Published in final edited form as:

J Proteome Res. 2010 June 4; 9(6): 3270–3279. doi:10.1021/pr100205k.

Identification of Tandem Mass Spectra of Mixtures of Isomeric Peptides

Xi Chen¹, Paul Drogaris², and Marshall Bern³

¹Dept. of Industrial Engineering, U. Washington, Seattle, WA 98195

²Dept. of Chemistry, University of Montreal, Montreal, QC H3C 3J7

³Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304

Abstract

Shotgun proteomics separates peptides by chromatography and precursor mass over charge, yet in almost any large data set of a complex sample there will be some tandem mass spectra containing more than one peptide. These mixture spectra contain two co-eluting peptides with close precursor mass over charge, and sometimes contain exact isomers, often the same peptide with the same modification in two different positions. Isomers present a problem when the position of the modification is of special interest, as in histone modification studies or “oxidative footprinting” studies of protein structure. Here we give algorithms for identifying isomeric mixtures, and present results on two different histones and four oxidative footprinting targets. Five of the six targets contain at least one peptide that appears in isomeric mixtures, but in none of the cases are mixtures so prevalent that they greatly impact the overall identification rate.

Keywords

oxidative footprinting; hydroxyl radical surface mapping; histone; acetylation; post-translational modification; modification site localization; propionic anhydride

Introduction

Mixture or “chimeric” spectra are surprisingly common in shotgun proteomics experiments. It has been estimated that in standard liquid chromatography tandem mass spectrometry (LC-MS/MS) pipelines, as many as 10 – 20% of the MS/MS spectra from a complex biological sample contain two or more coeluting molecular species^{1 2 3}. Simple samples, containing only a few proteins, typically produce many fewer mixture spectra, but if the proteins in the sample are heavily modified, the number of mixture spectra can again grow large. Moreover, mixture spectra in heavily modified simple samples are often exact isomers, for example, the same peptide carrying the same modification in two different locations, and hence are especially hard to recognize as mixtures, because the two components may have many fragmentation peaks in common. The goal of the work reported here was to determine the prevalence of isomeric mixtures, and the feasibility of identifying them. We consider two important cases of heavily modified simple samples: a biological sample enriched in histones, with roughly 30 detectable proteins and fairly good coverage (over 50%) of histones H4, H3, and H2A, and “oxidative footprinting” studies⁴ of nominal

Address Correspondence to: Marshall Bern, Ph.D., Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304, bern@parc.com, Voice: (650) 812-4443, Fax: (650) 812-4471.

one-protein samples, containing only small amounts of contaminant proteins. In both cases, the exact positions of modifications are of special interest.

In histones, there are digestively producible peptides with multiple modification sites, for example, R.GKGGKGLGKGGAKR.H near the N-terminus of histone H4, which has four potential post-translational modification sites, the four lysines. There are a number of possible isomers, for example, four singly acetylated forms of this peptide, and different isomers have different biological functions.⁵ Ideally we would like to determine how much of each isomer is present in each MS/MS spectrum. If this is not possible, we would at least like to know which isomers are possible and which can be ruled out.

Oxidative footprinting potentially presents an even more difficult problem than histones. Oxidative footprinting, more precisely termed “hydroxyl radical surface mapping”, correlates oxidation rates with solvent accessibility. In this structural biology technique, most amino acid residues are potential modification sites⁴, and indeed residue-level resolution of solvent accessibility is one of the advantages of the method over comparable techniques such as hydrogen/deuterium exchange or covalent labeling of lysines⁶. The abundance of modification sites, however, poses a challenge for data analysis, because there may be 100s of ways to add 31.990 Daltons (that is, a net change of two oxygen atoms) to a peptide. As in the histone case, we would ideally like to identify and quantify each isomer in each MS/MS spectrum, but more realistically, we would settle for less exact information, because the final goal is not exact peptide identification, but rather differential analysis of solvent accessibility in two different conditions, an analysis that can tolerate a certain amount of noise in the identifications.

There is already a data analysis tool called SLoMo⁷ that is specialized for modification site localization, and other tools that can localize phosphorylations^{8,9}, but to our knowledge none of these tools can identify mixture spectra. There are also some data analysis tools, such as ProbIDTree³ and ByOnic¹⁰, that can identify two peptides per MS/MS spectrum by removing peaks explained by a first identification, and then making a second identification from the remaining peaks. We have used this feature of ByOnic for data-independent MS/MS¹¹, which fragments all precursors within a wide m/z window, regardless of peak content, and hence produces a large number of mixture spectra. ByOnic’s algorithm works well enough for two completely different peptides, and even for some isomeric peptides, but it fails if the removal step leaves too few peaks to make a second identification by ordinary database search. For example, a first identification of GK[+42]GGKGLGKGGAKR would knock out all but six b- and y-ion peaks of GKGGK[+42]GLGKGGAKR, leaving too few peaks for a confident second identification. However, if we did observe two b2 ions (one with and one without +42 Daltons), two b3 ions, and so forth, we could be quite sure that both isomers are present, and we might even be able to estimate the relative abundance of the isomers from the ratios of peak heights or areas.

Several groups^{12–14} have developed tools for identifying mixtures of histone isomers from profile-mode ECD or ETD (electron capture and electron transfer dissociation) spectra. Moreover, Phanstiel et al.¹³ use least squares to estimate the relative abundance of up to four isomers per spectrum, assuming that peak intensities are unaffected by modifications, that is, the c5 ion of CK[+42]GGK has the same intensity as the c5 ion of CKGGK[+42], assuming equal isomer abundance. Dimaggio et al.¹⁴ make the same assumption, and use mixed integer linear optimization to estimate the relative abundance of any number of isomers.

