since the residual mass of the N-terminal gap is 131.10 Da, the last residue of this sequence is assigned as methionine. The 278.18 Da gap is therefore assigned as MF. In summary, the entire peptide sequence is interpreted as MFLSFPTTK*.

After determination of a candidate sequence, the algorithm evaluates the confidence scores of all amino acid assignments. In the previous example, the four residues identified with x/y ion

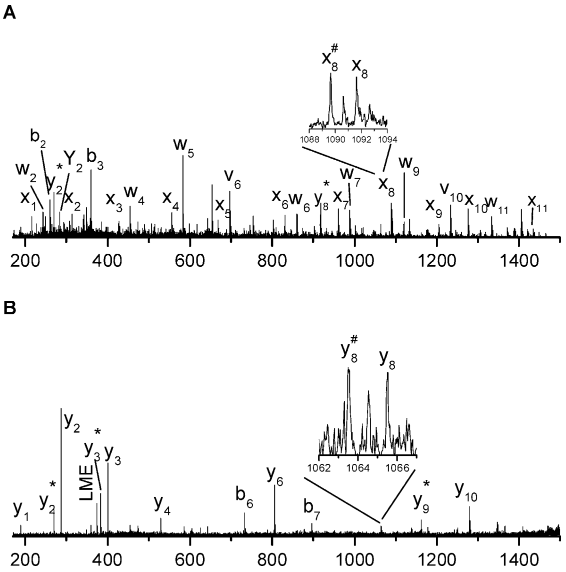


Figure 2. (A) Photodissociation and (B) PSD spectra of peptide EETLMEYLENPK after guanidination.

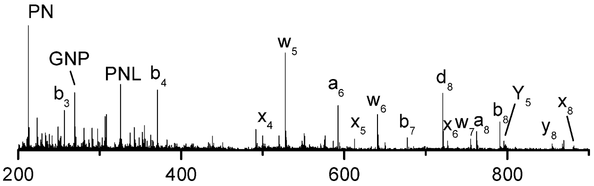


Figure 3. Photodissociation spectrum of peptide TGPNLHGLF after removal of PSD fragments.

pairs are assigned confidences of 1.0 as indicated above while other assignments are evaluated based on the occurrence of x-, v-, w-, and y-type fragments in the photodissociation spectrum. For example, at the eighth residue from the peptide C-terminus, two sequence ions (x_8 and v_8) are observed. Since three sequence ions are actually expected for phenylalanine (x_8, y_8, and v_8), the confidence score of this assignment is assigned as 0.67. With the average of the confidence scores of all amino acid assignments, the overall confidence score of the sequence is 0.906. These confidence scores are used to rank sequence candidates when multiple sequences are derived from photodissociation data.

Analysis of Tryptic Peptides from Model Proteins. Tryptic peptides are guanidinated prior to photodissociation for three reasons. First, lysine-terminated peptides yield higher precursor ion intensities since guanidination increases their gas-phase basicity.[44] Second, guanidinated peptides yield more abundant high-energy photofragments than unmodified lysine-terminated peptides because the guanidino group sequesters the charging proton.[40,42] Third, as noted before, guanidination makes it easier to differentiate lysine and glutamine. Lysine (128.095 u) and glutamine (128.059 Da) cannot be distinguished by low-resolution mass spectrometers, but guanidination converts lysine to homoarginine (170.13 Da). Although homoarginine has a mass close to several amino acid combinations (i.e., AV, IG, and LG), this rarely induces ambiguities during sequence interpretation because for tryptic peptides it is primarily located at the peptide C-termini.

To evaluate the accuracy of this sequencing approach, peptides from human hemoglobin, horse myoglobin, horse cytochrome c, and bovine ubiquitin were fragmented by photodissociation and PSD and then sequenced. Results are summarized in Table 2. A total of 29 out of 31 yielded sequences with 5 or more residues. Comparison with the protein database showed that 322 out of 355 (90.7%) amino acids were correctly identified. Of the remaining 33 amino acids, 28 were not identified at all and there were 3 incorrect amino acid assignments. Two of the incorrect assignments appeared in peptide EETLMEYLENPK from horse cytochrome c, in which LM was incorrectly interpreted as DE as shown in Table 2. After the peptide was identified by referring to the protein sequence, its photodissociation spectrum was further interpreted. PSD fragments were removed and the spectrum replotted as displayed in Figure 2A. Similar to Figure 1C, it is dominated by x-, v-, and w-type ions. During de novo sequencing, two candidate x_8 ions were found (labeled as $x_8{}^{\#}$ and x_8) that are terminated by glutamate and methionine, respectively. Remarkably, both of the corresponding y-type ions ($y_8{}^{\#}$ and y_8) also appeared in the PSD spectrum (Figure 2B), and this led to the incorrect assignment.

