Published in final edited form as:

Methods Mol Biol. 2013 ; 1007: 173–182. doi:10.1007/978-1-62703-392-3_7.

Improving Peptide Identification Using Empirical Scoring Systems

Robert J. Chalkley

Department of Pharmaceutical Chemistry, University of California San Francisco

Summary

Peptides and proteins are routinely identified from peptide fragmentation spectra acquired in a mass spectrometer, analyzed by database search engines. The types of fragments that can be formed are known, and it is also well appreciated that certain fragment types are more common or more informative than others. However, most search engines do not use detailed knowledge of peptide fragmentation, but rather consider a limited range of fragments, giving each an equivalent weighting in their scoring system that decides which results are likely to be correct. This chapter will discuss efforts to make use of information about the frequency of observation of different fragment ion types in order to produce more sophisticated and sensitive scoring systems and will demonstrate how these new scoring systems are particularly powerful for analysis of electron capture or electron transfer dissociation data.

Keywords

Database search engine; collision-induced dissociation; electron transfer dissociation; statistical analysis of fragmentation; charge-state dependent scoring; sequence dependent scoring

1. Introduction

The ability to identify proteins at high sensitivity using mass spectrometry, whether isolated from a gel spot or band, or as part of a complex mixture of proteins and peptides, has transformed many biomedical research fields. These developments have been driven by major improvements in mass spectrometry instrumentation over the last twenty years that have produced machines that are significantly more sensitive, have improved mass accuracy and are at least an order of magnitude faster at acquiring data. Concomitant with these instrument developments analysis tools have to be developed to effectively mine the information from the data acquired. The large-scale sequencing of genomes, started in the early 1990s, dramatically simplified the protein identification challenge, as data could be compared to reference protein sequences rather than considering all amino acid permutations. The first approaches for protein identification by database searching were by an approach referred to as peptide mass fingerprinting, which only used peptide masses[1]. However, it quickly became clear that if there was a mixture of proteins in a sample, then fragmentation data was additionally required. Hence, tools for analyzing peptide

fragmentation data have become the mainstay of proteomic data analysis. The first tool to perform this task was SEQUEST[2], but there are now more than twenty tools regularly in use that perform this type of analysis[3].

There are three types of strategies that are employed by these programs for identifying peptides. The most common approach is for the software to create a theoretical prediction of what fragment masses would be expected to be observed in a spectrum of a particular peptide, then compare this list to an observed spectrum. A second approach is to try to interpret an amino acid stretch (tag) from an observed spectrum, and then use this, often in combination with theoretical prediction of other fragments that would be expected, in order to identify the peptide. The third approach, which is not currently used as much, but will become increasingly popular in the near future, is to compare an observed spectrum to a library of previously identified peptides. This approach does not require any prediction of what the spectrum should look like and can also make optimal use of fragment ion intensity, so should be able to provide the highest reliability and sensitivity, although its limitation is the need for the peptide to be present in the library of peptides queried. The rest of this chapter is going to focus on strategies to improve the first of these three approaches; i.e. the theoretical prediction of what a peptide fragmentation spectrum should look like and how to use this information to increase peptide identification sensitivity and reliability.

The sequence of the peptide is obviously a major factor in determining the appearance of the fragmentation spectrum. In addition to the peptide's sequence, the fragmentation approach is equally as important. The most commonly used fragmentation method for peptides has been collision induced/activated dissociation (CID or CAD). This approach collides analyte ions with neutrally-charged gas molecules. The collision imparts internal energy into the analyte, causing the weakest bonds to be broken. There are two variants of CID that are heavily employed. In an ion trap the analyte is resonantly excited; i.e. a particular m/z is selectively given kinetic energy to cause collisions. In this approach, once the analyte has fragmented, then the products are no longer excited. Hence, most fragments observed in ion trap CID spectra are the product of a single bond cleavage. If the CID is employed in a multipole (quadrupole, hexapole, etc.) collision cell then both precursor and product ions are given kinetic energy, so products of multiple bond cleavages can also be observed. More recently, two other related fragmentation approaches have been developed; electron capture dissociation (ECD)[4] and electron transfer dissociation (ETD)[5]. Both of these approaches work by causing analytes to gain an electron (either directly from an electron beam or via a reaction with an anion). This creates an unstable free radical analyte that fragments at locations that are driven by the site of electron capture. CID and ECD/ETD produce different backbone cleavages. Using the nomenclature for describing peptide fragment ions[6], CID cleaves mainly the peptide link to produce 'b' and 'y' ions, plus neutral loss derivatives of these ions[7], whereas ECD/ETD cleaves at a different location in the peptide backbone to produce c and z• radical ions[8].

