BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i328–i336
doi:10.1093/bioinformatics/btm198

Using dynamic programming to create isotopic distribution maps from mass spectra

Sean McIlwain[1,*], David Page[2], Edward L. Huttlin[3] and Michael R. Sussman[3]

[1]Department of Computer Sciences, University of Wisconsin, Madison, WI, [2]Department of Computer Sciences and Department of Biostatistics, University of Wisconsin, Madison, WI and [3]Department of Biochemistry, University of Wisconsin, Madison, WI, USA

ABSTRACT

Motivation: This article presents a method to identify the isotopic distributions within a mass spectrum using a probabilistic classifier supplemented with dynamic programming. Such a system is needed for a variety of purposes, including generating robust and meaningful features from mass spectra to be used in classification.

Results: The primary result of this article is that the dynamic programming approach significantly improves sensitivity, without harming specificity, of a probabilistic classifier for identifying the isotopic distributions. When annotating isotopic distributions where an expert has performed the initial 'peak-picking' (removal of noise peaks), the dynamic programming approach gives a true positive rate of 96% and a false positive rate of 0.0%, whereas the classifier alone has a true positive rate of only 47% when the false positive rate is 0.0%. When annotating isotopic distributions in machine peak-picked spectra, which may contain many noise peaks, the dynamic programming approach gives a true positive rate of only 22.0%, but it still keeps a low false positive rate of 1.0% and still outperforms the classifier alone. It is important to note that all these rates are when we require *exact* matches with the distributions in annotated spectra; in our evaluation a distribution is considered 'entirely incorrect' if it is missing even one peak or contains even one extraneous peak. We compared to the THRASH and AID-MS systems using a looser requirement: correctly identifying the distribution that contains the mono-isotopic mass. Under this measure, our dynamic programming approach achieves a true positive rate of 82% and a false positive rate of 1%, which again outperforms the classifier alone. The dynamic programming approach ends up being more conservative than THRASH and AID-MS, yielding both fewer true and false peaks, but the F-score of the dynamic programming approach is significantly better than those of THRASH and AID-MS. All results were obtained with 10-fold cross-validation of 99 sections of mass spectra with a total of 214 hand-annotated isotopic distributions.

Availability: Programs are available via http://www.cs.wisc.edu/~mcilwain/IDM

Contact: mcilwain@cs.wisc.edu

1 INTRODUCTION

Analyzing proteomic data generated from mass spectrometry shows promise for predicting disease and finding biomarkers within bodily fluid samples (serum, urine, etc.). By taking mass spectra of affected and unaffected samples from subjects, potential features can be extracted that can separate the two groups using classification algorithms. One major challenge is generating predictive features that have statistical and biological relevance (Baggerly *et al.*, 2004; Coombes *et al.*, 2005). A number of different approaches have been explored (Dekker *et al.*, 2005; Hilario *et al.*, 2003; Li *et al.*, 2002; Qu *et al.*, 2002; Rai *et al.*, 2002; Soltys *et al.*, 2004; Schwegler *et al.*, 2005; Tibshirani *et al.*, 2003; Wu *et al.*, 2003 Zlatkis *et al.*, 1979; Tibshirani *et al.*, 2004).

There are many issues to consider when analyzing proteomic mass spectrometry data for classification. Detector saturation reduces the predictive power of the peak intensities of the sample's mass spectrum. There are mass-to-charge (m/z) shifts of common peaks between spectra. Within the spectra, there are noise peaks that are not indicative of the underlying biology. These noise peaks are caused by chemical and electronic noise during the sample acquisition. Also, there are redundant features due to isotopic distributions, various adducts, multiple charge states and peptide fragments occurring from proteolysis.

Isotopic distributions, which are collections of peaks occurring from the same molecular compound but having different compositions in their atomic isotopes, may be seen as help rather than a hindrance. These distributions give us multiple peaks, and hence multiple evidence sources, for specific peptides (Clauser *et al.*, 1999; Desiere *et al.*, 2004; Eng *et al.*, 1994; Keller *et al.*, 2002; Goldberg *et al.*, 2005). Isotopic distributions can be particularly helpful in cases where we are not mapping to a set of peptides obtained by a 'theoretical trypsin digest' of an organism's proteome. Such cases include experiments with metabolomics or with organisms whose genomes have not been sequenced and hence whose proteomes have not been predicted. Also, for quantitative proteomics experiments using isotopic labeling where the isotopic distribution patterns are non-standard (Beynon and Pratt (2005); Huttlin *et al.* (in press), Krijgsveld *et al.*, 2003; Ong *et al.*, 2002; Whitelegge *et al.*, 2004; Yao *et al.*, 2001), the current annotation methods may have trouble (Chen *et al.* (2006); Horn *et al.*, 2000).

While some of current algorithms are concerned with deconvolving spectra into the mono-isotopic peaks, it would be useful to have entire isotopic distributions for a number of applications. One such application would be in quantitative proteomics experiments involving isotopic labeling where the ratios of the peak heights or areas are used for quantitative

*To whom correspondence should be addressed.

measurements. Also, using isotopic distributions as features could possibly prove to be more robust and biologically significant when compared to single peak features. Experiments such as these would benefit from a map of isotopes within spectra rather than a mono-isotopic deconvolved counterpart.