Only four residues of peptide TGPNLHGLF from horse cytochrome c were properly sequenced, but it still could be identified by referring to the protein sequence. This peptide was generated as a side product of the chymotryptic digestion of larger tryptic peptides. As seen in Figure 3, the sequencing gap is induced because x_1-x_3 are not detected. This results because the most basic residue in this sequence, histidine, preferentially binds the ionizing proton[49] and thus cleavage of backbone bonds to its C-terminal side does not yield x-type ions.[40,42] Instead, N-terminal a_6, b_7, a_8, and b_8 ions are produced. However, they are not used to derive sequences by the present algorithm. To enable more residues from nontryptic peptides to be identified, the sequencing algorithm should interpret both N- and C- terminal fragments. This will also facilitate identification of tryptic peptides with missed cleavages that lead to abundant N-terminal fragments.[42] It is noteworthy that a few proline-terminated internal fragments also appear in Figure 3. Presumably internal fragment formation involves charge-directed fragmentation processes that are enhanced at proline.[27]

Compared with conventional de novo analyses of low-energy CID or TOF−TOF CID data,[28] the present method generates rather few incorrect amino acid assignments. This is directly attributable to the use of constraint ions to check each amino acid assignment. Tentative amino acid assignments with low confidence are not labeled with their identities but by mass gaps. In contrast, most conventional algorithms generate complete peptide sequences that contain multiple incorrect amino acid assignments.[28] Without reference to a protein database, these false assignments are difficult to recognize. This poses a challenge for protein identification using sequence homology searches in which multiple amino acid errors would need to be allowed, thereby increasing the search times and numbers of false identifications. To avoid incorrect amino acid assignments, Pevzner and coworkers proposed an alternative "spectral profile" approach to

[49] Harrison, A. G. *Mass Spectrom. Rev.* **1997**, *16*, 201–217.

Table 3. Differentiation of Leucine and Isoleucine in Four Model Proteins[a]

proteins	observed Xle	identified Xle	differentiated Xle	Incorrect Assignments
hemoglobin	12	12	9	0
myoglobin	20	18	16	1
cytochrome c	12	9	9	0
ubiquitin	13	11	11	0
total residues	57	50	45	1
percentage of distinguishing Xle (%)			90.0	2.0

[a] Xle denotes an amino acid that can be either leucine or isoleucine.

represent de novo sequencing results.[50] In their method, only highly confident amino acid assignments are identified; others are represented by mass gaps. As exemplified by some of the above discussion, this strategy is adopted by the present sequencing approach. A second advantage of the photodissociation/de novo sequencing method involves sequence coverage. With this four protein sample set, over 90% sequence coverage was achieved. By comparison, PSD spectra contain far fewer mass spectral peaks. Without additional information, it is difficult to directly assign these peaks, and even once they are assigned, their paucity typically limits sequence identification to 50−60% of residues. The latter is consistent with Bringans and co-workers' assessment of three commercial de novo sequencing software packages in which 66% or fewer of the amino acids in peptide sequences were correctly identified in the analysis of tryptic peptides from model proteins using low-energy CID or TOF−TOF CID.[28] In their experiment, lower peptide quantities (50 fmol) were analyzed by an ABI 4800 TOF−TOF mass spectrometer that is generally considered to be 10 times as sensitive as the ABI 4700 TOF−TOF mass spectrometer employed in this study. Thus, although the comparison is not exactly rigorous, the two experiments should be comparable. Much higher sequence coverage is achieved in the present work because of the wealth of information in 157 nm photodissociation spectra.[39,40,42,46]

It is noteworthy that the N-terminal two amino acids in peptide ALELFR from horse myoglobin were not identified. This is because the x_5 ion was not detected in the photodissociation spectrum obtained by the ABI 4700 TOF−TOF mass spectrometer due to the instrument's low sensitivity in the high mass region. However, this fragment ion was intense feature in previous photodissociation TOF−TOF spectra recorded with a home-built instrument.[39] This suggests that additional information should be extractable with improved instrument tuning, particularly in the low- and high- mass regions.

