Published in final edited form as:

Nat Protoc. 2006 ; 1(5): 2213. doi:10.1038/nprot.2006.330.

Verification of automated peptide identifications from proteomic tandem mass spectra

David L Tabb^{1,2}, David B Friedman¹, and Amy-Joan L Ham¹

¹Mass Spectrometry Research Center, Department of Biochemistry, Nashville, Tennessee 37232-8340, USA.

²Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee 37232-8340, USA.

Abstract

Shotgun proteomics yields tandem mass spectra of peptides that can be identified by database search algorithms. When only a few observed peptides suggest the presence of a protein, establishing the accuracy of the peptide identifications is necessary for accepting or rejecting the protein identification. In this protocol, we describe the properties of peptide identifications that can differentiate legitimately identified peptides from spurious ones. The chemistry of fragmentation, as embodied in the 'mobile proton' and 'pathways in competition' models, informs the process of confirming or rejecting each spectral match. Examples of ion-trap and tandem time-of-flight (TOF/TOF) mass spectra illustrate these principles of fragmentation.

INTRODUCTION

Proteomic tandem mass spectrometry

Tandem mass spectrometry (or MS/MS) has become a widespread technology for proteomics. In this technique, enzymatic digestion by a site-specific protease (usually trypsin) cleaves proteins into a mixture of peptides. Once inside the mass spectrometer, ions of a particular peptide are isolated in an initial step of mass spectrometry and collided with gas to produce daughter ions. These fragments are cataloged to produce a tandem mass spectrum. Identification of each MS/MS as a particular peptide sequence is typically performed by database search algorithms such as Sequest¹ or Mascot². These algorithms are essential to proteomics, but a substantial fraction of the identifications they produce are marginal; some are correct, and others are not. Protein identifications, in turn, may be supported by only a few observed peptides. To confirm or reject individual peptide identifications and thus the protein identifications based upon them is a challenging problem. The aim of this protocol is to provide an approach that can be used to differentiate legitimately identified peptides from spurious ones. Proteomic identification is a complex process affected by separation and ionization, the interplay of fragmentation mechanisms, instrument-dependent effects and the types and configurations of algorithms used for identification. These are discussed in more detail below.

Separation and ionization

Several types of instruments are often used for proteomics. Analysts that use quadrupole ion-trap instruments (such as the Thermo LCQ or LTQ)³ typically use a different work-flow than users of tandem TOF instruments (such as the Applied Biosystems 4700 or 4800 TOF/TOF)⁴. Both workflows typically begin with digestion of proteins into peptides. The key difference hinges on the ionization techniques that are most appropriate to these instruments. Ion-trap users typically separate the peptides by liquid chromatography directly in line with electrospray ionization of the peptides as they enter the mass spectrometer. TOF/TOF instruments generally use matrix-assisted laser desorption/ionization (MALDI), and this process requires crystallizing an acidic, UV-absorbing matrix with embedded peptides onto sample plates for analysis.

The type of ionization used will also affect the charges observed for peptides: electrospray typically produces singly, doubly and triply charged peptide ions, whereas MALDI generally produces singly charged peptide ions. The positive charge takes the form of one or more protons carried by the peptide. When the peptides are selected for fragmentation, the protons are associated with basic residue side chains or with the N terminus of the peptide.

Interplay of fragmentation mechanisms

Collision-induced dissociation (CID) energizes each peptide ion through collisions with gas molecules. As the internal energy increases for a particular peptide ion, it begins sampling new conformations, and an associated proton can become mobile if insufficient basicity is holding it in place. The resulting fragmentation is described by the mobile proton model⁵ (Fig. 1a). When a proton nears a peptide bond, the interaction enables an attack by the preceding peptide bond's carbonyl to create a ring structure. The intermediate then breaks to form an N-terminal *b*-series fragment and a C-terminal *y*-series fragment. If the original peptide carries at least two protons, both of these fragments are likely to emerge as ions. Because this mechanism requires attack by the preceding carbonyl, the initial peptide bond cannot break by this mechanism. Conditions are optimized for single cleavage events per molecule, and because many molecules of the same peptide undergo CID simultaneously, a heterogeneous collection of fragment ions will form, with different molecules breaking between different residues.

Although the above description explains the origin of *b*- and *y*-series ions under mobile-proton conditions, many other peptide fragmentation pathways are possible. The 'pathways in competition' model⁶ describes fragmentation as a competitive process in which some mechanisms can supersede others. In some cases, *b*- and *y*-series ions result from 'charge remote' mechanisms that do not require a proton's proximity. For example, if all protons are tightly bound to basic sites, aspartic acid can cleave the peptide bond to its C-terminal side to make this *b*-*y* pair dominant above the others⁷ (Fig. 1b). Other reactions are also possible that produce ions other than *b* and *y* types. Neutral losses from the peptide ion can occur for peptides that carry labile post-translational modifications. For example, phosphorylations on serines and threonines are prone to β -elimination⁸. Likewise, peptides containing oxidized methionine residues can lose methane sulfenic acid⁹ (Fig. 1c). Dominant alternative fragmentation pathways can greatly reduce the pool of fragment ions observed in a tandem mass spectrum, reducing its information content. Other pathways can lead to additional fragment series that complicate the spectrum. The complexity of this chemistry is inherent to CID; more regular fragmentation may be possible by techniques such as electron-capture dissociation or electron-transfer dissociation¹⁰.