We describe a method to annotate the isotopes within a mass spectrum. To accomplish this, we propose an algorithm analogous to an approach by Craven, Page, Shavlik, Glasner and Bockhorst to a very different problem: predicting operons within a DNA sequence from the *E.coli* K-12 genome (Craven *et al.*, 2000). Their algorithm employs dynamic programming, building upon using a naïve Bayes model that predicts the probability of an operon given the data.

Using expert-constructed peak lists from the spectra, we show that the dynamic programming map algorithm achieves a dramatically superior true positive/false positive rate when compared to the classifier used to score isotopic distributions. We also extend the algorithm to handle the many noise peaks present when using machine-constructed peak lists rather than expert-constructed peak lists, and again the dynamic programming approach outperforms the classifier. We do not address the issue of overlapping distributions, but an extension to handle this issue is proposed as future work.

2 APPROACH

Using probabilities from features of distributions, such as length, shape, inter-distribution distances, and intra-distribution distances, we can construct a naïve Bayesian model, illustrated in Figure 2, to estimate the probability that a proposed distribution of peaks is an isotopic distribution. By 'an isotopic distribution', we mean both that (1) every peak in the distribution arises from the same molecular compound with different combinations of isotopes and (2) no other peaks in the spectrum arise from the same molecular compound with the exception of charge state. We can estimate the parameters (probabilities) of the naïve Bayes model using either the literature or training data, i.e. some annotated spectra. In our work, we choose to estimate the parameters from hand-annotated spectra.

Given a probability for each potential isotopic distribution (each run of consecutive peaks), we would like to map all the peaks of the spectrum into their isotopic distributions. We take the score of any peak in such a map to be the log (base 2) probability of the distribution to which it is mapped, and we take the score of a map to be the sum of the peak scores.[1]

[1]We sum the log probabilities by peak rather than by distribution to avoid artificially raising the scores of maps with very long distributions. For a spectrum of length n the score of any map is now effectively the product of n values, regardless of the number of distributions used in the map. If we summed one score per distribution instead, a map placing all n peaks of a spectrum in a single distribution might score better than other maps simply because it is the sum of a single negative value (log probability) rather than the sum of many negative values. The same scoring approach was motivated and used by Craven, Page, Shavlik, Glasner and Bockhorst (Craven *et al.*, 2000) for the task of finding operons in *E.coli*.

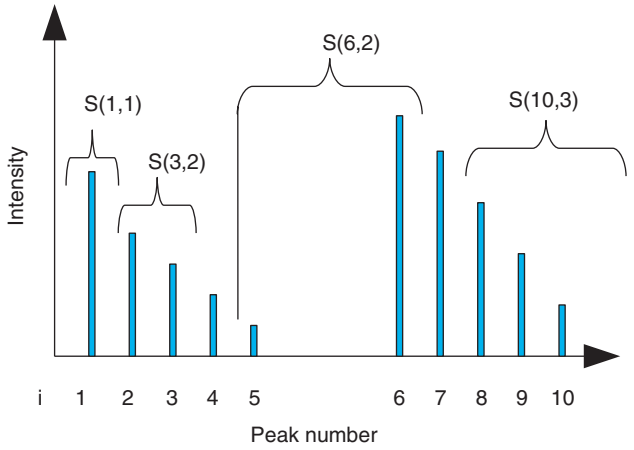

Fig. 1. Example S-Matrix entries that are collected from a peak list. Mass-to-charge ratio increases from left to right.

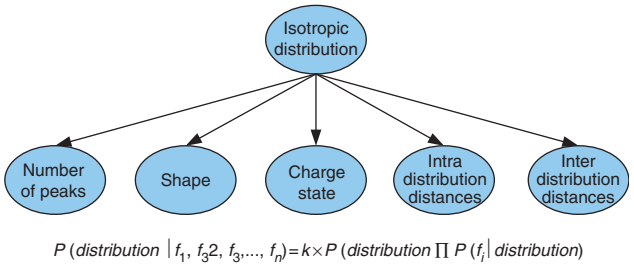

Fig. 2. Example naïve Bayes model for estimating isotopic Distribution Probabilities.

We now describe a dynamic programming approach to find the optimal map with respect to this scoring function.

In the dynamic programming approach, we use the distribution probabilities from the Bayes net to calculate $S(i,j)$, where $S(i,j)$ is the probability that the peak distribution ending at position (peak number) i and having length j is a true isotopic distribution (Fig. 1). $M(i)$ denotes the optimal distribution map, or sequence of isotopic distributions, beginning with the first peak and ending at peak position i. For successive values of i, we consider extending the current distribution or starting a new one. The overall score for $M(i)$ is given by:

$$M(0) = 0,\ M(i) = \max_i \begin{cases} 1 \times S(i,1) + M(i-1), \\ 2 \times S(i,2) + M(i-2), \\ \ldots, \\ i \times S(i,i) + M(i-i) \end{cases} \quad (1)$$

We also store the length of the current distribution at every $M(i)$ in order to recover the optimal map. The proof of the following theorem is analogous to the correctness proof for the operon finding algorithm (Craven *et al.*, 2000).