Centroided collision-induced dissociation (CID) spectra are likely to present a more difficult quantification problem than ECD or ETD, because CID fragment intensities generally show

more dependence upon residue side chains and modifications, and the centroiding (“peak picking”) process retains only local maxima, keeping peak heights but losing peak areas. Due to the challenges posed by centroided CID spectra, especially CID spectra of oxidative footprinting samples, we pursued a less ambitious goal than relative quantification of an arbitrary number of isomers. We developed algorithms and software to identify up to two peptides or modification forms per spectrum, and to classify each identified spectrum as

- Pure if there is separate evidence for only one form,
- Possible mixture if there is weak separate evidence for each of the two forms,
- Probable mixture if there is strong separate evidence for both forms, or
- Uncertain if there is no separate evidence for either form.

In the case of mixture spectra, we tested several simple algorithms for relative quantification of the two peptides. We did not attempt to identify mixtures of three or more peptides, because the case of two peptides is already difficult, and because two isomers often cover the third, leaving it with no unique peaks. For example, GK[+42]GGKGLGKGGAKR and GKGGKGLGK[+42]GGAKR together include all the peaks of GKGGK[+42]GLGKGGAKR, so that the presence of the hidden third isomer would be indicated by peak intensity ratios alone, for example, larger $b5' / b5$ than $b4' / b4$, where $b5'$ means the intensity of the peak including the +42 Da shift. We tested our algorithms on six proteins, two acetylated histones and four oxidative-footprinting targets. We find that five of the six proteins contain at least one peptide that gives mixture spectra, and that overall identification results can be improved, but not greatly improved, by consideration of mixture spectra.

Materials

We used MS/MS spectra of a sample that had been LC-fractionated before digestion to enrich for histone H4. The sample was treated with histone deacetylase (HDAC) inhibitor to preserve lysine acetylation, and was derivatized with propionic anhydride to propionylate free and mono-methylated lysines and block trypsin digestion at lysine15. The sample was then digested with trypsin and analyzed by LC-MS/MS on a Thermo LTQ Orbitrap, using Orbitrap MS scans and LTQ MS/MS scans.

We used oxidative footprinting MS/MS spectra from two sources: *E. coli* OmpF porin (outer membrane protein F) from Siu Kwan Sze (Nanyang Technical University, Singapore), and chicken lysozyme, horse myoglobin, and bovine ubiquitin from Carlee McClintock (Oak Ridge National Laboratory). The OmpF data were from a cell sample with *in vivo* oxidation by Fenton chemistry as described previously.¹⁶ The other data were derived from commercially available pure proteins (with trace contamination from human keratin and sampling carryover), electrochemically oxidized using a boron-doped diamond electrode¹⁷. All samples were digested with trypsin. OmpF was analyzed by LC-MS/MS on a Thermo LTQ FT instrument with FTICR single-MS and LTQ MS/MS. Lysozyme and ubiquitin were analyzed by LC-MS/MS with Orbitrap MS scans and LTQ MS/MS scans. Myoglobin was analyzed by LC (four offline SCX fractions)/LC-MS/MS with LTQ for both single- and tandem-MS.

Methods

We wrote a new peptide-scoring program named Xim. This program takes as input a tandem-MS spectrum, a small number of possible peptides without modifications, and a set of modification rules. Xim first preprocesses the spectrum to select a set of 50 – 100 **good peaks**. The preprocessing reweights peaks according to “local intensity” by comparing them

with other peaks with similar mass over charge (m/z). It upweights peaks that appear to be monoisotopic peaks within peak series, and downweights peaks that appear to be isotope peaks or water losses from other intense peaks. We have described this preprocessing elsewhere (see the supplemental information for ByOnic10).

Modification forms for Xim are specified by a set of rules. A rule such as “M $\rightarrow$ M[+15.995]” sets a variable modification: any methionine can be modified to oxidized methionine. The rule “N-terminal Q $\rightarrow$ Q[−17.027]” applies only to N-terminal glutamines. A fixed modification such as carbamidomethylated cysteine is set by simply changing the mass of cysteine in the table of residue masses. Xim includes one nonstandard feature: modifications can be designated as common or rare, and at most one rare modification of any kind is allowed per modification form. For example, “X $\rightarrow$ X[+21.982], rare” specifies sodiation of any residue, and the designation “rare” sets a realistic limit of one sodiation. The pair of rules “X $\rightarrow$ X[+21.982], rare” and “M $\rightarrow$ M[+31.990], rare” allows a peptide to have either sodiation or methionine sulfone, but not both. Modification forms are generated all at once for all observed precursor mass shifts (e.g. peptide mass + protons +16, peptide mass + protons + 32, etc.) of a certain peptide, and those that match the precursor mass shift of each individual spectrum within a tolerance are considered proper candidate forms for that spectrum. The default precursor mass tolerance is 2 Daltons, which fits the isolation window of the Thermo LTQ and Thermo LTQ Orbitrap instruments, but the tolerance can be adjusted for other instrument types. Of course, Orbitrap precursor measurements give much greater accuracy than 2 Da, but only one mixture component needs to match the measured precursor.

Xim scores each modification form of each peptide by a **total score**, which integrates over all the good peaks. In addition, it computes a **separation score** for the top 2 distinct modification forms P and P'. The separation score for modification form P integrates over all the **separation peaks**, meaning those good peaks explained by P that are not explained by P'.

For the total score, we used X!Tandem's hyperscore18. The hyperscore $H(P,S)$ for peptide P and spectrum S is $H(P,S) = N_b!N_y! \text{MatchInt}(P,S)$, where N_b is the number of b-ions of P matching good peaks in S, N_y is the number of y-ions matching good peaks in S, and $\text{MatchInt}(P,S)$ is the sum of the intensities of all the good peaks in S matching ions of P. A good peak is considered matched if its mass over charge (m/z) is within a tolerance (usually 0.4 Da) of a theoretical ion. We included doubly charged y-ions as well as singly charged ions when counting N_b and N_y . The factorials in the expression for hyperscore emphasize b- and y-ion peak series rather than separate peaks. For the separation score, we simply counted the number of good peaks explained by one of P and P' but not the other.