The fragility or stability of an individual peptide bond is dependent upon the amino acid residues flanking it. The 'proline effect' describes the intense fragments resulting from cleavage on the N-terminal side of proline residues, but the magnitude by which this cleavage

exceeds others is highly dependent upon the residue on the N-terminal side of the bond¹¹. Glycine and serine also enhance N-terminal cleavage (to a lesser extent), whereas valine, leucine and isoleucine produce more fragment ions from the bonds at their C-terminal sides¹².

Other effects are layered on top of these residue-specific trends. In ion traps, the peptide bonds near the middle of the peptide tend to produce more intense fragment ions than bonds near the terminus¹². The positions of basic residues can also affect fragment intensities: in general, the fragment ions that contain basic residues are more likely to generate intense peaks than those that do not^{13,14}. On the other hand, fragments that contain serine and threonine residues are more likely to be diminished by neutral loss of water^{12,15}, whereas fragments that contain asparagine, glutamine, arginine or lysine may lose ammonia^{12,16}. These effects cumulatively determine the pattern of intensities observed in fragment ions.

Instrument-dependent influences

The conditions under which CID occurs affect the peptide fragmentation patterns and are different for ion traps, QqTOFs (hybrid instruments combining quadrupoles, collision cells and time-of-flight mass analyzers) and TOF/TOF analyzers.

Ion traps generally employ low energy (eV) to induce fragmentation for ions at a target m/z ¹⁷ through a series of collisions with gas molecules. Ions outside the target mass/charge ratio (m/z) range do not show the effects of this added energy; once a peptide ion has dissociated, the resulting fragment ions are not excited to fragment further. Use of 'broadband activation', on the other hand, broadens the m/z range in which subsequent fragmentation takes place. In general, the precursor ion is completely fragmented in this process. Physical constraints inherent to ion-trap mass analyzers result in the loss of ions below one-third of the precursor ion m/z during MS/MS.

In QqTOFs, the fragmentation is similar, with a few departures¹⁸. First, CID takes place in a radio-frequency collision cell rather than an ion trap, enabling the retention of low-mass fragment ions. Second, all ions passing through the collision cell are excited, though this excitation takes place during a shorter interval than in an ion trap, and so secondary fragmentation of *b* and *y* ions may be induced. In brief, this fragmentation is quite similar to the process observed with triple-quadrupole instruments.

Fragmentation is less standardized in TOF/TOF instruments, ranging from high-energy (keV) collisions that can break both side-chain bonds and peptide bonds¹⁹ to low-energy collisions. TOF/TOF analyzers do not exclude the low m/z region, and this part of the spectrum can include not only the smallest *b* and *y* ions but also immonium ions that can give information about the composition of the peptide sequence (Fig. 1d). Also, the resolution and mass accuracy of TOF/TOF spectra are generally an order of magnitude greater than that of ion traps, though the diversity of fragment ions is somewhat reduced.

Limitations of database searches

Despite their vital importance to protein identification, database search algorithms generally use simplistic models of fragmentation. These tools work by comparing an observed spectrum to a spectrum model based upon a particular peptide sequence. Typically, cleavage of each peptide bond is modeled as equally likely. In Sequest, a theoretical spectrum is populated with peaks corresponding to the fragment ions (where all *y* ions are the same intensity and all *b* ions are the same intensity), and the projected intensity and m/z data are compared to the observed spectrum in the Fourier domain¹. The X!Tandem algorithm simulates fragmentation of peptide sequences to produce more accurate theoretical spectra, and it computes a dot product that

characterizes the similarity between the intensities in the observed spectrum and those of the theoretical one^{20,21}. Mascot² and OMSSA²², on the other hand, produce a list of m/z values where fragment ions may be expected for a given sequence, and the numbers of matches and mismatches are used to produce a score for the sequence.

A MudPit3²³ experiment may produce more than 100,000 tandem mass spectra; in the case of the eight-cycle MudPit data used here as an example, 87,533 MS/MS were produced. Because many of these spectra will be identified under multiple precursor charge-state assumptions, the raw number of identifications will be even higher (in this case, 138,325). Clearly, manually evaluating these identifications is not feasible; generally, this task is reserved for peptides important for identifying a protein of particular biological interest. Necessarily, most peptide identifications are passed or rejected by algorithms rather than by researchers. Monitoring the percentage of total spectra that yield passing identifications can generate a useful statistic for characterizing experiments. On rare occasions, 50% of the spectra generated in a MudPit experiment will be confidently identified. More typically, 10% of spectra will be identified. This measure, however, also reflects many other properties of the experiment, such as sample complexity.