THEOREM 1. *Given any spectrum of n peaks, the algorithm returns an optimal—maximum scoring—map of that spectrum into isotopic distributions.*

The run-time of the algorithm is quadratic in the number of peaks, n, in the spectrum. Nevertheless, we can further reduce the run time in practice because it is unnecessary to consider possible isotope distributions with peaks of more than, say, 12 or 15 peaks. Maintaining the validity of the theorem while making this reduction requires an assumption that an optimal map contains no distributions of length greater than this bound; in practice distributions of this length are given such low probabilities according to the model that they are unlikely to be returned anyway.

Within a mass spectrum there are many peaks that occur due to chemical and electrical noise. These spurious peaks can occur within an isotopic distribution from a real peptide. To handle noise peaks, we devise a modification to the score matrix $S(i,j)$. $S(i,j)$ contains the best distribution between peak i and peak $i-j$, inclusive. This distribution can now be any subset of the peaks between $i-j$ and i. We introduce a penalty for the exclusion of peaks between $i-j$ and i. We then run dynamic programming as before, using the same formula as discussed above. At every $M(i)$, we must also store the optimal peak distribution that starts at position $i-j$ and ends at j. We can then build the map using these optimal distributions at the $M(i)$'s. To prevent exponential explosion, we place a bound on j. Assuming this bound is accurate, the revised dynamic programming algorithm maintains optimality. Due to the number of possible noise peaks within a spectrum, we also allow the algorithm to label an entire run of peaks as noise. If the maximum probability of the optimal distribution within a run of peaks is less than a certain threshold, called the noise threshold, the algorithm labels these peaks as a run of noise peaks, having probability of the noise threshold plus 0.1.

2.1 Naïve Bayes model

We use a naïve Bayes model to calculate the isotopic distribution probabilities. The features employed by the model encode data about charge state, distances between peaks within a potential distribution, distance between the first peak of the potential distribution and the last peak before it, distance between the last peak of the potential distribution and next peak after it, the shape of (sequence of peak heights in) the potential distribution, the number of peaks in the potential distribution and the relative ratio of intensities of the two highest peaks in the distribution (repeated also for the two lowest peaks). The remainder of this section describes these features in more detail. The naïve Bayes model assumes these features are independent of one another given the class (true isotopic distribution or not an isotopic distribution). Even when this assumption is violated, naïve Bayes models often work better in practice than more complicated Bayesian models because the conditional independence assumption means the model needs to estimate fewer parameters from the data, often resulting in better parameter estimates.

2.2 Determining charge state

To determine the charge state, we calculate the best-fitting line on a plot of peak number (increasing from 1 to n peaks in the proposed distribution) versus the m/z value for that peak. We require the slope of the line to be $1.0028/Z$, where $Z = 1, 2, 3, \ldots, N$. We also require the line to pass through the first peak. Z is the best-fitting charge state for the distribution. The reason for calculating the charge state in this way is the following. The value of 1.0028 is the mass of a neutron and Z is the charge of the molecule. Owing to the nature of the measurement made by the mass spectrometer of a molecular compound, the mass peak values are divided by the number of charges within the molecule (hence the m/z ratios for the x-axis of mass spectra). We can calculate probabilities of the various charge states by counting the frequencies of the charge states within the training peak spectra.

For our data, we decided to use a maximum allowable charge state of 3. It is worth noting that for high mass molecules, other methods for determining charge state such as the Patterson or Fourier method may be more robust (Senko *et al.* 1995a). Incorporating these methods to improve our results is listed as future work.

2.3 Inter- and intra-distance features

The intra-distance feature captures the degree to which the distances between peaks within a potential isotopic distribution look like the distances we would expect in an isotopic distribution. The inter-distance feature represents the degree to which the distances at the borders of the potential distribution appear appropriate; specifically, the distance between the last peak before (first peak after) the distribution and the first peak (last peak) within the distribution should not be similar to distances we would expect *within* a true isotopic distribution. A complication is that the distances we expect within a true isotopic distribution vary with charge state. We therefore use the fitted line from the previous subsection (charge state feature) to determine the actual inter-distribution (inter) and intra-distribution (intra) distance features, through a method described in the remainder of this section. This determination involves estimating the error from the line (see Fig. 3). We explain the intra distance feature calculation first and follow up with the inter-distance probability.

For a hypothetical isotopic distribution, we calculate the maximum squared error of the mass-to-charge values to the expected mass-to-charge values given the charge state as determined above. This error can be thought of as the maximum deviation of a distance between two consecutive peaks in the distribution from the distance we would expect in an ideal isotopic distribution, given the charge state. The computed squared error is then used as a feature, called the intra-distanc- feature, for our model.

For the inter-distance feature, we take the charge state fit from the proposed isotopic distribution and calculate the squared error of the best-fit peak 'outside' of this distribution. This error is the minimum of distances between either the last peak before the distribution and first peak inside or the last peak inside the distribution and first peak afterward. We also refer to the intra-distance and inter-distance features as Inter/Intra Error.