For validation we implemented Ascore8, which gives a probability score (p-value) based on a Bernoulli trial model of random peak matching. The chance of matching k good peaks to a set of N theoretical peaks is given by $(N \text{ choose } k) p^k(1-p)^{N-k}$, where p is the chance of matching a single peak at random, which can be derived from the number of observed peaks, the mass range, and the mass matching tolerance. Ascore sums this expression over all values of k from the actual number of matched peaks n up to the maximum value N, in order to estimate Pvalue, the probability of obtaining at least n matched peaks by chance. The classifier described below used X!Tandem's hyperscore for the total score and simple peak counting for the separation score. These choices gave sensitive, robust, and human-interpretable results, with sufficient sensitivity for the purposes of our study.

Xim classifies each identified spectrum (that is, each spectrum whose top modification match has sufficiently high hyperscore) into four classes based on the separation score

(number of separation peaks chosen as good peaks). The four classes are: pure, possible mixture, probable mixture, and uncertain. For an identification of modification form P, Xim computes a real-valued threshold $T = \max \{ 1.0, N_s / 10.0 \}$, where N_s is the number of theoretical separation peaks (b- and y-ions only) in P relative to P'. For example, if P contains 14 b- and y-ions not also found in P', then $T = 1.4$. If the separation score for P exceeds T, then P has evidence, and if the separation score exceeds 2T, then P has strong evidence. A **pure** spectrum has one modification form with strong evidence and the other with no evidence; a **probable mixture** has strong evidence for both forms; a **possible mixture** has evidence for both forms but not both strong; and an **uncertain** identification has evidence for neither form.

For spectra classified “mixture”, we also experimented with relative quantification of the two identified components. The two components of an identified mixture usually have separation peaks in one-to-one correspondence. For example, GK[+42]GKGGLGKGGAKR has six separation peaks (b2 – b4 and y10 – y12) with respect to GKGKK[+42]GLGKGGAKR, and vice versa. One simple algorithm for relative quantification computes the ratios of the heights or areas of corresponding peaks (b2 versus b2', b3 versus b3', and so forth), and sets the ratio of the two components to be the median of these ratios. Missing peaks give ratios of zero, infinity, or zero divided by zero, but these should not cause a problem so long as there are not too many missing peaks. Another simple algorithm computes the sum of peak heights or areas over all separation peaks and then takes the ratio. This algorithm is also applicable to the case of two distinct peptides in the same mixture, that is, peptides with separation peaks that are not in one-to-one correspondence.

Xim is not a full-functionality search engine like Mascot or SEQUEST, but rather an adjunct search engine like P-Mod19, because it expects input containing only a few peptides, at least one of which, properly decorated with modifications, is correct. To obtain a list containing only a few peptides (in fact, usually only one peptide), we used high-accuracy precursor masses and/or ByOnic10, a database-search tool similar to Mascot20 and SEQUEST21. Unlike Mascot or SEQUEST, ByOnic includes specialized support for oxidative footprinting, so that its oxidative-footprinting identifications are relatively good, but Xim discards ByOnic's placement of modifications and scores modification forms anew, so Xim would work almost equally well with any search engine as a front end.

Results

We first describe an experiment to estimate Xim's error rate. Then we present results on isomeric mixtures of distinct peptides, the same peptide with different locations of post-translational modifications, and the same peptide with different oxidative labels.

False Positive Rate

We selected ~320 spectra from OmpF, myoglobin, lysozyme, and histones that ByOnic had confidently identified as unmodified, fully tryptic peptides. These spectra are very likely to contain only one peptide each, because in simple protein mixtures, the chance of chimeric spectra in which the primary component is unmodified is small. We gave Xim four possible identifications for each spectrum: the correct (ByOnic) peptide along with three fictitious peptides formed by transposing two amino acid residues of the correct peptide. For example, if the correct peptide is LSYDTEASIAK, then there are nine distinct isomers formed by swapping residues two residues apart: YSLDTEASIAK, LDYSTEASIAK, and so forth. We discarded duplicate and near-duplicate isomers resulting from swaps such as I/L, Q/K, and N/D with masses within 2 Da, and randomly selected three fictitious sequences from the remainder. Xim's real problems range from 2 to 6 isomers per histone peptide, and from 2 to

~100 for oxidative footprinting, depending upon the number of modification sites considered and the total mass of modifications.

Figure 1 shows a plot of Xim's classifications as a function of the distance between the two swapped residues. Regardless of distance, Xim classified over 90% of the spectra as pure or uncertain, with uncertain the more likely classification at swap distance one (adjacent residues) and pure more likely at larger distances. Possible mixture classifications ranged from 0 to 8%, and probable mixtures from 0 to 1%. Xim also made ~5% false identifications, not shown in Figure 1, in which the top scorer was a fictitious sequence, rather than the correct sequence; false identifications are most often classified as uncertain and very few as pure.

