How do shotgun proteomics algorithms identify proteins?

Edward M Marcotte

Instrumentation aside, algorithms for matching mass spectra to proteins are at the heart of shotgun proteomics. How do these algorithms work, what can we expect of them and why is it so difficult to find protein modifications?

Shotgun proteomics is a remarkably powerful technology for identifying proteins, whether individually or in samples as complex as cell lysates. Take, for example, three notable recent applications—mapping the major protein complexes of yeast cells[1,2], systematically identifying proteins in mammalian subcellular organelles[3] and discovering diagnostic biomarkers for disease[4]—all analyses that would have been difficult, at best, by other technologies. How does shotgun proteomics work and why is it of interest in computational biology?

The concept

Named after shotgun DNA sequencing, in which long DNA sequences are computationally reconstructed from many short sequencing reads, shotgun proteomics identifies proteins from tandem mass spectra of their proteolytic peptides[5]. To explain by analogy, imagine you find a box of old-fashioned geared watches in an antique store and want to know which watch models you've stumbled across. The shotgun proteomics approach would be to dismantle each watch into its major assemblies (e.g., the movements, dials and escapements), smash each assembly into component parts, weigh the parts and look up the weights in a parts catalog, reading out which parts, and therefore which watches, you have—or rather, had. Shotgun horology might not be the method of choice for discriminating watch collectors.

Actual shotgun proteomics experiments like those mentioned above proceed in roughly the same manner and involve taking a complex mixture of proteins (e.g., purified from an organelle or associated with a particular disease state), cutting the proteins into peptides by sequence-specific proteolysis and then analyzing the mixture of peptides using mass spectrometry. Each peptide is isolated in the mass spectrometer and characterized by tandem mass spectrometry (MS/MS or MS^2), which involves breaking the peptide into many smaller fragments and measuring the mass spectrum. The component peptides, and therefore the parent proteins, are identified from the tandem mass spectra. Matching tandem mass spectra to peptide sequences is the central computational challenge of mass spectrometry–based proteomics. This identification is a difficult computational problem and ultimately determines the success of shotgun proteomics.

Edward Marcotte is at the Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, 2500 Speedway, MBB 3.210, Austin, Texas, 78712, USA.
e-mail: marcotte@icmb.utexas.edu

First, smash your peptide and weigh each piece

To better understand the algorithms involved, let's start by considering how peptides are analyzed using a mass spectrometer. In the typical approach, the mass of the peptide is measured first (more properly, a mass spectrometer actually measures the mass/charge ratio; in the common electrospray ionization approach, most charges are limited to +1, +2 or +3). As a peptide's mass is often not unique—for example, two peptides with the same amino acids in different order will share the same mass—we need information about the amino acid sequence. Therefore, the peptide is isolated in the mass spectrometer and energetically excited, usually by collision with inert gas molecules, to break it into smaller fragments. The mass/charge ratios of these smaller fragments are measured, and the resulting fragmentation mass spectrum (called an MS/MS or MS^2 spectrum, which effectively is just a list of the counts of each of the fragment ions) provides a fingerprint characteristic of that peptide's amino acid sequence.

The fragmentation process is reasonably well behaved. In the most common mode of operation, the peptide breaks along its backbone between adjacent amino acids. Such random breaks across a population of identical peptides create a sequencing ladder of peptide fragments differing in size by successive removal of amino acids. The peptide sequence can ultimately be determined from this ladder. For a number of reasons, however, directly reading the sequence turns out to be more complicated than one might expect.

First, unlike DNA sequencing, two ladders are created—one starting from the amino terminus and one from the carboxyl terminus—and both ladders are represented in the MS/MS spectrum. Second, there are three chemical bonds connecting consecutive amino acids; different experimental approaches favor breakage at different bonds (**Fig. 1a**). Breakage at the bond between the alpha carbon and the carbonyl carbon produces the *a*-ion and *x*-ion series (denoting peptide fragments containing the amino-terminal amino acid and the carboxy-terminal amino acid, respectively). Breakage at the carbonyl carbon-amide nitrogen bond produces the *b*- and *y*-ion series (as in **Fig. 1b**), and breakage at the amide nitrogen-alpha carbon bond produces the *c*- and *z*- ion series. **Figure 1c** shows such a ladder of peptide fragments in the fragmentation spectrum of the peptide NQWFFSK (one-letter amino acid code). The particular linear ion-trap mass spectrometer protocol that produced this spectrum tends to favor *b* and *y* ions. As a third complication, note that there are other peaks in the spectrum besides the *b* and *y* ions, such as *b* and *y* ions that have lost ammonia (b–NH_3, y–NH_3) or water (b–H_2O, y–H_2O), as well as occasional *a* ions and additional peaks, not all of which are easily explained. Finally, the fragments have widely varying abundances, stemming from different efficiencies of bond breakage and fragment production.