Another factor that influences the appearance of a fragmentation spectrum is the charge state of the precursor ion. The fragmentation mechanisms in CID are heavily studied and have been reviewed elsewhere[7]. One of the fragmentation mechanisms that occurs in CID requires what is referred to as a mobile proton, where if the precursor ion has more charges

than basic residues then one of the protons, rather than being sequestered to a basic residue, can migrate along the length of the peptide and initiate backbone cleavage[9]. Electrospray ionization (ESI) of tryptic peptides typically produces doubly charged precursor ions containing a single basic residue. Hence, in these peptides the mobile proton fragmentation mechanism can dominate. However, if the charge does not exceed the number of basic residues; for example if there is a histidine or a missed tryptic cleavage in the peptide sequence, or if the ion was created by matrix-assisted laser desorption ionization (MALDI), which produces singly charged ions, then this fragmentation mechanism will not occur and charge-remote fragmentation will dominate[7]. Large-scale empirical studies of sequence-dependent fragmentation have shown that for fragmentation spectra of precursors where there is no mobile proton, the fragmentation has different sequence-dependence[10, 11]. Despite this, to the best of the author's knowledge, no search engine tries to score fragmentation spectra of precursors with no mobile proton differently based on the different fragmentation characteristics. Potential reasons for this anomaly will be mentioned later.

2. Methods

2.1 Improving ECD/ETD MS Scoring using Statistical Analysis of Spectra

There is also a significant effect of the precursor charge-state on the fragmentation observed in ECD and ETD MS data. These fragmentation approaches do not impart internal energy into the analyte. Hence, if a doubly-charged precursor ion captures an electron, there is now only one charge present, so once fragmentation occurs there is no internal energy or charge repulsion to dissociate the fragments. Hence, they can remain associated together by non-covalent interactions and are not observed as separate fragments in the mass spectrum. For higher charge-state precursors this is less of an issue, probably because the presence of a charge on both fragments helps them to dissociate. To overcome this dissociation issue, a small amount of internal energy is typically added (often referred to as supplemental activation)[12]. This allows the non-covalently attached fragment ions to dissociate, providing fragmentation spectra with many more peaks than otherwise observed. However, closer analysis of these spectra leads to the observation of ions one mass unit higher than an expected z^+ ion and one mass unit lower than an expected c ion. It is believed that during the extended association of the two fragment ions prior to dissociation by supplemental activation a hydrogen transfer reaction occurs from the c to radical z^+ ion. Hence, these $c-1$ and $z+1$ ions are commonly observed in spectra of doubly-charged precursor ions. In addition, the introduction of internal energy into the ions can also cause the increased formation of fragment ions formed by the CID process. A final difference between ETD spectra of doubly-charged precursors and those of a higher charge state is that after electron capture there is only a single charge left and in a tryptic peptide this is going to be localized on the C-terminal basic residue. Hence, for tryptic peptides the spectrum is dominated by fragments containing the C-terminus of the peptide (as mass spectrometry only detects charged components). This is illustrated in Figure 1, which shows ETD spectra of doubly-charged and triply-charged precursors of the same peptide. The spectrum of the doubly-charged precursor contains a series of ten consecutive $z+1$ fragment ions, but only one N-terminally derived fragment ion: the $c-1_{12}$ ion. In contrast, the spectrum of the triply-

charged precursor contains nine c ions in addition to the z ions that were also observed from the doubly-charged precursor.