Two Distinct Peptides

One of the oxidative footprinting targets, horse myoglobin, fortuitously contains a pair of isomeric tryptic peptides, due to a pair of lysine repeats with 13 intervening residues: KKHGTVVLTALGGILKKK. The two 15-residue peptides starting after the initial two lysines apparently co-elute. ByOnic identified six spectra as containing peptide P = HGTVVLTALGGILKK, and after knocking out the peaks of this peptide, ByOnic identified one or two of the knockout spectra as also containing P' = KHGTVVLTALGGILK, depending upon score thresholds and size of the search. Figure 2 shows the spectrum that ByOnic most confidently identified as a mixture; this spectrum has a ByOnic score of 698 (approximately equivalent to Mascot 70) for P and 327 for P' (approximately equivalent to Mascot 33). Dynamic range is an issue here, because P has about 5 times the total intensity of P'. Here we count the total intensity as the sum of the heights of all identified peaks, including neutral losses, a-ions, and doubly charged y-ions. The total intensity ratio roughly agrees with ratios obtained from "corresponding" peaks. For example, the tallest peak of a singly charged b- or y-ion from P is y6 and it has about 5 times the intensity of y5', the tallest b- or y-ion peak from P'. These peaks both reflect the favored cleavage after leucine.

Xim proved more sensitive than ByOnic at identifying this co-eluting pair, classifying five of the six spectra as probable mixtures and one as a possible mixture. To be fair, Xim has an easier task than ByOnic, because Xim scores only two peptides and ByOnic scores thousands. Xim's classifications agree with manual inspection, but of course the border between probable mixture and possible mixture is fuzzy. Xim's evidence for the second identification ranged from 4 observed b- and y-ion separation peaks (out of a possible 28), up to 14 b- and y-ion peaks. The second identification was KHGTVVLTALGGILK three times, HGTVVLTALGGILKK once, and the two peptides were approximately tied twice. Much more common than either isomer is HGTVVLTALGGILK, a tryptic peptide with no missed cleavages.

Histones

We show results for three histone peptides: GKGGKGLGKGGAKR from the tail (residues 4 – 17) of histone H4, KTVTAMDVVYALKR from the middle (residues 79 – 92) of histone H4, and AGGKAGKDSGKAKAKAVSR from the tail (residues 2 – 20) of histone H2A. Histone modifications are a subject of intense study, and hence the possible modification forms are relatively well understood.

Figure 3 shows a well-identified mixture spectrum, containing two different triply acetylated forms: P = GK[+42]GGK[+56]GLGK[+42]GGAK[+42]R and P' = GK[+42]GGK[+42]GLGK[+56]GGAK[+42]R. Lysines that were unmodified or singly methylated before the addition of propionic anhydride will generally be propionylated and appear as K[+56] or K[+70] respectively. Dimethylated and acetylated lysines will not be

propionylated, and will appear as K[+28] and K[+42.011], respectively. Lysine can also be trimethylated, K[+42.047], but we could rule out this possibility, both because the lysines in this peptide are known acetylation sites, and because the Orbitrap precursor mass errors were below ~5 ppm, accurate enough to distinguish trimethylation from acetylation on a peptide of mass 1453 Da.

The ratios of heights of corresponding separation peaks in Figure 3 are in fairly good agreement. The y-ions agree quite closely, $y_6/y_6' = 47/19 \sim 2.5$, $y_7/y_7' = 236/81 \sim 2.9$, $y_8/y_8' = 46/17 \sim 2.7$, and $y_9/y_9' = 178/70 \sim 2.5$. The b-ions agree less closely, with $b_5/b_5' = 40/30 \sim 1.3$, $b_6/b_6' = 16/0 = \infty$, $b_7/b_7' = 52/26 = 2.0$, and $b_8/b_8' = 0/0$. Excluding ∞ and 0/0, the P/P' peak ratios have mean 2.38 and standard deviation 0.54. We discuss the implications of these numbers in Conclusions.

Altogether the data set included 35 spectra with Orbitrap precursor masses consistent with modification forms of GKGGKGLGKGGAKR. Of these, ByOnic identified 24 as modification forms of GKGGKGLGKGGAKR. Xim classified 10 of the 35 spectra as probable or possible mixtures. Table 1 lists the six spectra for which Xim found at least two separation peaks, by descending order of “mixture-ness”, that is, the real-valued separation score of the second identification. Score and Sep Score in Table 1 are $-\log_e(\text{Pvalue})$, where Pvalue is computed using Ascore's formula above with $p = 0.8$, so that a Sep Score of 3.0 should arise by chance with probability about $\exp(-3.0) = 0.05$. The spectrum shown in Figure 3 is the one on the second row of the table. Xim found sufficiently strong separation peaks for 1 b-ion and 3 y-ions to support the second identification of GK[+42]GGK[+42]GLGK[+56]GGAK[+42]R, encoded as 42.42.56.42 in the table. As shown in Figure 3, when all peaks are taken into account, not just locally intense “good peaks”, there are actually 2 b-ion separation peaks (b_5' and b_7') and 4 y-ion separation peaks (y_6' , y_7' , y_8' , and y_9'), so reported Ascore Pvalues may be underestimates of statistical significance due to imperfect peak selection.

Altogether the data set included 12 spectra with Orbitrap precursor masses consistent with modification forms of AGGKAGKDSGKAKAKAVSR from Histone 2A. Both ByOnic and Xim could obtain first identifications for all 12 spectra. Xim found only one spectrum with 2 or more separation peaks for a second identification. This spectrum had P = AGGK[+70]AGK[+70]DSGK[+56]AK[+56]TK[+56]AVSR as its first identification and P' = AGGK[+42]AGK[+42]DS[+56]GK[+56]AK[+56]TK[+56]AVSR as its second identification, with one b-ion and one y-ion separation peak. These two peptides are not exact isomers ($+70 = 70.0417$, $+42 = 42.0106$, and $+56 = 56.0261$), but they are within 0.04 Da, so both would be fragmented together. Manual inspection, however, identified this spectrum as pure P* = [+56]AGGK[+42]AGK[+42]DSGK[+56]AK[+56]TK[+56]AVSR. Xim had not considered propionylation of the N-terminus, and the separation peaks for Xim's top two modification forms are actually peaks of P*, for example, P* and P share a b_8 peak at 710, and the +1 isotope peak for b_5 of P* coincides with b_6 of P'.