Then, find your peptide in a database based on its fragments' masses

Although we might prefer to sequence peptides directly from the MS/MS mass differences[6], because of the many confounding factors (e.g., multiple ion series, noise, missing peaks and additional peaks), an alternative approach is usually taken[7]: given an organism's genome sequence, we can computationally identify which proteins, and therefore which peptides, could in principle be present. Then, using the rules of peptide fragmentation as we currently understand them, a database is created of predicted MS/MS spectra (e.g., see the predicted fragmentation spectrum of the peptide NQWFFSK in **Fig. 1d**).

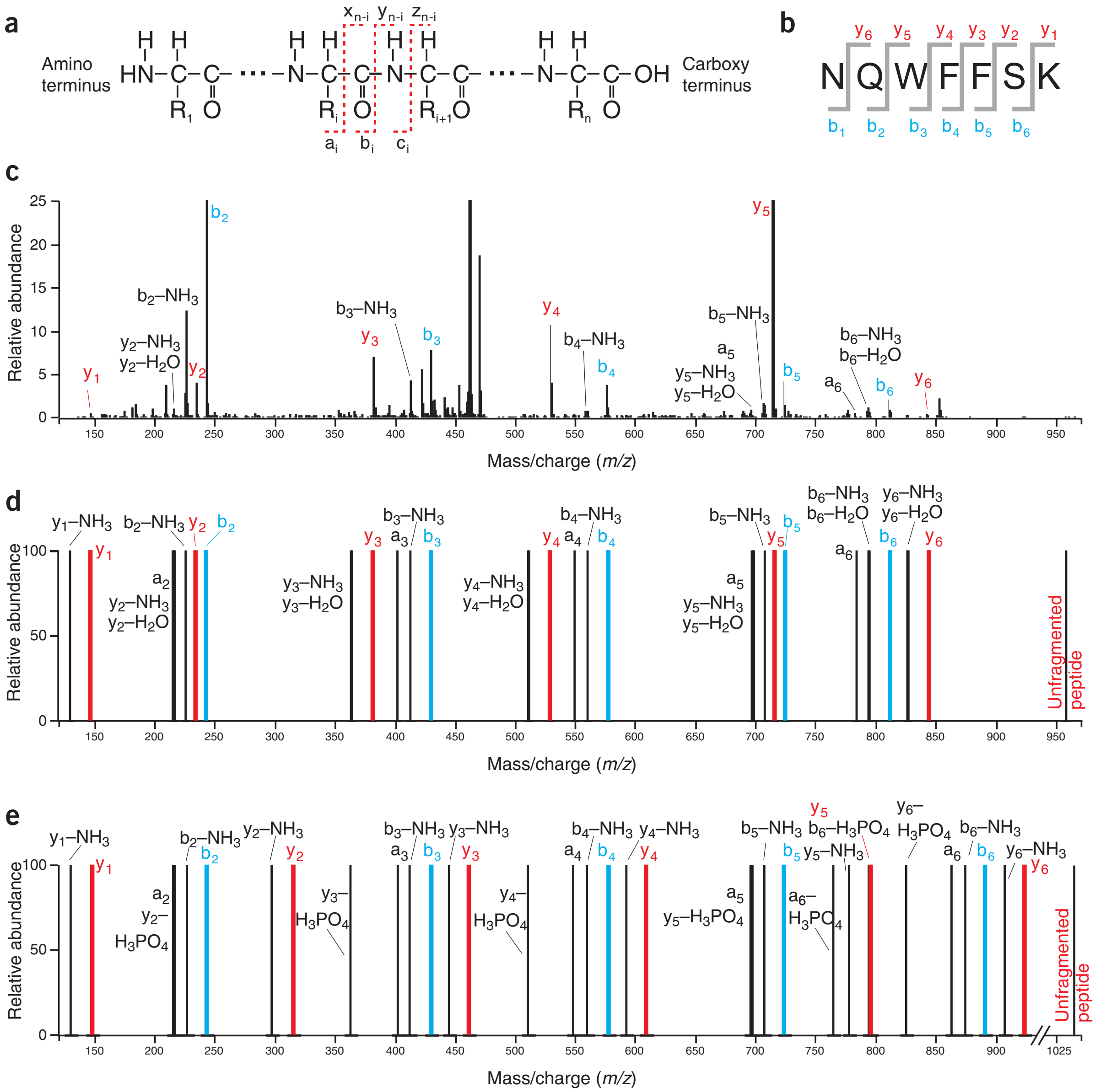

Figure 1 Shotgun proteomics identifies proteins from the fragmentation mass spectra of their constituent peptides. (**a**) Peptides are broken into smaller fragments in the mass spectrometer, producing families of fragments of differing masses, as described in the text. (**b**) The *b*- and *y*-ion series, generated by breaking peptides within peptide bonds, are commonly observed in ion-trap mass spectrometers. For example, fragmenting the peptide NQWFFSK between W and F produces the b_3 ion NQW (mass 429.19) and the y_4 ion FFSK (mass 528.28). (**c**) The resulting mass ladder of many such fragments can then be measured by the mass spectrometer, shown here in an experimental MS/MS spectrum of the peptide NQWFFSK. (**d**) The experimental spectrum is identified by computationally matching it to predicted MS/MS spectra, such as the one shown here for NQWFFSK. A typical database might contain MS/MS spectra predicted for all tryptic peptides from all proteins encoded by a particular genome sequence. Although experimental fragments clearly have varying abundances, predicted MS/MS spectra may not, depending on the methods used. The rules governing fragment abundance are only generally understood[8]. (**e**) Phosphorylation of the serine in NQWFFSK increases the mass of all serine-containing fragments by 79.97, as shown in this predicted MS/MS spectrum. For example, the b_6 fragment NQWFFS shifts from mass 810.36 to mass 890.32, whereas the b_1 to b_5 fragments remain unchanged. Also, 'neutral loss' ions might now be observed in which the phosphate group is removed during mass spectrometry (e.g., y_2–H_3PO_4).

Unfortunately, our understanding of peptide fragmentation is incomplete[8], and therefore our models for what to expect can still be improved, as a comparison of the spectra in **Figure 1c** and **d** shows. Nonetheless, the experimental MS/MS spectra are compared to these predicted spectra, and the best matches are found that meet minimum criteria for statistical significance. In this manner, the experimental spectra are associated with peptide sequences. Because we know which protein sequences contain these peptide sequences, we have therefore also identified the proteins. The support for each protein is based on the composite evidence for its component peptides.