An in-depth statistical analysis of ion types observed in ETD spectra was performed[13]. Peptides with different characteristics were created by digestion of proteins with a selection of enzymes: LysC and trypsin created peptides with a basic C-terminal residue, LysN created peptides with a basic N-terminal residue and chemical cleavage with cyanogen bromide created peptides with a random distribution of basic residue locations within the peptide products. Fragmentation spectra results were separated on the basis of precursor charge state (2+ compared to 3+ and higher) and enzyme used to create the peptide. A summary of the results analyzing the frequency of different fragment ion types is shown in Table 1. Doubly charged tryptic peptides contained three times as many C-terminally derived fragment ions as N-terminally derived ions. However, for precursors that were triply-charged or higher there was a roughly equal balance of N-terminal and C-terminal fragments. There were also twice as many hydrogen transfer products (c-1 and z+1 ions) in spectra of doubly-charged precursor ions compared to precursors of a higher charge state.

When database search engines were adapted from analyzing CID to ETD/ECD, most programs simply switched from searching for b and y ions to c and z• ions. It was quickly recognized that search engines that did this were performing poorly with spectra of doubly charged precursors. In a study comparing the performance of several search engines at analyzing ETD data it was shown that the number of doubly-charged precursor ion spectra identified was extremely varied, with some search engines identifying very few spectra of this charge state[14]. Based on the results of the statistical study of ETD fragmentation data, it is not surprising that without considering z+1 and c-1 ions performance would suffer.

To investigate the potential improvement that could be made by utilizing these fragmentation statistics, a scoring system that gives varied weighting for the different fragment ion types depending on the precursor charge state and also the presence of a basic residue at N-terminus, C-terminus, both termini or neither termini was developed in the search engine Protein Prospector[15]. The performance of this scoring was compared to the previous generation of Protein Prospector that also used a weighted fragment ion scoring, but this was based on average occurrence of the different ion types without separating by charge and sequence. The comparison was made by analyzing both the data used to create the fragment ion statistic results[13], but also querying the same dataset that was used in the previously mentioned ETD search engine comparison[14].

For the data produced by different enzyme cleavages the relative improvement in identifications at a fixed false discovery rate of 1% varied from 15 % for Lys-C data to 39 % for Lys-N data[15]. The larger improvement for Lys-N data was due to the original scoring system being heavily biased toward peptides with a basic C-terminal residue. When comparing the Protein Prospector performance to the other search engines the new scoring system tripled the number of IDs compared to the worst performing search engine and nearly doubled the results for the most successful competing tool compared to. Figure 2 shows these results broken down by charge state of identified spectrum. As can be seen, OMSSA identified a vanishingly small number of spectra that were doubly-charged and X!

Tandem also struggled with these. The new scoring system in Protein Prospector (v5.4) led to a 69 % increase in identification of doubly-charged spectra compared to the previous generation (v5.3).

Another group subsequently performed a similar statistical analysis of ETD data for developing a scoring system in pFind[16]. In their study they made use of high-resolution ETD data that allowed easier differentiation between c and z• ions and their hydrogen transfer equivalents (c-1 and z+1), but their overall conclusions were the same, and the charge-dependent scoring that they employed led to 63-122 % more IDs in comparison to Mascot[17], which is a similar level of improvement to that reported using Protein Prospector[15].

2.2 Improving CID MS Scoring Using Statistical Analysis of Spectra

Given the dramatic improvement these two software report for ETD data, one would expect statistical analysis of CID MS data would have led to improved scoring systems for this type of data as well. There have been many statistical analyses of fragment ions observed in CID MS data [10, 18, 19] and even studies that analyze cleavage between different amino acid combinations[11], so there is plenty of published information that could be utilized. However, one limitation of many search engine scoring systems is that their probability/reliability calculation requires giving equal weighting to all peak matches, as they simply count the number of peaks matched and calculate a probability of this happening at random e.g. Mascot, OMSSA. This precludes their ability to utilize this information. As soon as you start varying the weighting for different fragment ion types it becomes more difficult to predict the meaning/relevance of a given score. The only practical strategy to address this is to observe the distribution of scores for incorrect answers, model them to a distribution, and then calculate the probability of a given score being part of this distribution[20].