2.4 Shape probabilities feature

For shape, we mean the relative intensity patterns of the peak distribution. The ratio patterns of the isotope peaks are

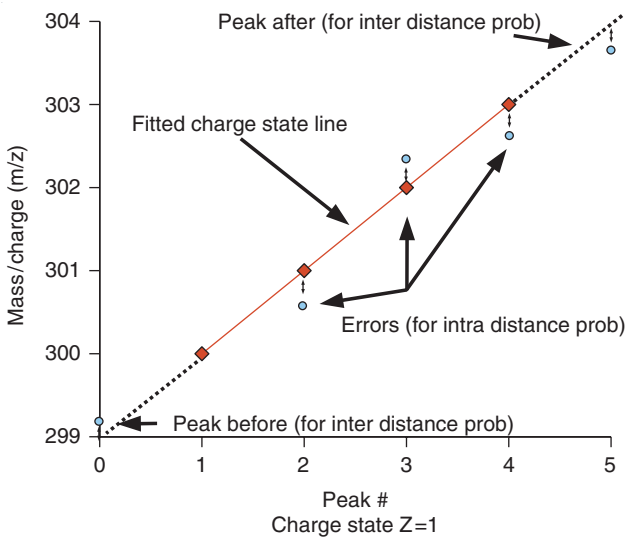

Fig. 3. Fitted charge state line.

dependent upon the molecular composition. For estimating this shape probability, we use an array of weighted nearest neighbor classifiers that are trained based upon the ratios of the peaks relative to the highest peak within the proposed isotopic distribution. Each classifier is trained using a specific number of ratios. Classifier 3 has three ratios, Classifier 4 has four ratios, etc. The number of ratios that a particular distribution has determines which classifier, to use for the prediction. For training, we generate a negative example from each positive example by randomly permuting the peak ratio order. From this classifier, we get a value in the range (0...1) that is indicative of whether these ratios fit a positive isotopic distribution. For distribution of one or two peak ratios, we return a value of 0.5 to signify undecided. We use this value as a feature for the overall classifier. It is important to note that these classifiers are re-trained on every fold of cross-validation (discussed later), using only the training data for that fold, to avoid an overly optimistic estimate of the performance of the method.

We realize that another shape feature that could be used is to calculate extact isotopic distributions using a method such as Mercury using averagine molecular compounds around the mass of the observed isotopic distribution (Rockwood *et al.*, 1996; Rockwood and Haimi 2006; Senko *et al.*, 1995b). This method is utilized in THRASH and is listed as future work, to see if including it as a feature in the classifier improves upon our current results (Horn *et al.*, 2000).

2.5 Remaining features and overall classification

Additional features include the number of peaks in the distribution. Others are the ratios of the most intense peak to the mean and median of the entire peak list. We also include mean and median intensity ratios for the second highest peak and least intense peak. Another feature is the ratio of the highest peak intensity over the number of peaks in the distribution (Max Npeaks Ratio). We build a naïve Bayes classifier using these features. Our naïve Bayes classifier takes both discrete and continuous features, where it assumes continuous features follow a Gaussian distribution. Because not all continuous features are Gaussian, we also consider binning the continuous features such that each bin contains an equal number of the data points. The number of bins is tuned to maximize the area under the precision-recall curve (APR) of each individual feature. Tuning is repeated on every fold of cross-validation, to avoid over-optimistic estimates of performance. Tuning is performed by an inner loop of 10-fold cross validation. If the APR for a Gaussian feature using the same cross-validation fold set is better than the binned feature, we use the Gaussian feature instead (Davis and Goadrich, 2006). We use counts for the discrete features. The log of the probability is used for the $S(i,j)$. Using a naïve Bayes model, we can assign probabilities to the importance of each of these features for determining the overall probabilities for the $S(i,j)$ matrix. We then can build the $M(i)$ matrix from Equation (1) to yield the isotopic distribution map.

3 METHODS

Mass spectra sections were obtained from labeled plant data in which Arabidopsis was grown in liquid culture with either natural abundance or ^{15}N-labeled MS salts. Labeled and unlabeled samples were then combined, fractionated and analyzed via LC-MS on a Micromass QTOF-II mass spectrometer (Nelson *et al.*, 2007). A random selection of isotopic envelopes was taken from the LC-MS analysis of a single digested protein fraction. Isotopic-labeled pairs were selected based on visual inspection for their clarity and an effort was made to include spectra representative of 1+, 2+ and 3+ charge states at a range of m/z values. This resulted in 99 sections of peak picked mass spectra that contain 214 hand-annotated distributions. Note, some of the sections came from the same spectrum, but these do not overlap in their mass-to-charge values. For each isotopic distribution in the training set, we generate a positive example and a set of 'near-miss' negative examples. These 'near-misses' include distributions:

- having an additional peak in the beginning or end of the distribution (2 negatives)
- missing a peak in the beginning or end of the distribution (2 negatives)
- consisting of a single peak from the true distribution (1 negative)
- having one or two additional noise peaks within the distribution (1 negative for each peak, and 1 negative for each pair of peaks)

We build our naïve Bayes model using a training set built from the features generated as described previously. To score the generated isotopic map, we use three different metrics. We call them the *absolute*, *coarse* and *mono-isotopic* scores. The absolute metric is over entire distributions. The second metric is over peaks, to essentially capture the fraction of peaks that are grouped correctly into an isotopic distribution. The third metric is also over peaks, with additional context restraints for finding the correct peak and distribution charge. For each score, we define the four quadrants of a contingency table, or confusion matrix. From these confusion matrices, we can calculate performance points for ROC and PR curves.