Differentiating Leucine and Isoleucine. Leucine and isoleucine can be distinguished based on their unique side chain fragments produced by 157 nm photodissociation as shown in Table 1. Leucine primarily leads to w-type ions that are 85.05 Da lighter than the corresponding x-type ions. Isoleucine yields both v- and w-type ions that are 84.06 and 71.04 Da lighter than the corresponding x-type ions, respectively. Since the two w-type ions are spaced by 14.01 Da, they have been used to distinguish the two isomers in high-energy fragmentation experiments.[36,37,51]

The present de novo approach uses both w- and v-type ions to distinguish leucine and isoleucine as follows. The algorithm first looks for the w-type ions that are either 71.04 or 85.05 Da lighter than the x-type ion. When only the former peak is observed, the amino acid is assigned as isoleucine; the corresponding v-type ion that is 84.06 Da lighter than the x-type ion is then sought to confirm this assignment. When only the latter peak is detected, leucine is identified.

Results for distinguishing leucine and isoleucine in the four model proteins are summarized in Table 3. In total, 57 leucine or isoleucine (Xle) residues appeared in the 31 observed peptides. A total of 45 of these were correctly differentiated, 4 were identified as one or the other of these and 7 were not identified because the corresponding x-type ions were not detected in the photodissociation spectra. One isoleucine, in peptide VEADIAGHGQEVLIR, was mistakenly assigned as leucine. Two of the four indistinguishable residues were located at peptide N-termini that did not yield side chain fragments while the other two were near peptide termini where side chain fragments are often not detected as noted above. The error in the isoleucine assignment resulted because neither w_2 nor v_2 were observed in the photodissociation spectrum due to low sensitivity in the low-mass region. However, in this particular spectrum, an abundant b_2 ion at 229.14 Da was misassigned as a w_2 ion terminated by leucine (expected at 229.13 Da). Thus 98.0% of the identified leucines and isoleucines were correctly distinguished, but fortuitous errors such as this one may be difficult to completely eliminate. Observation of both w- and v-type ions appears to improve the accuracy of distinguishing leucine and isoleucine.

Differentiating Phenylalanine and Oxidized Methionine. Phenylalanine (147.068 Da) and oxidized methionine (147.040 Da) are a challenge to distinguish in a MALDI TOF−TOF instrument. However, they can be easily identified by their side chain fragments in photodissociation experiments. As displayed in Table 1, methionine primarily leads to w-type ions that are shifted from the corresponding x-type ions by 103.01 Da. Oxidized methionine yields w-type ions that are 119.00 Da lighter than the corresponding x-type ions. In contrast, phenylalanine leads to abundant v-type ions that are 118.04 Da lighter than the x-type ion. Since these two side chain fragments differ by 0.96 Da, they can easily distinguish phenylalanine and oxidized methionine.

As shown in Table 2, all phenylalanine residues were correctly identified based on observation of v-type ions. This is consistent with the fact that v-type ions terminated by aromatic amino acids are usually abundant in photodissociation spectra.[40,42,52] Only one oxidized methionine residue was in this sample set in peptide

(50) Kim, S.; Bandeira, N.; Pevzner, P. A. *Mol. Cell. Proteomics* **2009**, *8*, 1391–1400.

(51) Fernandez-de-Cossio, J.; Gonzalez, J.; Betancourt, L.; Besada, V.; Padron, G.; Shimonishi, Y.; Takao, T. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 1867–1878.

M*QIFVK from bovine ubiquitin. Because it resides at the peptide N-terminus where side chain fragments are often not generated, it was not possible to distinguish between the two assignments. However, the native, unoxidized form of this peptide was also in the sample set and in this case the methionine was identified, as noted in the table.

Protein Identification by Homology Searches. The ability to identify proteins using peptide de novo sequencing was investigated by matching the sequencing results against a protein sequence database. All of the interpreted sequences in Table 2 were submitted to the MS-Homology program to match against the SwissProt.2008.06.10 database containing 389 046 proteins. A total of 26 of the 31 sequences identified a unique peptide in the database. Since these peptides can be generated from protein homologues in different organisms, each of them typically matched multiple proteins in the database. When the organism under study was further used to constrain the matching results, each of the 26 sequences matched a unique protein.

The remaining five sequences did not lead to identification of a unique protein for two reasons. In three cases, [184.184]ELFR from horse myoglogbin, TGPN from horse cytochrome c, and EGIP from bovine ubiquitin, the short sequences matched numerous peptides from nonhomologous proteins in the database. (Mass gaps in the latter two sequences were not included in the search because they were too large to be accepted by the searching algorithm.) To address these ambiguities, another search was performed with a smaller database that contained only equine or bovine proteins. The first two sequences each matched a unique equine peptide that translated into correct protein identifications while the third matched several peptides from different bovine proteins. Although the nontryptic peptide TGPNLHGLF yielded a four-residue sequence that led to an ambiguous protein identification, all residues of the corresponding tryptic peptide TGPNLHGLFGR were assigned and this translated to a unique protein identification. A second reason for unsuccessful protein identification is that two sequences, EETDEEYLENPK and VEADIAGHGQEVLLR did not match with any peptide in the SwissProt database because of the one or two incorrect amino acid assignments noted in Table 2. However, a sequence homology search allowing up to two errors successfully matched unique proteins in the database.