When an experiment has resulted in an unsatisfactory number of identified peptides, many researchers will respond by lowering the score thresholds to include a larger number of identifications. The danger of this practice, however, is that false identifications are particularly concentrated at lower scores. Reducing thresholds will increase the overall number of identifications, but it can also substantially increase the error rate of the passing identifications. The convention to which shotgun proteomics is currently adapting is that the false positive rate should always be characterized in these experiments. To publish their results, researchers should be able to estimate the percentage of passing identifications that are incorrect, whether by using reversed or randomized database searches^{24,25} or by using software trained on identifications by the appropriate database identifier²⁶.

Because database search algorithms use simple models of fragmentation, the identifications they produce can be confirmed or rejected on the basis of spectral information that their scoring routines do not use. If researchers simply count the numbers of matched ions to evaluate the match, they are only repeating an operation already conducted as part of match scoring. Instead, manual evaluation should be based on chemical fragmentation rules that are not incorporated into the automated match scoring. In previous work, some of these rules have been assembled for manual validation²⁷ or automated analysis of identifications²⁸ (see http://www.proteomesoftware.com/Proteome_software_pro_protein_id.html and http://www.cebi.sdu.dk/Steen_Mann_NRM_Suppl_PeptValid.pdf). In this protocol, we present a step-by-step strategy for performing this validation from a fragmentation point of view (Fig. 2).

MATERIALS

REAGENT SETUP

Protein structure disruption—In most shotgun proteomics experiments, protein structures are disrupted through reduction and denaturation. The reduction of disulfide bonds by DTT is commonly followed by alkylation with iodoacetamide, a process that causes cysteine residues to be carboxyamidomethylated (thus giving them a mass of 160 Da rather than 103 Da). Cysteine residues can be problematic for proteomics because they can become chemically modified in response to sample handling. Denaturation can be managed by several means. Researchers should be aware that some denaturants (particularly urea) can lead to chemical modification of peptides.

Protease digestion—Digestion of proteins to peptides can be achieved using several enzymes. Trypsin, however, is by far the most commonly used, in part because of its specificity and robustness. Peptides resulting from trypsin digestion are nearly ideal for mass spectrometry because they tend to be between 5 and 30 residues and produce dominant y-series ions owing to the basic residue generally found at the C terminus. As a result, most attempts to characterize peptide fragmentation from sets of identified peptides have used trypsin digest data. Chymotrypsin and other enzymes, however, can be valuable when maximization of sequence coverage is a priority.

In the two examples discussed in the anticipated results, the samples were prepared as follows.

***Mus musculus* brush borders**—Proteins were extracted from mouse brush borders. They were reduced and alkylated before trypsin digestion.

***Helicobacter pylori* proteins**—*H. pylori* proteins were isolated from membrane fractions, resolved and analyzed using 2D difference gel electrophoresis²⁹. Proteins were reduced and alkylated during gel separation, and an integrated Spot Handling Workstation (GE Healthcare) excised each spot, subjected it to in-gel trypsin digestion and mixed the resulting peptides with matrix (5 mg ml⁻¹ α -cyano-4-hydroxycinnamic acid supplemented with 1 mg ml⁻¹ ammonium citrate) on a MALDI target.

EQUIPMENT SETUP

Sample separation—Complex samples require separation before their introduction to a tandem mass spectrometer. In general, a reversed-phase liquid chromatography (RPLC) column is sufficient to separate peptides produced from mixtures of fewer than 100 proteins. More complex mixtures can be more sensitively examined when additional dimensions of separation are used. If MALDI is used, researchers have generally found that 5 mg ml⁻¹ α -cyano-4-hydroxycinnamic acid should be supplemented with 1 mg ml⁻¹ ammonium citrate to reduce background ionization.

Instrument settings—Because each instrument manufacturer provides different options for configuring tandem mass spectral collection, instruments from different manufacturers cannot generally be set up identically. This protocol is most applicable to low-energy CID fragmentation, which uses fragmentation energies in the tens of eV. Thermo instruments express ‘normalized collision energy’ as a percentage rather than as a kinetic-energy value. Other manufacturers express this value in terms of the voltages used to propel ions through collision cells. We recommend examining the fragment ions produced for a variety of peptide sequences that vary in basicity and length to determine the optimal collision-energy settings for a particular instrument. Test peptides can be obtained by digesting a single protein or a defined mixture of a few proteins. Too little energy will leave a dominant intact peptide ion, but too much may produce secondary fragments from the *b*- and *y*-series ions or decrease the overall signal of the spectrum.

Data analysis—Identification of tandem mass spectra can be managed by several major database search algorithms. For ion-trap data, Sequest¹ has long been a standard tool, but many others have been shown effective in this context as well. Mascot² is usable on a variety of data types and offers the additional feature of using TOF MS data for peptide mass-mapping identification of proteins. Open-source, freely available algorithms such as X!Tandem²¹ and OMSSA²² have also proven themselves as viable contenders for database identification.

Conducting identification in such a way as to enable determination of false-positive rates is invaluable. The most common way this is done is by searching sequence databases that contain

both legitimate matching sequences and sequences that are known not to be present in the mixture. 'Decoy' sequences are proteins that are from a species unrelated to the one from which the sample was produced (such as Arabidopsis proteins when a bacterial sample is being processed). 'Reversed' databases contain two versions of each sequence in the database, the normal orientation and a version that has been reversed to read from the protein C terminus to the protein N terminus. The appearance of decoy or reversed proteins in the final list of identifications can be used to determine the overall error rate of identification^{24,25}.