Altogether the data set included 24 spectra with Orbitrap precursor masses consistent with modification forms of KTVTAMDVVYALKR. Xim identified these as K[+56]TVTAMDVVYALK[+56]R, K[+56]T[+56]VTAMDVVYALK[+56]R (which may really be [+56]K[+56]TVTAMDVVYALK[+56]R), K[+56]TVTAM[+16]DVVYALK[+56]R, and K[+56]TVT[+56]AMDVVYALK[+56]R. Xim found no probable mixtures.

Incidentally, ByOnic identified a number of apparent amidations ($D \rightarrow N$ and $E \rightarrow Q$ conversions) in the histone data set: ME[−1]ELHNQEVQK[+56]R from NonO protein, D[−1]AVTYTEHAK[+56]R from histone H4, and NK[+56]YE[−1]DEINK[+56]R from

keratin 1. Figure 4 shows K[+56]TVTAMD[−1]VVYALK[+56]R. Amidation may be a chemical artifact of propionic anhydride treatment^{11 17} or simply a result of high pH.

Oxidative Footprinting

In oxidative footprinting studies, almost every peptide has multiple modification forms, and likely total modification masses, such as +16, +32, and +48, can almost always be achieved in numerous ways. Table 2 gives ByOnic's list of oxidations; this list is compiled both from the literature²² and from our own empirical studies. Table 2 does not in fact represent the full complexity of the search, because many oxidative footprinting samples also include common non-oxidative modifications such as pyro-glu conversions, deamidation, and sodiation.

Examples from Olga Charvatova, Carlee McClintock, and others (shown in posters at ASMS, and currently in submission for journals) have demonstrated that recognizable mixture spectra occur in oxidative footprinting studies, and Xim provides a way to estimate their frequency. ByOnic is not a good tool for finding mixture spectra in oxidative footprinting studies, because the mixture components generally contain too many peaks in common. Here we present a survey of all the tryptic peptides in each of four proteins, and an experiment in relative quantification of two isoforms.

Full Protein Statistics—We show mixture identification results for four proteins: Myoglobin (Figure 6), Ubiquitin (Figure 7), Lysozyme (Figure 8) and OmpF (Figure 9). In each figure, (a) lists the number of spectra in each of the four categories (pure, uncertain, possible mixture, probable mixture), and (b) draws a bar chart of the above statistics on a unified scale for the whole protein. The table that labels a bar lists representative modification forms of “probable mixture” for each precursor mass group. For all four proteins we used a precursor mass tolerance of 2 Da and a rich set of modification rules: everything in Table 2, along with N-terminal pyro-glu, deamidation of N and Q, and sodiation on any residue. More restrictive tolerance or rules may result in fewer uncertain and possible mixture identifications, but we would like to find as many probable mixtures as we can, in order to assess the prevalence of the phenomenon.

Peptides give uncertain, possible, and probable mixture classifications in various ways. All the components of the probable mixtures of LFTGHPETLEK in myoglobin (Figure 6) are simply different locations of sodium adducts, simultaneously enabled by our rich set of rules. Most sodiated peptides give uncertain or possible mixture classifications, because there is not strong evidence to localize the sodium. For GLSDGEWQQVLNVWGK (Figure 6), the probable mixtures identified for four different mass shifts, +4, +16, +32, +36, all involve either [+4] or [+32] at one or both of the two tryptophan residues (as in Figure 5), combined with [+16] at various sites. These mixtures are more interesting than a mix of sodium adducts, because they give information about solvent accessibility. In general, uncertain classifications are most likely for sodiated peptides and for oxidized peptides with total mass shifts of +16, +32, or +48. Possible and probable mixture classifications are most likely for oxidized peptides containing well-spaced instances of the six most reactive residues (C, M, W, F, Y, H), such as GLSDGEWQQVLNVWGK (Figure 6), IVSDGNGMNAWVAWR (Figure 8), and YADVGSFDYGR (Figure 9). In lysozyme, the long peptide IVSDGNGMNAWVAWR exhibits various forms with mass shifts of +16, +22, +32 and +48, usually with the commonly seen M[+16] and W[+16] combined with [+16] or [+32] on other sites.

Spectra containing modification forms of HPGDGFADAQGAMTK (Figure 6) produce 752 classifications of uncertain, more than half of all the classifications. Many of these uncertain classifications occur because we allow H[−22] and sodiation (+22 Da) to coexist, so that

both the likely HPGDFGADAQGAM[+16]TK and the implausible H[−22]P[+22]GDFGADAQGAM[+16]TK score well. Excluding or penalizing two closely spaced opposite modifications in the same peptide causes 142 of the 752 uncertain classifications to switch to pure. The remaining uncertain classifications almost all involve +16 on AMTK, where few separation peaks, corresponding to low mass y-ions or high-mass b-ions, are observed. Most of these peptides probably contain M[+16] rather than the less likely A[+16], T[+16], or K[+16], but Xim uses no chemical knowledge to disambiguate the identifications. Besides −22/+22, another example of canceling positive and negative modifications occurs when S/T[−2] (oxidation followed by loss of water) and deamidation (N/Q[+1]) are simultaneously enabled. This combination of modification rules can give implausible identifications such as FES[−2]N[+1]FN[+1]TQATNR. To avoid these anomalies, one might turn off the non-oxidative modifications, but without sodiation enabled we observe spurious identifications combining one H[−10] with two nearby [+16]’s to give a mass shift of +22. An improved, but less pure, version of Xim would use special-case rules to reclassify uncertain classifications for which one identification is *prima facie* more plausible than the others, as judged by the number and proximity of modifications.