With the exception of a lack of cleavage N-terminal to a proline, all other cleavages between amino acids occur at fairly similar frequency in ECD/ETD data[21]. In contrast, N-terminal to a proline is the most common cleavage in CID spectra with a mobile proton, whereas C-terminal to proline and glycine are usually of low intensity. In spectra with no mobile proton cleavage C-terminal to acidic residues (aspartic and glutamic acid) dominates, although cleavage N-terminal to proline is also more intense than average[10]. Hence, sequence-dependent scoring is likely to lead to the greatest benefit for CID data.

Ion fragmentation statistical information has been used to write software that predicts the appearance of CID fragmentation spectra[22, 23]. This software works reasonably well. However, it is far too slow to be utilized by a search engine on-the-fly. This brings up one of the major reasons why sequence dependent scoring for CID data has not taken over: with the large amount of spectra acquired on modern mass spectrometers, search engine speed has become a major deciding factor in the popularity of a search engine. Any search engine that is unable to analyze data in less time than it took to acquire is not going to be practical. For this reason, most of the use of sequence or peak intensity is in software for re-scoring top matches[24].

Probably the only sequence-dependent scoring algorithm for CID data that has been described is SQID[25]. This software includes a scoring component that gives weighting based on the likelihood of the amino acid cleavages formed by the most intense four or five peaks, using a training dataset to produce these statistics. The authors observed 5-25 % increased identifications compared to Sequest[2], depending on the dataset. However, as there were other differences in their scoring in addition to the sequence-dependent aspect, it is impossible to tell the specific benefit of this component. It is also unclear from the article what effect the sequence-dependent scoring had upon search speed.

3. Conclusions

By using results from statistical analysis of fragmentation spectra it is possible to develop improved scoring systems for database search engines. This can lead to dramatic improvements in performance for ETD/ECD spectra, particularly the implementation of a charge-state and sequence-dependent scoring system, which make a large difference for analysis of spectra from doubly-charged precursors. This type of scoring should also be beneficial for CID spectra. However, for CID spectra the charge compared to the number of basic residues, rather than charge *per se* should be the key parameter. Sequence dependent scoring has also been preliminarily investigated, but it is not yet clear how valuable this will be.

Reasons why these types of scoring are not more common include the challenge of modeling the scoring to determine a measure of reliability for a given score and the fact that if you create a complex scoring system the algorithm is likely to be relatively slow, and with the large datasets now being typically analyzed, speed is an important parameter.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences, NIH NIGMS P41GM103481, from the National Institutes of Health, and the Vincent J. Coates Foundation.

References

1. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A*. 1993; 90(11):5011–5. [PubMed: 8506346]
2. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*. 1994; 5:976–989. [PubMed: 24226387]
3. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics*. 2010; 73(11):2092–123. [PubMed: 20816881]
4. Zubarev RA, Horn DM, Fridriksson EK, Kelleher NL, Kruger NA, Lewis MA, Carpenter BK, McLafferty FW. Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal Chem*. 2000; 72(3):563–73. [PubMed: 10695143]
5. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A*. 2004; 101(26):9528–33. [PubMed: 15210983]
6. Biemann K. Appendix 5. Nomenclature for peptide fragment ions (positive ions). *Meth. Enzymol*. 1990; 193:886–888.