How is this comparison of experimental and predicted MS/MS spectra performed? As one might wish to compare many experimental spectra (perhaps 50,000–100,000 MS/MS spectra in a typical shotgun proteomics experiment) to many predicted spectra (e.g., ~900,000 predicted peptides for the ~5,800 proteins of yeast, if one takes into account the fact that the protease used to make the peptides may occasionally miss a cleavage site or two), the method should be relatively fast.

In the simplest approach, which is used by the TurboSequest program (http://fields.scripps.edu/sequest/) distributed with ThermoFinnegan (San Jose, CA, USA) mass spectrometers, the similarity between experimental and predicted spectra is calculated as the background-corrected, cross-correlation function of the spectra[7], restricting the comparison to peptides with predicted overall masses close to that of the peptide under experimental analysis. Many other approaches have been developed, such as using postprocessing filters to reduce the false-positive identifications by TurboSequest[9], or probabilistic scores for matching spectra (as in the programs Mascot[10] (http://www.matrixscience.com/search_intro.html) and X!Tandem (http://www.thegpm.org/TANDEM/)). Some approaches take into account the abundances of peaks in the MS/MS spectra, whereas some consider only their positions. In spite of this variety of algorithms, the 'dirty little secret' of shotgun proteomics is that <20% of MS/MS spectra from typical experiments are successfully identified—although not all of the spectra are likely to be interpretable (or even to correspond to peptides), there is still a need for improved database-matching algorithms.

Given that the technique ultimately relies on comparisons to a database of predicted spectra, how do we know if the identifications are correct? This is a fundamental issue for mass spectrometry. One popular and reasonably effective approach is to shuffle the order of amino acids in each protein encoded by the genome, generating a database of predicted spectra from these shuffled sequences. Using this database instead of the real database should drastically reduce the number of peptides identified; from the number of peptides identified at a given scoring threshold in the shuffled versus real database, a false-positive identification rate can be calculated.