Relative Quantification—We performed an experiment in relative quantification of oxidized isomers using the spectrum shown in Figure 5 of peptide GLSDGEWQQVLNVWGK in Myoglobin. The two forms, P = GLSDGEW[+32]QQVLNVW[+4]GK and P' = GLSDGEW[+4]QQVLNVW[+32]GK, each have 14 distinct theoretical b/y-ions: b7–b13 vs. b7'–b13' and y3–y9 vs. y3'–y9'. Among all observed good peaks, 13 are explained by distinct ions of P (b7, b8, b10–b13, y3–y9) and 13 are explained by distinct ions of P' (b7', b8', b10'–b13', y3'–y9'). A peak with mass matching to y5-18 is also observed. Figure 10 shows the relative intensity of these peaks after preprocessing (the original intensities show similar relations) of separation peaks explained by both forms. The two intensities for the same ion roughly follow the same trend. Note that the largest deviations from this trend occur at b7/b7' and y9/y9', which correspond to a cleavage on the C-terminal side of the first modification site. Taking out these two outlier pairs of intensities, we obtain an average P / P' intensity ratio of 1.22 with standard deviation 0.55 (or, alternatively, ratio of the sum of intensities, 1.12).

Conclusions

We have found that automatically identifying mixtures of isomeric peptides is quite feasible. When given the right set of modifications, Xim’s ranking of “mixture-ness” agrees well with expert manual evaluation. We caution, however, that mixture-ness is a matter of degree, and when the second identification has much lower abundance, or when it has too few separation peaks, neither Xim nor manual inspection can decide the case with any certainty. Xim is surely improvable, but we found that its success is not highly dependent upon the details of the scoring function; its success, as judged by human experts, depends mainly upon the interactions among the modification rules. All the scoring functions we tried (peak count, Ascore, and hyperscore) worked almost equally well. In retrospect, this is not surprising, as Xim has a relatively easy search task, namely searching 10s or at most 100s of possible modification forms with the same precursor mass.

We also found that each protein has its own “personality”, due to its distribution of modification and trypsin cleavage sites. For example, myoglobin has the pair of KK’s that produce a mixture of distinct peptides; it also has the peptide GLSDGEWQQVLNVWGK, which contains two exposed tryptophan residues, and those are the only two highly reactive residues in the peptide. OmpF gives many uncertain spectra, but few possible or probable mixtures. This is partly due to its longer peptides and also partly due to its overall lower oxidation rate, because it was oxidized *in vivo* by Fenton chemistry rather than by fast

photo-oxidation and hydrogen peroxide. In this case, residue-level resolution of oxidations was not necessary to draw conclusions concerning the voltage-gating mechanism of the porin16.

Clear mixture spectra, for which the two modification forms have strong evidence and roughly equal abundance, are not quite as common as we expected before carrying out this study. Out of roughly 50 detected tryptic peptides in histone H4, H3, and H2A, only one peptide, GKGGKGLGKGGAKR from the histone H4 tail, gave clear mixture spectra.

Oxidative footprinting gave more possible and probable mixtures (~7% of all spectra over all four target proteins) than histones, but also a much larger number of uncertain spectra (~31% of all spectra), for which even the primary identification is in doubt. Some of the uncertainty can be eliminated by rejecting identifications with a canceling pair of modifications, and most of the remaining uncertainty is “minor”, meaning that the possible identifications differ only in the exact placement of an oxidation or sodiation. The overall level of mixtures and uncertainty is not so high as to invalidate the usual oxidative footprinting data analysis, which assumes that each spectrum contains a pure and identifiable modification form. We believe that identification of both components of mixture spectra, however, and “fuzzy” localization of oxidations within uncertain spectra down to an interval of, say, two to four residues, would improve the data analysis and give less noisy measurements of solvent accessibility by hydroxyl radical surface mapping.

Finally, we conclude that rough estimation of relative abundance of two isomeric mixture components is possible with centroided CID spectra. The ratios of corresponding fragment peak intensities tended to agree in all the mixture spectra we examined, but the coefficient of variation of the ratio is fairly high, 0.23 for Figure 3 (histone isomers) and 0.45 for Figure 5 (oxidation isomers), and the use of centroided rather than profile-mode spectra means that some relevant peaks are lost, and some peak ratios cannot be determined. The variation is large enough that detection and quantification of a hidden third component, such as GKGGK[+42]GLGKGGAKR in a mixture of GK[+42]GGKGLGKGGAKR and GKGGKGLGK[+42]GGAKR, would not be reliable if the third component constitutes less than ~20% of the ion current. This contrasts with the reported detection and quantification of up to four isomers, even hidden isomers, down to fractions below 10% from profile-mode ETD spectra.^{13, 14}

Acknowledgments

We thank Heinrich Anhold, Robert Hettich, Carlee McClintock, Siu Kwan Sze, and Pierre Thibault for access to data. This work was supported by NIH grant GM085718.

References

1. Hoopmann MR, Finney GL, MacCoss MJ. High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Analytical chemistry*. 2007; 79:5620–5632. [PubMed: 17580982]
2. Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical chemistry*. 2006; 78:5678–5684. [PubMed: 16906711]
3. Zhang N, et al. ProbiDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics*. 2005; 5:4096–4106. [PubMed: 16196091]
4. Xu G, Chance MR. Hydroxyl radical-mediated modification of proteins as probes for structural proteomics. *Chem Rev*. 2007; 107:3514–3543. [PubMed: 17683160]