7. Paizs B, Suhai S. Fragmentation pathways of protonated peptides. *Mass Spectrom Rev.* 2005; 24(4): 508–48. [PubMed: 15389847]
8. Zubarev RA. Electron-capture dissociation tandem mass spectrometry. *Curr Opin Biotechnol.* 2004; 15(1):12–6. [PubMed: 15102460]
9. Wysocki VH, Tsaprailis G, Smith LL, Breci LA. Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom.* 2000; 35(12):1399–406. [PubMed: 11180630]
10. Kapp EA, Schutz F, Reid GE, Eddes JS, Moritz RL, O'Hair RA, Speed TP, Simpson RJ. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal Chem.* 2003; 75(22):6251–64. [PubMed: 14616009]
11. Huang Y, Triscari JM, Tseng GC, Pasa-Tolic L, Lipton MS, Smith RD, Wysocki VH. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Anal Chem.* 2005; 77(18):5800–13. [PubMed: 16159109]
12. Swaney DL, McAlister GC, Wirtala M, Schwartz JC, Syka JE, Coon JJ. Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Anal Chem.* 2007; 79(2):477–85. [PubMed: 17222010]
13. Chalkley RJ, Medzihradszky KF, Lynn AJ, Baker PR, Burlingame AL. Statistical Analysis of Peptide Electron Transfer Dissociation Fragmentation Mass Spectrometry. *Analytical Chemistry.* 2010; 82(2):579–584. [PubMed: 20028093]
14. Kandasamy K, Pandey A, Molina H. Evaluation of several MS/MS search algorithms for analysis of spectra derived from electron transfer dissociation experiments. *Anal Chem.* 2009; 81(17): 7170–80. [PubMed: 19639959]
15. Baker PR, Medzihradszky KF, Chalkley RJ. Improving software performance for peptide ETD data analysis by implementation of charge-state and sequence-dependent scoring. *Mol Cell Proteomics.* 2010; 9(9):1795–1803. [PubMed: 20513802]
16. Sun RX, Dong MQ, Song CQ, Chi H, Yang B, Xiu LY, Tao L, Jing ZY, Liu C, Wang LH, Fu Y, He SM. Improved peptide identification for proteomic analysis based on comprehensive characterization of electron transfer dissociation spectra. *J Proteome Res.* 2010; 9(12):6354–67. [PubMed: 20883037]
17. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis.* 1999; 20(18):3551–67. [PubMed: 10612281]
18. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol.* 1999; 6(3-4):327–42. [PubMed: 10582570]
19. Tabb DL, Smith LL, Breci LA, Wysocki VH, Lin D, Yates JR 3rd. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem.* 2003; 75(5): 1155–63. [PubMed: 12641236]
20. Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem.* 2003; 75(4):768–74. [PubMed: 12622365]
21. Li W, Song C, Bailey DJ, Tseng GC, Coon JJ, Wysocki VH. Statistical analysis of electron transfer dissociation pairwise fragmentation patterns. *Anal Chem.* 2011; 83(24):9540–5. [PubMed: 22022956]
22. Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal Chem.* 2004; 76(14):3908–22. [PubMed: 15253624]
23. Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal Chem.* 2005; 77(19):6364–73. [PubMed: 16194101]
24. Gibbons FD, Elias JE, Gygi SP, Roth FP. SILVER helps assign peptides to tandem mass spectra using intensity-based scoring. *J Am Soc Mass Spectrom.* 2004; 15(6):910–2. [PubMed: 15144981]
25. Li W, Ji L, Goya J, Tan G, Wysocki VH. SQID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry. *J Proteome Res.* 2011; 10(4):1593–602. [PubMed: 21204564]

Figure 1.

Comparison of ETD MS spectra of (a) doubly and (b) triply charged precursors of the peptide QLWGLLIEETEK. The fragmentation spectrum of the doubly-charged precursor is dominated by C-terminally derived z+1 ions, whereas the spectrum from the triply-charged precursor contains extensive fragment ion series from both ends of the peptide.

Figure 2.

Pie charts breaking down the precursor charge state of spectra identified by a selection of search engines when analyzing a common ETD fragmentation dataset[14]. Total number of spectra identified at each charge state by each search engine are also indicated.

Table 1

Frequency of observing different ion types in ETD fragmentation spectra of peptides produced by different enzyme digestions. Results are separated for doubly charged precursors and those with a higher charge state.

Enzyme Prec. Charge	LysC 2+	LysC > 2+	LysN 2+	LysN >2+	CNBr 2+	CNBr >2+	Trypsin 2+	Trypsin >2+
b	0.02	0.03	0.06	0.04	0.04	0.05	0.03	0.03
c-1	0.11	0.05	0.17	0.06	0.14	0.08	0.10	0.04
c	0.20	0.33	0.49	0.42	0.31	0.36	0.11	0.34
y	0.13	0.09	0.05	0.06	0.06	0.06	0.14	0.11
z	0.30	0.29	0.17	0.27	0.27	0.25	0.31	0.31
z+1	0.23	0.20	0.07	0.15	0.18	0.21	0.32	0.17