In the two examples discussed in the anticipated results, the samples were prepared as follows.

M. musculus brush borders—Fractions from strong cation-exchange chromatography were subjected to RPLC in 100-mm columns *en route* to a Thermo Finnigan LTQ linear ion-trap tandem mass spectrometer. Spectra were identified with the International Protein Index database for mice, version 3.13 (50,489 proteins; <http://www.ebi.ac.uk/IPI/>) by Vanderbilt's Myri-Match database identifier.

H. pylori proteins—MALDI-TOF MS and data-dependent TOF/TOF MS/MS was performed using a Voyager 4700 (Applied Biosystems). Functioning as part of GPS Explorer (Applied Biosciences), The Mascot² algorithm used both MALDI-TOF MS spectra (representing peptide mass maps) and TOF/TOF tandem mass spectra of intense peptide ions to identify proteins from the NCBI database (<http://www.ncbi.nlm.nih.gov/RefSeq/>). Searches allowed for one missed cleavage and carbamidomethylation of cysteine but did not constrain molecular weight, pI or taxonomy.

PROCEDURE

1| Evaluate tandem mass spectral quality

A spectrum that has been collected from a peptide present at too small a concentration has an elevated chance of being identified incorrectly (see Fig. 3). Features that may be helpful in recognizing low quality spectra include the following: (i) The total ion current for the spectrum may be low, (ii) the number of peaks reported for the spectrum is small, (iii) the observed peaks are all similar in intensity and (iv) the precursor ion is present at low intensity in the MS scan.

? TROUBLESHOOTING

2| Consider other matches to this peptide

The quality of a spectral match can be difficult to evaluate in isolation. Many times, however, a peptide is identified from several spectra, increasing confidence in each match. Look for other spectra that have been identified as the same sequence. If the peptide sequence is identified at multiple charge states, the probability of correct identification is higher. If the peptide exists in both normal and modified forms (such as one with an oxidation on a methionine residue), the two variants can help to confirm each other.

▲ CRITICAL STEP Some identification algorithms will attempt to identify each spectrum at multiple precursor charges; if multiple high-scoring identifications have resulted from a particular spectrum, choose one to accept.

3| Examine the overall match visually

Use software to superimpose the fragments expected for this peptide upon the observed spectrum (this feature is generally part of database identification software packages). Many users count the numbers of matched ions to verify identifications, but the number of matched

peaks is generally part of the score for the identification, giving this practice little additional value. Consider the following (Fig. 4 and 5):

1. Are there numerous intense, unexplained peaks throughout the spectrum? These mismatches are symptoms of an incorrect identification. Alternatively, this may be a chimeric MS/MS, containing fragments from multiple peptides (Fig. 4).
2. What proportion of the total intensity for this spectrum can be accounted for by this sequence? Accounting for more than half of the intensity should only be expected when noise peaks have been removed from the spectrum. High-intensity spectra are generally populated with numerous low-intensity fragments, reducing the proportion of intensity in fragments.
3. Is the fragment ion series from one terminus well-matched, whereas the other is almost entirely unmatched (Fig. 5)? For singly charged peptide ions, this may result because a terminal arginine residue assures that one ion series dominates the other. In multiply charged peptide ions, this phenomenon may indicate an incorrectly determined precursor charge.

▲ CRITICAL STEP Small peptides (6 residues or less) produce very few fragment ions. This will not only limit the ability of the algorithm to identify the sequence correctly but also reduce a researcher's ability to validate the identification manually.

4| Recognize neutral-loss fragments

Both the precursor ion and fragment ions may decompose by the neutral loss of small molecules. Losses from the precursor can result in major peaks at a slightly lower m/z than the precursor. Losses from fragment ions, in contrast, can lead to pairs of peaks for each fragment ion that produces the loss. Finding these losses can be useful in confirming or rejecting peptide identifications on the basis of sequence (Fig. 6):

1. Fragment ions that contain Ser or Thr may lose water (-18 Da)¹².
2. Fragment ions that contain Asn or Gln may lose ammonia (-17 Da)¹².
3. b ions may lose carbon monoxide (-28 Da) to form a ions as secondary fragment ions; this is enhanced for b ions containing Arg (D.L.T., unpublished data).
4. Peptides with Gln residues at the N terminus can lose ammonia (-17 Da)³⁰.
5. Peptides with oxidized Met residues can lose methane sulfenic acid (-64 Da)⁹.
6. Peptides and fragment ions containing phospho-Ser or phospho-Thr can lose phosphoric acid (-98 Da)⁸ (Fig. 6).

▲ CRITICAL STEP Database searches that allow for multiple post-translational modifications, that expect peptides to result from nonspecific cleavages or that make use of very large sequence databases may overwhelm an identification algorithm's ability to discriminate correct sequences from incorrect ones. These search results require greater scrutiny than those from ordinary tryptic searches against a more limited sequence database.