3.1 Absolute scores

Improving the robustness of two-class feature selection using proteomics measured via mass spectrometry requires this feature as well.

Therefore, we utilize an absolute score method yielding the following counts for a confusion matrix.

- True positive—exact distribution appears in the map.
- True negative—generated negative distribution does not appear within the map.
- False positive—exact distribution does not appear within the map.
- False negative—generated negative distribution appears within the map.

3.2 Coarse scores

The absolute score counts and isotopic distribution wrong even if it misses one peak of the distribution or includes one extraneous peak (for e.g. a noise peak). Nevertheless, a mostly-correct distribution is often very useful as feature for machine learning or in approaches to quantitative mass spectrometry. The coarse score provides a way of giving credit for such mostly-correct distributions:

- True positive—peak is actually in an isotopic distribution and predicted in an isotopic distribution.
- True negative—peak is actually not in an isotopic distribution and predicted as not in an isotopic distribution.
- False positive—peak is actually not in an isotopic distribution and predicted in an isotopic distribution.
- False negative—peak is actually in an isotopic distribution, but predicted not in an isotopic distribution.

3.3 Mono-isotopic scores

For some applications, only the mono-isotopic mass is needed. Algorithms related to ours, such as THRASH (Horn *et al*., 2000) and AID-MS (Chen *et al*., 2006), return the mono-isotopic peak for every predicted isotopic distribution. So to measure performance for determining the correct mono-isotopic masses and to allow comparison of our method against others, we introduce the mono-isotopic mass metric.

- True positive—peak is the mono-isotopic peak, and is predicted in a distribution having the same charge.
- False positive—peak is not the mono-isotopic peak, but is predicted as one in a distribution, or the charge is not the same.
- False negative—peak is a mono-isotopic peak, but not found in an isotopic distribution.
- True negative—peak is not a mono-isotopic peak and is not found as a mono-isotopic peak in a distribution.

Our data set is a mixture of normal peptides and their ^{15}N isotopically labeled counterparts. For natural abundance peptides, the mono-isotopic peak corresponds to the peak resulting from only the most common isotopes of each atom: ^{1}H, ^{12}C, ^{16}O, ^{14}N and ^{32}S. Since each of these is the lowest mass isotope, the mono-isotopic peak is that peak in each natural abundance envelope with the lowest m/z value. The situation is more complicated for ^{15}N-labeled samples: ^{15}N-labeled peptides do not contain a truly mono-isotopic peak because every peak in the distribution can result from different combinations of labeled isotopes. However, when nearly complete ^{15}N enrichment is achieved, the isotopic envelope of the labeled peptides takes on a shape that is similar to its natural abundance envelope, but shifted by one mass unit for each nitrogen in the peptide. By analogy with the unlabeled isotopic envelope, we will define the heavy mono-isotopic peak to be the peak within the labeled distribution which results predominantly from ^{1}H, ^{12}C, ^{16}O, ^{15}N and ^{32}S. The charge state is calculated using the previously described algorithm. We then can use the mono-isotopic mass-to-charge and the charge values to calculate the mono-isotopic mass.

We train the machine peak-picked (MP) algorithm using a grid of values for the noise threshold (0.0–0.9, step 0.025) and the noise peak penalty (0.0–0.2, step 0.005). Using this grid, we maximize the function $0.5 \times (P + R)$ for the mono-isotopic scores of the training sections. We perform 10-fold cross-validation generating ROC and PR curves for the classifier and ROC and PR points for each of the score metrics. The *curve* can be generated for the probabilistic classifier, but the isotopic distribution map commits to a particular set of distributions, not a ranking or probability estimate over distributions. Therefore, scoring the isotopic distribution map algorithm's predictions using the absolute metric yields a *point* on the graph that signifies the overall ROC and PR scores for the isotope distribution map. These curves and points are generated using the MP algorithm. We also build a curve and point for the expert peak-picked, (EP) algorithm. The EP algorithm only considers the peaks that are marked as being in an isotopic distribution.

4 DISCUSSION

Our resulting ROC and PR curves for the model and map using leave-one-out cross-validation are shown in Figures 4–11. The statistical results are in Table 1. The absolute score performance measurement from the EP map gives a true positive rate of 96% and a false positive rate of 0.0% while the MP map gives a true positive rate of 22.0% and a false positive rate of 1.0%. The coarse score from the EP map gives a true positive rate of 100% and a false positive rate of 0.0%. When tuning upon the mono-isotopic score, the MP map obtains a coarse score true positive rate of 85% and a false positive rate of 4%. The corresponding rates for the mono-isotopic scores using the MP map are a true positive rate of 82% and a false positive rate of 1%. The coarse and mono-isotopic scores for the MP map show that the algorithm can annotate peaks that belong to an isotopic distribution and obtain enough information to determine the mono-isotopic masses reasonably well.