5. Dang W, et al. Histone H4 lysine 16 acetylation regulates cellular lifespan. *Nature*. 2009; 459:802–807. [PubMed: 19516333]
6. Fitzgerald MC, West GM. Painting proteins with covalent labels: what's in the picture? *Journal of the American Society for Mass Spectrometry*. 2009; 20:1193–1206. [PubMed: 19269190]
7. Bailey CM, et al. SLoMo: automated site localization of modifications from ETD/ECD mass spectra. *Journal of proteome research*. 2009; 8:1965–1971. [PubMed: 19275241]
8. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature biotechnology*. 2006; 24:1285–1292.
9. Bern M, Goldberg D. Improved ranking functions for protein and modification-site identifications. *J Comput Biol*. 2008; 15:705–719. [PubMed: 18651800]
10. Bern M, Cai Y, Goldberg D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Analytical chemistry*. 2007; 79:1393–1400. [PubMed: 17243770]
11. Bern M, et al. Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Analytical chemistry*. 82:833–841. [PubMed: 20039681]
12. Pesavento JJ, Mizzen CA, Kelleher NL. Quantitative analysis of modified proteins and their positional isomers by tandem mass spectrometry: human histone H4. *Analytical chemistry*. 2006; 78:4271–4280. [PubMed: 16808433]
13. Phanstiel D, et al. Mass spectrometry identifies and quantifies 74 unique histone H4 isoforms in differentiating human embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105:4093–4098. [PubMed: 18326628]
14. DiMaggio PA Jr, Young NL, Baliban RC, Garcia BA, Floudas CA. A mixed integer linear optimization framework for the identification and quantification of targeted posttranslational modifications of highly modified proteins using multiplexed electron transfer dissociation tandem mass spectrometry. *Mol Cell Proteomics*. 2009; 8:2527–2543. [PubMed: 19666874]
15. Drogaris P, Wurtele H, Masumoto H, Verreault A, Thibault P. Comprehensive profiling of histone modifications using a label-free approach and its applications in determining structure-function relationships. *Analytical chemistry*. 2008; 80:6698–6707. [PubMed: 18671409]
16. Zhu Y, et al. Elucidating in vivo structural dynamics in integral membrane protein by hydroxyl radical footprinting. *Mol Cell Proteomics*. 2009
17. McClintock C, Kertesz V, Hettich RL. Development of an electrochemical oxidation method for probing higher order protein structure with mass spectrometry. *Analytical chemistry*. 2008; 80:3304–3317. [PubMed: 18351783]
18. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics (Oxford, England)*. 2004; 20:1466–1467.
19. Hansen BT, Davey SW, Ham AJ, Liebler DC. P-Mod: an algorithm and software to map modifications to peptide sequences using tandem MS data. *Journal of proteome research*. 2005; 4:358–368. [PubMed: 15822911]
20. 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]
21. Yates JR 3rd, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical chemistry*. 1995; 67:1426–1436. [PubMed: 7741214]
22. Xu G, Chance MR. Radiolytic modification and reactivity of amino acid residues serving as structural probes for protein footprinting. *Analytical chemistry*. 2005; 77:4549–4555. [PubMed: 16013872]

Figure 1.

Xim's classifications of pure spectra, for which the possible identifications were the correct peptide along with three different isomers, each with two amino acid residues transposed. The x-axis shows the distance between the two transposed residues, for example, 2 in the case of LSYDTEASIAK (correct), YSLDTEASIAK, LDYSTEASIAK, and so forth, and 3 in the case of DSYLTEASIAK, LTYDSEASIAK, and so forth. At distance one uncertain is the most common classification, but the correct classification predominates at distances two or greater. False possible-mixture classifications are uncommon (always below 10%), and probable-mixture classifications remain negligible throughout the range, never exceeding 1%.

Figure 2.

A spectrum containing a mixture of two overlapping isomeric peptides from Apomyoglobin, P = KK.HGTVVLTALGGILKK.K and P' = K.KHGTVVLTALGGILK.K. These peptides co-elute, and according to Xim every spectrum containing P also has evidence for P' and vice versa. In this spectrum, peptide P accounts for ~80% of total peak intensity and P' accounts for ~15%, so it is necessary to plot intensity on a log scale to see all the peaks of P'.

Figure 3.

A spectrum containing a mixture of two isomeric modification forms of histone H4 peptide GKGKGLGKGGAKR. Here [+56] is propionylation (a chemical derivative) and [+42] is *in vivo* acetylation. The two modification forms, P and P', are both triply acetylated, with P unacetylated at the second lysine and P' unacetylated at the third lysine. Separation peaks are shown with larger font than shared peaks. Notice that the peaks unique to P are all higher intensity than the corresponding peaks in P', although b5' with cleavage after K[+42] is almost as intense as b5 with cleavage after K[+56]. The only missing separation peaks in this centroided spectrum, which has only 53 peaks in its entire peak list, are b8, b6', and b8'.

Figure 5.

A spectrum containing a mixture of two isomeric modification forms of Apomyoglobin peptide GLSDGEWQQVLNVWGK. In this case, the two forms are of roughly equal abundance, as shown by the peak ratio comparison in Figure 9. Cleavage after W[+32] seems favored as b7 and y9 from P and b14 (small font) from both P and P' are especially strong, with y9 the strongest peak in the entire spectrum.