? TROUBLESHOOTING

5| Examine low-mass reporter ions

If the MS/MS gives information for low-mass ions, the sequence composition may be checked against the immonium ions. These reporter ions appear at 27 Da less than the amino acid residue masses and are most readily observed for proline, valine, isoleucine/leucine, histidine, phenylalanine, tyrosine and tryptophan. Similarly, the C-terminal residue of the peptide can

be checked in this region. If a peptide's C terminus is lysine, the first *y* ion will appear at 147 Da (though this fragment does not always ionize efficiently in TOF/TOF). If the C terminus is arginine, the first *y* ion will appear at 175 Da.

6| Evaluate sequence effects on fragment intensity

Multiple trends affect the intensity of each fragment ion. Peptide bonds in the middle of each peptide tend to cleave more readily than those nearer the termini. In ion traps, ions from the *y* series typically have double the intensity of ions from the *b* series. Particular residues can also have direct influences on fragmentation (Fig. 7 and 8a):

1. The 'proline effect' describes the intense fragment ions formed to the N-terminal side of a Pro residue (Fig. 7). In particular cases (such as Asp-Pro), these ions can completely dominate the tandem mass spectrum¹¹.
2. The side chain of an Asp residue can cleave the peptide bond to its C-terminal side if the peptide's ionizing protons are immobilized by basic residues⁷ (Fig. 8a). In some cases (such as when Arg 'sequesters' the proton), these ions can dominate the tandem mass spectrum.
3. Branched hydrophobic side chains (Val, Ile, Leu) favor fragmentation of the bonds to their C-terminal sides. Gly and Ser, on the other hand, favor fragmentation of the bonds to their N-terminal sides¹².
4. Peptide bonds between Asn and Gly are very labile³¹, potentially owing to their enabling an intermediate succinimide structure³².
5. The side chain of His can attack its own C-terminal bond, enhancing the formation of this *b* ion^{12,33}.

7| Evaluate basicity effects on fragment intensity

When a peptide dissociates, the basicity and size of the two produced fragments determine the number of protons each fragment will retain^{13,14}. If the peptide sequence contains only one basic residue (arginine, histidine or lysine), the fragments that contain the basic residue will be more intense than the others (see Fig. 8b). When a triply charged peptide dissociates, fragments that contain multiple basic residues are those most likely to adopt a double charge.

8| Assess protein evidence

Proteomics is generally focused on proteins, not peptides. In species with compact genomes, most peptides will be unique to a particular protein. In species where gene duplication has taken place, peptide identifications are more difficult to assemble into protein identifications. The MIAPE guidelines^{34,35} direct researchers to do the following:

1. Differentiate between peptides that are shared among multiple proteins and peptides that are distinct to a particular protein.
2. Identify indistinguishable protein groups when it is unclear which protein(s) of the group has led to the observed peptides.
3. Report the smallest number of proteins and protein groups necessary to explain the observed peptides. For example, if protein 1 explains the presence of peptides A, B and C, whereas protein 2 explains the presence of only peptides A and B, then claiming the identification of protein 1 alone is more parsimonious than claiming that both proteins 1 and 2 are present^{36,37}.

? TROUBLESHOOTING

Step 1: low-quality spectra—Technical replicates can be a valuable resource for validation. If multiple mass spectrometry data sets are collected from a single sample, a marginal identification may be reinforced by its appearance in multiple replicates. If an identification of low signal is found in only one replicate, the identification is more likely to be false. Researchers may need to increase the sample concentration or improve sample separation to improve the sensitivity of the mass spectrometry. Although not practical on a global scale, an alternative approach to the confirmation of the match is to reanalyze the sample and target a particular m/z value for tandem MS fragmentation, rather than using traditional data-dependent analysis. If the m/z value for a given peptide (or small collection of peptides) is targeted, a spectrum will be sampled across a chromatographic peak, and a high-quality spectrum may result when the peptide is at its highest concentration. In addition to targeting the m/z value, the scan time or number of scans averaged for a spectrum can be increased to achieve a higher-quality spectrum that can confirm or exclude a particular assignment of a spectrum to a sequence.

Step 4: MS/MS/MS—Multiple fragmentations in series can be used to probe peptide structures. If an MS/MS spectrum is dominated by a particular fragment (as is characteristic of phosphoserine and phosphothreonine-containing peptides), MS/MS/MS of this dominant fragment may be helpful to reveal peptide structure. This type of mass spectrometry, generally available only with ion-trap mass analyzers, can often produce nearly complete fragment series from a peptide once the initial loss has taken place. These spectra can help confirm the initial peptide identification. Some instruments provide data-dependent methods to scan for common neutral-loss ions and automatically fragment them if they are observed.

ANTICIPATED RESULTS

As an example of this evaluation process, we examined the mouse brush border sample. The biologist who submitted the sample is interested in the presence of calmodulin. The evidence for this 16.7-kDa protein, however, is tenuous; only two peptides were observed among 11,251 tandem mass spectra. If they were clear matches, we could confidently identify this protein, but both are from triply charged precursors. Manual examination is necessary to confirm or reject the presence of this protein.