Looking at the ROC plots that use the full set of features (Fig. 4 and 6), the curves lie substantially above the diagonal.

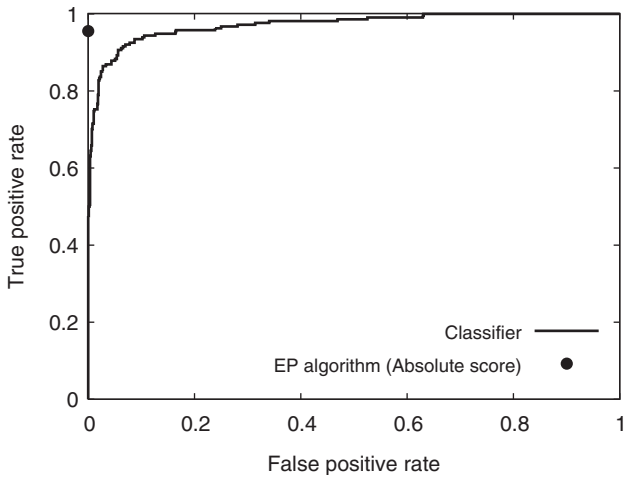

Fig. 4. Curves and points using all features with EP data ROC curve.

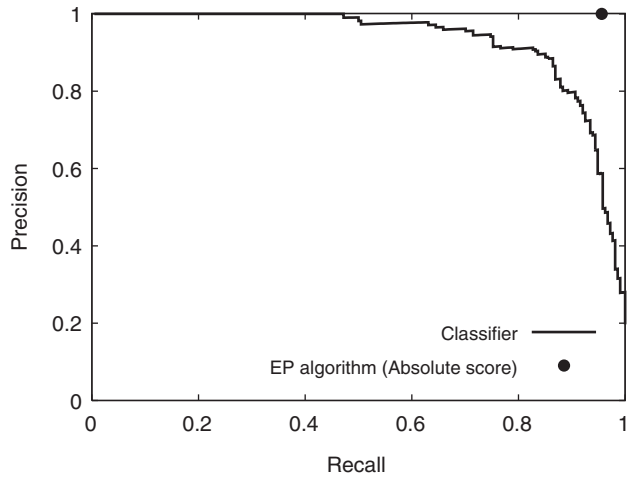

Fig. 5. Curves and Points using all features with expert EP PR curve.

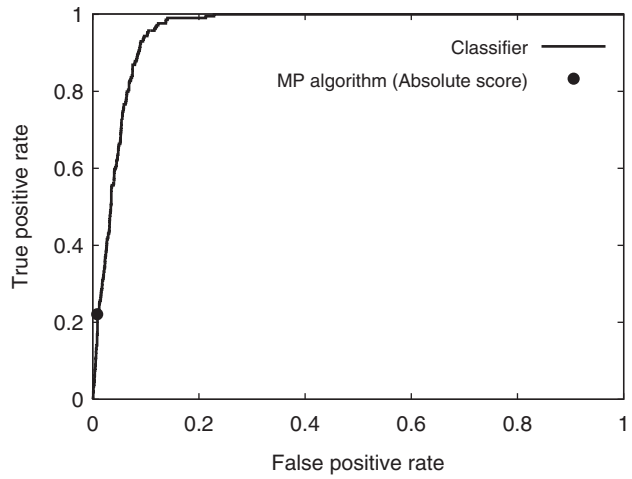

Fig. 6. Curves and Points using all features with MP data (ROC curve).

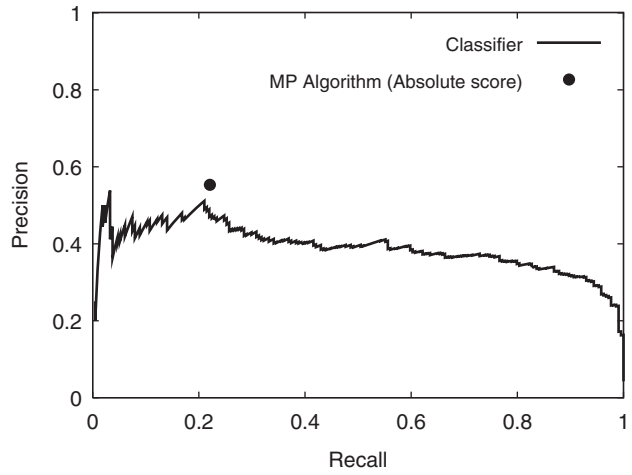

Fig. 7. Curves and Points using all features with MP data (PR curve).