Assessing the Scoring Approach. To more thoroughly evaluate the present sequencing approach, the 23 proteins listed in Table S1 in the Supporting Information were each digested separately. The resulting tryptic peptides from each protein were guanidinated and then purified by microextraction tips as described in the Supporting Information. Each tryptic digest was deposited to create one or two MALDI spots and each spot contained 2.5 pmol of materials. A total of 266 tryptic peptides were isolated, fragmented by photodissociation and PSD, and then sequenced.

A total of 167 of the 266 peptides whose precursor ions were fragmented led to sequences with 5 or more residues. This corresponds to 62.8% of the photodissociation spectra. The longest sequences for the 167 peptides are listed in Table S2 in the Supporting Information. Of the 167 peptides, only 5 sequences contain one or two incorrect amino acid assignments, corresponding to a false identification rate of 3.0%. The other 99 peptide ions that were photofragmented yielded limited sequence information for two principal reasons. First, many peptides contained more than one arginine or homoarginine due to a missed tryptic cleavage and these typically led to incomplete x-type ion series.[42] For example, peptide IQDKEGIPPDQQR from bovine ubiquitin yielded a sequence of only four consecutive amino acids as noted in Table 2. A second problem for sequencing was that some peptide precursor ions were low in intensity. This resulted because 10 precursor ions were isolated for fragmentation from each spot, and the last few of these often had low abundances since the spot was significantly depleted after thousands of MALDI shots.

The top ranked sequences for the 167 peptides (which were not always the same as the longest sequences listed in Table S2 in the Supporting Information), were used in homology searches against a database in order to assess their accuracy. A total of 163 (97.6%) of these sequences contained all correct amino acid assignments. A total of 68 (or 40.7%) of them were shorter than the longest sequences by one or more amino acids. This resulted because long sequences usually had fewer confirmations than short sequences, leading to lower scores. However, the top ranked sequences were still able to identify unique proteins since they were accurate and usually contained five and more residues. In addition, gaps in these sequences could be used as additional constraints during sequence matching as pointed out by Pevzner and co-workers.[50]

Completeness of the Derived Sequences. Although the present approach yields high peptide sequence coverage, it still leads to numerous gapped sequences as shown in Table 2 and Table S2 in the Supporting Information. Of the 167 successfully sequenced peptides, 96 (or 57.6%) yielded complete sequences. Of the remaining 71 peptides, 62 had an N-terminal mass gap and 30 peptides had a C-terminal mass gap. A total of 22 peptides had both N- and C-terminal gaps. However, only one sequence contained a gap in the middle. This gap distribution is consistent with the mass spectrometer's sensitivity, which is highest in the middle of the mass range and drops toward each end.[42] Accordingly, most x-type ions in the middle mass region are detected while some of them are missing in the low- or high-mass regions. Although gaps do not reveal any sequence information, they are still useful to protein identification by providing additional constraints as noted above.

CONCLUSIONS

A de novo sequencing algorithm was developed to interpret 157 nm photodissociation spectra. By combination of photodissociation and PSD data, this algorithm identifies x/y ion pairs and derives peptide sequences. Observation of y-, w-, and v-type ions provides additional constraints to amino acid assignments. In the analysis of 31 tryptic peptides from 4 model proteins, 322 (or 90.7%) of the amino acids are correctly identified, which is excellent by comparison with conventional de novo methods that interpret low-energy fragmentation data. Of the remaining 33 amino acids, 30 of them are not identified at all and there are three mistakes. The present de novo sequencing approach also allows leucine and isoleucine to be

[52] Zhang, L.; Reilly, J. P. *Proceedings of the 55th ASMS Conference on Mass Spectrometry and Applied Topics*, Indianapolis, IN, June 3–7, 2007.

differentiated using side chain v- and w-type fragments. A total of 45 of the 50 identified leucine and isoleucine residues in the data set are successfully distinguished. Of the remaining five residues, four are not distinguished because of undetected v- and w-type ions and there is only one mistake. These derived sequences are shown to be capable of identifying proteins from a large database using homology searches. In the analysis of a larger data set containing 266 tryptic peptides, 167 (or 62.8%) led to sequences with five or more identified amino acids. Only five sequences contained one or two false amino acid assignments, which is about 3.0% of the total identified sequences.

ACKNOWLEDGMENT

This work was supported by National Science Foundation Research Grant CHE-0518234.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review September 11, 2009. Accepted December 7, 2009.

AC902050Y