The first peptide is M*KDTDSEEEIREAFR, where M* represents an oxidized methionine (Fig. 9). For Step 1, the intensity for the spectrum is acceptable, with considerable dynamic range present. For the second, no other spectra matched to this peptide sequence, so confirming it by other spectral evidence is not possible. Visually, the match is not impressive, and three of the four most intense peaks appear to correspond to neutral losses from the precursor rather than b or y fragments. Of the intense ions, a few do not align to expected fragments. For Step 4, we note that the precursor does appear to lose 64 Da owing to an oxidized methionine and to lose ammonia or water. These loss ions could also be subjected to an MS/MS/MS experiment to confirm their assignment. Since this is an ion-trap MS/MS, we cannot see the low-mass ions for Step 5. For Step 6, we see that the y_{13} ion is the most intense observed ion, but a breakage between lysine and aspartate should not be particularly favored. In Step 7, we note that the two arginine residues are close to the C terminus, making a doubly charged y series most likely, and this is what we observed. In total, this peptide match is plausible, but the identification is not entirely reliable as it is based on a small number of matched fragments.

The second peptide is VFDDKDGNGYISAAELR. Although the overall intensity is slightly lower than that of the former spectrum, the spectrum still shows an acceptable signal and dynamic range. In Step 2, we note that a second spectrum (data not shown) has also been identified with this triply charged sequence, increasing our confidence that this identification is correct. Visual inspection reveals that y_2 through y_7 are observed in singly charged form,

and several doubly charged fragments are observed as well, with a large ion just below the precursor m/z left unmatched. For Step 4, we observe that ammonia losses are abundant among the doubly charged b -series ions, corresponding to succinimide formation in Asn-Gly at positions 7 and 8. In Step 6, we note that the most intense ions are $b_{10} + 2$ and $y_6 + 1$. This complementary pair has resulted from augmented fragmentation on the N-terminal side of serine and the C-terminal side of isoleucine. For Step 7, we note that the doubly charged b ions can be explained by the presence of a basic N terminus, a lysine residue and an asparagine residue in these fragments. Although the first peptide we examined left many doubts, this match is a confidently identified triply charged peptide.

Advancing to Step 8, we evaluate our confidence in the presence of calmodulin from examining these two peptides. We observe that M*KDTDSEEEIREAFR is a sequence found in four different proteins in the mouse database. For two of these proteins, this peptide was the only one identified, and so the data-mining software removed the two proteins as artifacts. Peptide VFDDKDGNGYISAAELR, in contrast, is found in only the remaining two proteins of which the other peptide was part. We are left with two protein identifiers, either of which explains the presence of the two peptides we have observed. Because we cannot differentiate the two proteins, we group them together as indistinguishable, equivalent protein identifications. Both are annotated as calmodulin sequences. On the basis of our review of the two identifications collected for this protein, we judge that one or both of these proteins is present.

These rules are most useful for evaluating uncertain identifications or those that will be used for follow-up experiments. They can be used to reject protein identifications for which the peptide evidence is lacking, but they can also be used to reinforce protein identifications when the peptide identifications stand up to careful analysis. In a simple 1D RPLC run, it may be possible to identify a protein by a single tandem mass spectrum for a unique peptide, but when millions of tandem mass spectra have been identified, this is almost always inappropriate. For proteins that are on the line between acceptance and rejection, manual validation of peptide identifications is an essential step.

Acknowledgments

This work was supported by US National Institutes of Health grant P30 ES000267, by an American Cancer Society Institutional Research Grant (no. IRG-58-009-48) through the Sartain-Lanier Family Foundation and Vanderbilt-Ingram Cancer Center Discovery Grant Programs, and by the Vanderbilt Academic Venture Capital Fund. M. Tyska (Vanderbilt University; supported by a Career Development Award from the Crohn's and Colitis Foundation of America) contributed the spectra of mouse brush border proteins, and R. Peek and A. Franco (Vanderbilt University) contributed *H. pylori* membrane samples.

References

- Eng JK, McCormack AL, Yates JR III. 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.
- Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;20:3551–3567. [PubMed: 10612281]
- Washburn MP, Wolters D, Yates JR III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol* 2001;19:242–247. [PubMed: 11231557]
- Daikoku T, et al. Proteomic analysis identifies immunophilin FK506 binding protein 4 (FKBP52) as a downstream target of Hoxa10 in the periimplantation mouse uterus. *Mol. Endocrinol* 2005;19:683–697. [PubMed: 15528267]
- Wysocki VH, Tsaprailis G, Smith LL, Brei LA. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom* 2000;35:1399–1406. [PubMed: 11180630]
- Paizs B, Suhai S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev* 2005;24:508–548. [PubMed: 15389847]