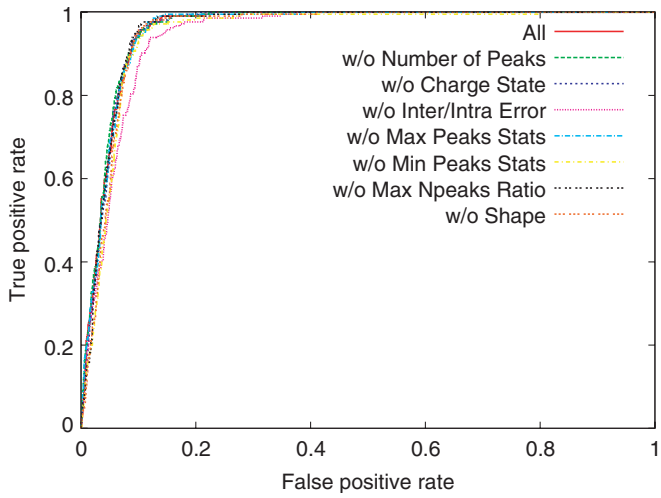

Fig. 8. Curves without one feature (ROC curve).

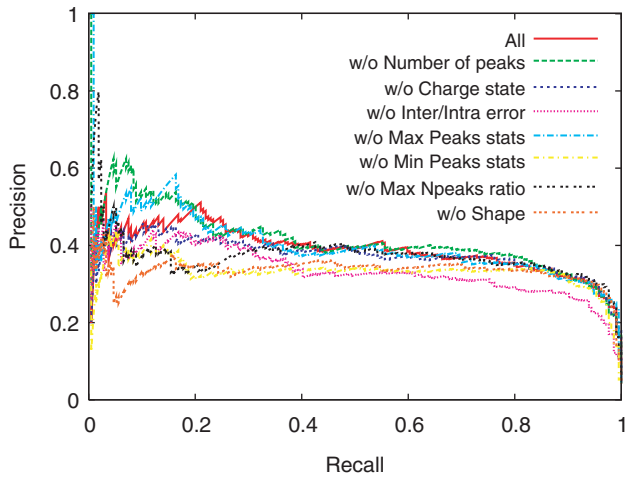

Fig. 9. Curves without one feature (PR curve).

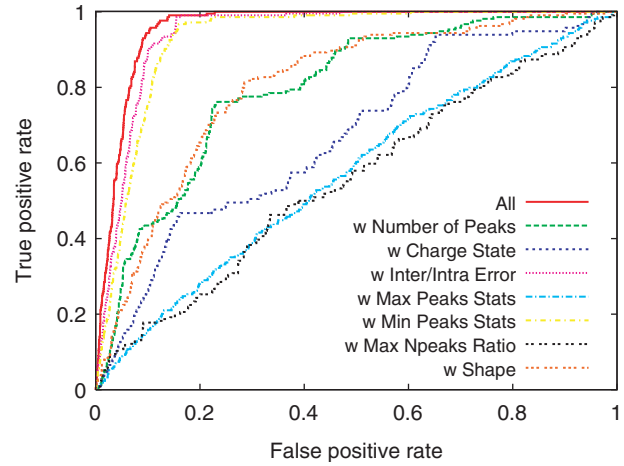

Fig. 10. Curves using one feature (ROC curve).

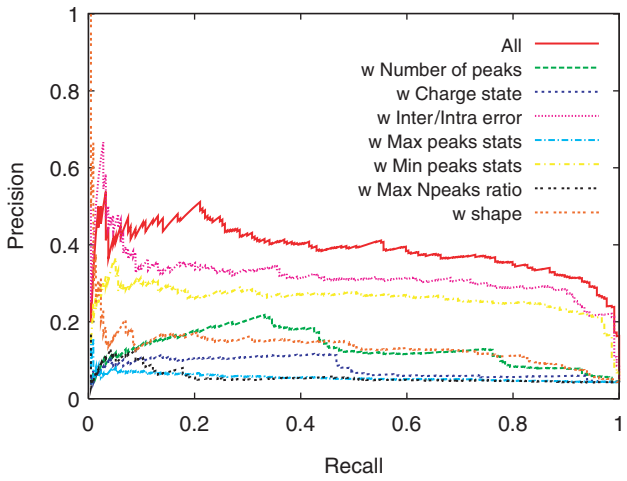

Fig. 11. Curves using one feature (PR curve).

Table 1. Statistical Results of Classifier and Dynamic Programming Algorithms (see Section 3 for the distinction between absolute and coarse scores)

	EP Algo	MP Algo
Classifier APR	0.93	0.39
Classifier AROC	0.97	0.96
Absolute		
Precision	1.00	0.55
Recall	0.96	0.22
F-Score	0.98	0.32
False positive rate	0.00	0.01
Coarse		
Precision	1.00	0.76
Recall	1.00	0.85
F-score	1.00	0.80
False positive rate	0.00	0.04
Mono-Isotopic		
Precision	0.98	0.62
Recall	0.96	0.82
F-score	0.97	0.71
False positive rate	0.00	0.01

This indicates that the probabilistic (naïve Bayes) classifier is performing much better than chance. The point corresponding to the isotopic distribution map in turn lies well above the ROC curve, indicating the construction of the map further improves performance. In domains that have many potential false positives, such as this one, precision-recall (PR) curves are often used instead of ROC curves because they more clearly show differences between algorithms, especially in terms of the number of false positive predictions. The PR curves (Figs 5 and 7) show that building the map yields a profound bonus in detecting isotopic distributions. Since our dynamic programming approach uses the classification algorithm for building the map, any improvements to the probabilistic classification scheme should in turn improve the map building performance.