A second popular approach for estimating error involves using a classification algorithm that can recognize the characteristics of correct and incorrect matches[9], analyzing such features as the overall score and the difference in scores between the top match and the next best match. By training on spectra from peptides of known identity, the algorithm can estimate probabilities of correct identifications when applied to samples of unknown identity. However, readers should note that a low probability of identification indicates only a lack of evidence. Because of imperfect database matching, the somewhat stochastic acquisition of MS/MS spectra and other such issues, one should be cautious in interpreting failure to identify a peptide as evidence for its absence.

Modifications changing peptide mass complicate database searching

One area in which shotgun proteomics shows particular promise is the large-scale identification of post-translational modifications of proteins. This is important, as most cellular proteins are modified at some point, whether enzymatically (e.g., phosphorylation, ADP ribosylation or ubiquitination) or sporadically (e.g., oxidation). Luckily, although there are exceptions (e.g., certain types of glycosylation), modified peptides can often be observed by mass spectrometry with high efficiency. Mass spectrometry, after all, detects single molecules. As each individual copy of a peptide contributes independently to the counts of ion abundance, a mixture of modified and unmodified peptides doesn't present an intrinsic detection problem. Nonetheless, the identification of post-translational modifications has proven difficult. Given the database lookup approach to shotgun proteomics, it should now be obvious why this is the case—modifications change the mass of all fragments containing them (as in **Fig. 1e**), and the database has to be modified to anticipate these changes.

The fundamental difficulty is therefore that peptides are identified by comparing experimental to expected spectra, and our analysis must be modified to take into account the mass differences from post-translational modifications. One solution would seem to be to simply add spectra for modified peptides to the database. When only a single modification is considered, especially if it is universal (e.g., all cysteines are alkylated), this strategy works well. Even so, because there are >200 naturally occurring post-translational modifications[11], each of which can occur in combination with others, one can't account for all possible combinations of modifications without the database quickly becoming unmanageable. For example, allowing any two modifications per peptide (out of >200 possible) results in a database that is roughly 200*200, or ~40,000, times larger. And, as the number of false positives grows with the size of the database searched (more comparisons means more chances for a false match), the search efficiency drops with each additional modification considered. Thus, ironically enough, although there isn't a significant problem for the mass spectrometer to observe most post-translational modifications, there is a strong computational barrier to identify them.

Several solutions have been suggested, including using dynamic programming (the basis of protein-sequence alignments, among many other applications) to align experimental and predicted spectra while allowing post-translational modifications[12]. But the generalized identification of post-translational modifications by shotgun proteomics is still largely an unsolved problem, ripe for exploration and progress by budding computational biologists.

COMPETING INTERESTS STATEMENT

The author declares no competing financial interests.

1. Gavin, A.C. *et al. Nature* **440**, 631–636 (2006).
2. Krogan, N.J. *et al. Nature* **440**, 637–643 (2006).
3. Foster, L.J. *et al. Cell* **125**, 187–199 (2006).
4. Decramer, S. *et al. Nat. Med.* **12**, 398–400 (2006).
5. Hunt, D.F., Yates, J.R. 3rd, Shabanowitz, J., Winston, S. & Hauer, C.R. *Proc. Natl. Acad. Sci. USA* **83**, 6233–6237 (1986).
6. Fischer, B. *et al. Anal. Chem.* **77**, 7265–7273 (2005).
7. Eng, J., McCormack, A.L. & Yates, J.R. 3rd *J. Am. Soc. Mass. Spectrom.* **5**, 976–989 (1994).
8. Wysocki, V.H., Resing, K.A., Zhang, Q. & Cheng, G. *Methods* **35**, 211–222 (2005).
9. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. *Anal. Chem.* **74**, 5383–5392 (2002).
10. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. *Electrophoresis* **20**, 3551–3567 (1999).
11. Gooley, A. & Packer, N. in *Proteome Research: New Frontiers in Functional Genomics* (eds. Wilkins, W. *et al.*) 65–91 (Springer, New York, 1997).
12. Tsur, D., Tanner, S., Zandi, E., Bafna, V. & Pevzner, P.A. *Nat. Biotechnol.* **23**, 1562–1567 (2005).