7. Huang Y, Wysocki VH, Tabb DL, Yates JR III. The influence of histidine on cleavage C-terminal to acidic residues in doubly protonated tryptic peptides. *Int. J. Mass Spectrom* 2002;219:233–244.
8. Schlosser A, Pipkorn R, Bossemeyer D, Lehmann WD. Analysis of protein phosphorylation by a combination of elastase digestion and neutral loss tandem mass spectrometry. *Anal. Chem* 2001;73:170–176. [PubMed: 11199962]
9. Reid GE, Roberts KD, Kapp EA, Simpson RI. Statistical and mechanistic approaches to understanding the gas-phase fragmentation behavior of methionine sulfoxide containing peptides. *J. Proteome Res* 2004;3:751–759. [PubMed: 15359728]
10. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. USA* 2004;101:9528–9533. [PubMed: 15210983]
11. Brei LA, Tabb DL, Yates JR III, Wysocki VH. Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal. Chem* 2003;75:1963–1971. [PubMed: 12720328]
12. Tabb DL, et al. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem* 2003;75:1155–1163. [PubMed: 12641236]
13. Paizs B, Suhai S. Towards understanding some ion intensity relationships for the tandem mass spectra of protonated peptides. *Rapid Commun. Mass Spectrom* 2002;16:1699–1702. [PubMed: 12203239]
14. Tabb DL, Huang Y, Wysocki VH, Yates JR III. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem* 2004;76:1243–1248. [PubMed: 14987077]
15. Ballard KD, Gaskel SJ. Dehydration of peptide $[M+H]^+$ ions in the gas phase. *J. Am. Soc. Mass Spectrom* 1993;4:477–481.
16. 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]
17. Jonscher KR, Yates JR III. The quadrupole ion trap mass spectrometer—a small solution to a big challenge. *Anal. Biochem* 1997;244:1–15. [PubMed: 9025900]
18. Chernushevich IV, Loboda AV, Thomson BA. An introduction to quadrupole-time-of-flight mass spectrometry. *J. Mass Spectrom* 2001;36:849–865. [PubMed: 11523084]
19. Medzihradsky KF, et al. The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal. Chem* 2000;72:552–558. [PubMed: 10695141]
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;75:768–774. [PubMed: 12622365]
21. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;20:1466–1467. [PubMed: 14976030]
22. Geer LY, et al. Open mass spectrometry search algorithm. *J. Proteome Res* 2004;3:958–964. [PubMed: 15473683]
23. Link AJ, et al. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol* 1999;17:676–682. [PubMed: 10404161]
24. Cargile BJ, Bundy JL, Stephenson JL Jr. Potential for false positive identifications from large databases through tandem mass spectrometry. *J. Proteome Res* 2004;3:1082–1085. [PubMed: 15473699]
25. Higdon R, Hogan JM, Van Belle G, Kolker E. Randomized sequence databases for tandem mass spectrometry peptide and protein identification. *OMICS* 2005;9:364–379. [PubMed: 16402894]
26. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem* 2002;74:5383–5392. [PubMed: 12403597]
27. Chen Y, Kwon SW, Kim SC, Zhao Y. Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *J. Proteome Res* 2005;4:998–1005. [PubMed: 15952748]
28. 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;15:910–912. [PubMed: 15144981]

29. Lilley KS, Friedman DB. All about DIGE: quantification technology for differential-display 2D-gel proteomics. *Expert Rev. Proteomics* 2004;1:401–409. [PubMed: 15966837]
30. Tabb DL, Narasimhan C, Strader MB, Hettich RL. DBDigger: reorganized proteomic database identification that improves flexibility and speed. *Anal. Chem* 2005;77:2464–2474. [PubMed: 15828782]
31. Kapp EA, et al. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem* 2003;75:6251–6264. [PubMed: 14616009]
32. Geiger T, Clarke S. Deamidation, isomerization, and racemization at asparaginy and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation. *J. Biol. Chem* 1987;262:785–794. [PubMed: 3805008]
33. Tsaprailis G, et al. A mechanistic investigation of the enhanced cleavage at histidine in the gas-phase dissociation of protonated peptides. *Anal. Chem* 2004;76:2083–2094. [PubMed: 15053674]
34. Bradshaw RA. Revised draft guidelines for proteomic data publication. *Mol. Cell. Proteomics* 2005;4:1223–1225. [PubMed: 16160100]
35. Carr S, et al. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol. Cell. Proteomics* 2004;3:531–533. [PubMed: 15075378]
36. Yang X, et al. DBParser: web-based software for shotgun proteomic data analyses. *J. Proteome Res* 2004;3:1002–1008. [PubMed: 15473689]
37. Tabb DL, McDonald WH, Yates JR III. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res* 2002;1:21–26. [PubMed: 12643522]

Figure 1.

Pathways in competition. **(a)** A 4-residue tryptic peptide fragments to form a b_2 and y_2 ion pair under mobile-proton conditions. During CID, one proton of the doubly charged peptide migrates to a peptide bond while the other remains at the C-terminal basic residue. Electrons from the initial carbonyl attack the partially positive carbon of the second carbonyl, leading to an intermediate that separates into two fragments. In a singly charged peptide, only one fragment can take on the proton to become a fragment ion, but in multiply charged peptides, both typically ionize. R_1 – R_4 denote side chains. **(b)** When all ionizing protons are sequestered, aspartate can promote the cleavage of the peptide bond to its C-terminal side. The side chain attack forms a five-member ring intermediate, and a hydrogen rearrangement and peptide bond breakage occur to produce a b ion with a succinic anhydride cyclic structure and a normal y ion. **(c)** Oxidized methionine side chains can lose neutral methane sulfenic acid. This *cis*-1,2 elimination reaction can outcompete the production of b - y ion pairs when protons are sequestered by basic residues. The end result can be a substantial neutral-loss ion signal appearing 64 Da lower than the precursor ion position at the same charge as the original peptide. **(d)** Immonium ions give information about peptide composition rather than sequence. These small fragments are typically absent in ion-trap tandem mass spectra because these analyzers do not retain ions at less than one-third of precursor m/z values. Masses (Da) of the most common intense immonium ions are listed.