Figure 6.
Per-peptide mixture classification results for Apomyoglobin. The tables associated with the bar chart give lists of identifications for selected peptides.

position	peptide	pure	uncertain	possible mixture	probable mixture
0	-.MQIFVK.T	21	49	0	1
6	K.TLTGK.T	13	0	0	0
11	K.TITLEVEPSDTIENVK.A	291	168	28	18
33	K.EGIPPDQQR.L	21	0	0	0
42	R.LIFAGK.Q	62	14	4	1
42	R.LIFAGKQLEDGR.T	4	0	0	0
48	K.QLEDGR.T	0	1	0	0
48	K.QLEDGRTLSDYNIQK.E	2	0	0	0
54	R.TLSDYNIQK.E	21	3	0	0
63	K.ESTLHLVLR.L	130	55	5	5
	TOTAL	565	290	37	25

(a)

Peptide Statistics

(b)

Figure 7.
Mixture identification results for Ubiquitin.

position	peptide	pure	uncertain	possible mixture	probable mixture
1	K.VFGR.C	65	11	0	0
5	R.CELAAAMK.R	14	52	1	4
14	R.HGLDNYR.G	512	231	7	2
21	R.GYSLGNWVCAAK.F	233	42	8	3
33	K.FESNFNTQATNR.N	176	125	13	15
45	R.NTDGSTDYGILQINSR.W	151	81	10	12
61	R.WWCNDGR.T	453	141	11	6
68	R.TPGSR.N	14	2	0	0
97	K.IVSDGNGMNAWVAWR.N	323	310	48	79
114	R.CKGTDVQAWIR.G	2	0	0	0
116	K.GTDVQAWIR.G	443	107	43	160
TOTAL		2386	1102	141	281

(a)

(b)

Figure 8.
Mixture identification results for Lysozyme.

position	peptide	pure	uncertain	possible mixture	probable mixture
28	K.DGNKVDLYGK.A	9	0	0	0
38	K.AVGLHYFSK.G	55	30	0	0
47	K.GNGENSYGGNGDMTYAR.L	13	106	18	0
64	R.LGFKGETQINSDLTGYGQWEYNFQGNNSGADAQTGNK.T	3	18	3	0
68	K.GETQINSDLTGYGQWEYNFQGNNSGADAQTGNK.T	7	26	3	0
104	R.LAFAGLK.Y	24	17	1	0
111	K.YADVGSFDYGR.N	83	117	9	7
154	R.VGGVATYR.N	47	5	0	0
162	R.NSNFFGLVDGLNFAVQYLK.N	1070	58	0	0
189	R.RSNGDGVGGSSISYEYEGFGIVGAYGAADR.T	1	1	0	0
190	R.SNGDGVGGSSISYEYEGFGIVGAYGAADR.T	1	0	0	0
218	R.TNLQEAQPLGNGKK.A	12	0	0	0
231	K.KAEQWATGLK.Y	5	1	1	0
241	K.YDANNIYLAANYGETR.N	723	170	7	0
257	R.NATPITNK.F	3	0	0	0
257	K.NATPITNKFTNTSGFANK.T	21	5	0	0
265	K.FTNTSGFANK.T	33	5	0	0
275	K.TQDVLLVAQYQDFGLRPSIAYTK.S	418	84	5	0
303	K.DVEGIGDVLVNYFEVGATYYFNK.N	70	19	0	0
327	K.NMSTYVDYIINQIDSDNK.L	21	205	17	1
345	K.LGVGSDDTVAVGIVYQF.-	21	1	0	0
		2640	868	64	8

(a)

(b)

Figure 9.
Mixture identification results for OmpF.

Figure 10.

Peak-by-peak relative quantification for Figure 4, a mixture spectrum of GLSDGEWQQVLNVWGK from myoglobin. The traces show the relative intensities of the peaks for P and P'. The x-axis shows the peaks sorted by mass, and the y-axis shows relative intensity in arbitrary units. There are two pairs of corresponding peaks (circled) that give outlier ratios; these are the peaks corresponding to cleavage on the C-terminal side of the W in position 7. For CID fragmentation, W[+32] gives more intense peaks at this cleavage than does W[+4].

Xim identified six mixture spectra for GKGGKGLGKGAKR from Histone H4. The modifications give the mass deltas, so that 56.42.56.42 means that the first and third lysines are propionylated and the second and fourth are acetylated. Real-valued Score and Sep Score are derived from Ascore, and generally track the number of matched b- and y-ion peaks, in the columns labeled “b,y” and “Sep peaks”. The results are sorted by decreasing “mixture-ness” (Sep Score of the second identification); the last three rows have statistically insignificant (Ascore p-value > 0.1) second identifications.

Table 1

Lysine mods	Score	b,y	Sep Score	Sep peaks	Lysine mods	Score	b,y	Sep Score	Sep Peaks
56.42.56.42	26.8	7,11	3.87	2,3	56.56.42.42	25.0	6,11	2.96	1,3
42.56.42.42	34.0	8,13	6.73	3,4	42.42.56.42	27.5	6,12	2.83	1,3
56.56.56.70	25.0	6,11	4.51	1,4	56.56.70.56	22.5	6,10	2.78	1,3
56.42.42.56	15.7	5, 8	2.78	1,3	56.42.56.42	13.5	5, 7	1.25	1,2
56.42.42.56	29.5	7,12	5.53	2,4	56.42.56.42	21.0	5,10	0.95	0,2
56.56.56.42	24.1	4,12	4.36	0,4	56.56.42.56	19.0	4,10	0.52	0,2

Table 2

This table gives modifications observed in oxidative footprinting experiments. Amino acid residues are listed roughly in order of reactivity. ByOnic considers the modifications shown in bold to be common. In order to control the combinatorial explosion, ByOnic allows only one uncommon modification per peptide.

Residues	Modification Masses
Cys	+15.99 , +31.99 , +47.98 , -15.98
Met	+15.99 , +31.99 , +33.97, -32.01, -48.00
Trp	+3.99 , +15.99 , +19.99, +31.99 , +47.98
Phe, Tyr	+15.99 , +31.99 , +47.98
His	+4.98, +15.99 , -10.03 , -22.03 , -23.02
Arg	+13.98, +15.99, -43.05
Leu, Ile, Val, Lys, Gln	+13.98, +15.99
Ser, Thr	+15.99, -2.02
Pro	+13.98, +15.99, +31.99, -30.01
Glu	+13.98, +15.99, -30.01
Asp	+15.99, -30.01
Ala, Asn	+15.99