Figure 2. These nine stages of analysis can enable a researcher to differentiate correct peptide identifications from false matches.

Figure 3.

These two spectra were both identified as the doubly-charged sequence EIIGVVSQEPVLFATTIAENIR. Their appearance is quite different because of the amount of signal represented; TIC denotes the sum of fragment ion intensities for each, reported using a natural-log scale. The upper spectrum contains almost 55 times as much intensity as the lower spectrum, and the count of peaks is 11 times higher. The upper spectrum shows a broad range of intensities, with noise peaks that are conspicuously absent in the lower spectrum. The low-intensity spectrum would be an insufficient match by itself, but the presence of the high-intensity spectrum reinforces the accuracy of the identification. Visualizations of peptide

identifications from the ion trap were produced by SVGSpecView, a tool developed at Vanderbilt; inset key shows colors and peak labels used by this software.

Figure 4.

This chimeric spectrum is displayed two different ways because it contains fragment ions from two different peptides, DSDYYNMLLK (top) and M*DFSDYDLLK (bottom). The two sequences together comprise a much more complete explanation of the observed fragment ions than either individually. The y2 and y3 ions for these sequences, however, are identical. Samples of high complexity are more likely to result in spectra where fragments of multiple peptides are represented.

Figure 5.

This spectrum was identified as a doubly charged SSLKAGALR (top). This identification can be recognized as false because (i) this sequence is found only in a reversed protein sequence, (ii) only the y series is matched and (iii) multiple major peaks are unexplained by this sequence. Manual interpretation revealed that the true sequence is &SLDQLR (bottom), where ‘&’ indicates an N-terminal acetylation. This peptide is the N terminus of a protein for which several other peptides were identified. Because acetylation was not considered in the search, the algorithm did not identify the spectrum correctly. The ‘M2^’ ion represents neutral loss of water from the doubly charged precursor ion.

Figure 6.

Phosphopeptides can show distinct losses of phosphoric acid both from the precursor and from fragment ions. This spectrum represents SAS@SDTSEELNSQDSPK (where S@ is phosphoserine). The 'M2@' peak is the doubly charged precursor after loss of phosphoric acid. Because the modification was near the N terminus, the *b*-ion series shows this loss throughout the spectrum. The presence of these ions can make interpretation complex because the mass of phosphoric acid (98 Da) falls within the range of the amino acids' masses, placing neutral-loss ions near to normal *b* ions.

Figure 7.

This TOF/TOF spectrum was identified as the sequence AVREAAAGLSGPGR by the Mascot algorithm. The sequence QIQLDAGIPNDK (from a different protein) was also matched with a lower score. Although the presence of a weak y1 ion (Step 5) and strong y5 (C-terminal cleavage of leucine, Step 6c) is consistent with the first sequence (green annotation), the height of the ion peak, $m/z = 473.2$, is more consistent with the expected preferential breakage between isoleucine and proline of the second sequence (blue annotation). Data resulted from an in-gel digest of a resolved protein, and several high-confidence matches to other MS/MS spectra were also matched to the protein containing the second sequence.

Figure 8.

Special TOF/TOF fragment ions. **(a)** This TOF/TOF spectrum demonstrates that when all available protons are sequestered by basic residues, aspartate can cleave its C-terminal bond in a dominant fragmentation reaction. In this peptide, arginine prevents the migration of the proton, and the y ion on the C-terminal side of aspartate dwarfs even the y ion generated on the N-terminal side of proline. In cases where the overall signal intensity is weak, this preferential cleavage may predominate (Step 6b), but these products of charge-remote fragmentation can also lend credibility to peptide identifications. **(b)** Tryptic peptides generally produce more intense y ions than b ions because of the basic residue at the C terminus of the peptide. This TOF/TOF spectrum shows how the presence of the arginine at the N terminus of this peptide generates b series ions preferentially (Steps 3c and 7). The most prominent cleavages are breaks on the C-terminal side of aspartate (Step 6b).

Figure 9.

These two peptide identifications are evidence for the presence of calmodulin in a mouse sample. The top spectrum is dominated by multiple neutral losses from the precursor. The sequence M*KDTDSEEEIREAFR would lend itself to doubly charged y -ion production because of a pair of arginine residues in the C terminus, but the intensity of y_{13} is not obviously traceable to sequence-specific effects. Compared with the uncertainties of the top spectrum, the lower spectrum is remarkably clear-cut. The sequence VFDKDGNGYISAAELR produces complementary fragments of both single and double charge. The ammonia losses from the doubly charged b ions can be traced to the asparagine at position 7. Together, these two identifications support the claim that calmodulin is found in the mouse sample.