The lesion and one-feature curves in Figures 8–11 give a hint of the relative value of each feature for determining an isotopic distribution. Lesion tests show that omitting any one of the Inter/Intra Error, Min Peaks Stats and the Shape features reduces the classifier's performance. The one-feature tests show that the Inter/Intra Error and the Min Peaks features classify well when used individually, though not as well as the full classifier. From this, we conclude that given the current list of the available features, the Inter/Intra and the Min Peaks features are probably the most important for the classifier.

5 RELATED WORK

Two systems that supply similar annotations are THRASH (Horn *et al.* 2000) and AID-MS (Chen *et al.*, 2006). THRASH is a widely used algorithm for interpretation of high-resolution mass spectra. THRASH is focused on fitting theoretical 'averagine' isotopic clusters through least squares in a moving window. The AID-MS algorithm implements a top-down (by decreasing spectral intensity) peak selection approach supplemented with novel charge state determination and other features to reduce false positive rates.

THRASH and AID-MS return the mono-isotopic peak for every predicted isotopic distribution, rather than returning the full isotopic distribution. Therefore, to compare our method against these two methods, we decided to compare the performance of finding distributions that contain the mono-isotopic mass peak.

Key distinctions between our algorithm and both AID-MS and THRASH are the following. First, construction of a unique map has a tendency to increase precision, but at the cost of decreased recall. Second, our algorithm can be *trained*. This property has the potential to make it more robust for use on data from a wider array of experimental conditions, with varying machines and with isotopic distributions that occur as the result of isotopic labeling or other 'un-natural' conditions, provided annotated training data for the conditions are available.

For both the THRASH and AID-MS algorithms, the results are given in a table. Included in this table is the mono-isotopic mass-to-charge, charge state and mono-isotopic mass for each distribution found in the spectra. For both these methods, we use the smoothed spectrum from which the peak list section come from.

To calculate statistics for THRASH and AID-MS, we compile a list of mono-isotopic masses from our annotated spectra sections and the corresponding results from THRASH or AID-MS. The one challenge in computing these metrics is that THRASH and AID-MS may place a mono-isotopic peak in a slightly different mass position than the expert-annotated data. To adjust for these offsets between the expert list and THRASH (or AID-MS) list, we employ the following algorithm.

Do: Find closest match between the two lists (distance is the difference in mass) if the difference is less then delta, then accept as a match and remove the matching masses from the two lists. Repeat until no more matches are found.

Table 2. Comparison of Mono-Isotopic peak finding

	Our method	THRASH	AID-MS
Precision	0.62	0.39	0.21
Recall	0.82	0.88	0.66
F-score	0.71	0.54	0.32
False positive rate	0.01	0.04	0.07

To try to be as fair as possible to THRASH and AID-MS we vary the delta from 0.0 to 0.5 to obtain the best F-score. After this algorithm completes, the following statistics are collected:

- True positive—number of matches found.
- False positive—number of THRASH/AID-MS peaks left.
- True negative—number of peaks in peak list minus (TP + FN + FP).
- False negative—number of annotated mass peaks left.

The results of these metrics applied to our annotated spectra set are given in Table 2. It appears that our algorithm is competitive with THRASH and AID-MS for this data set. In the future, we plan to obtain data obtained from different mass spectrometers and compare these algorithms' results. It is also worth noting that THRASH and AID-MS were originally developed to work with raw spectra, but we only used smoothed data. Using raw data may change the results for THRASH and AID-MS.

6 FUTURE WORK

As seen in Figures 4 and 5, the algorithm performs well in the absence of noise peaks (Figs 6 and 7). Due to the differences in mass of the amino acids, there are combinatorially many possible peptides near a particular mass. Overlapping distributions from different peptides are inevitable. We plan on modifying the algorithm to handle overlapping distributions. Trying other types of classifiers might also improve performance; possible classifiers include support vector machines, Bayesian network learning algorithms, tree-augmented naïve Bayes and probabilistic decision trees.

Also, improving the features, (such as distribution shape), or introducing more features into the model could improve performance of the classifier and subsequently the map building algorithm. Additional features that capture some of the logic from AID-MS and THRASH would be beneficial. Using the Patterson or the Fourier method for determining charge state can be implemented as a set of features. Estimating isotopic distributions using averagine compounds around the mass of interest may in fact provide a better shape feature than the one presented in this article. Isotopic distribution peaks generally increase in number as the mass-to-charge ratio increases. These features can possible produce an improved classifier. This new classifier will then allow the dynamic programming algorithm to produce better isotopic maps.

We also propose to augment and test our algorithm upon other annotated MS data sets that could include raw, more complex or less resolved data. We also plan on extending the algorithm to intergrate LC-MS data. To handle this, we would include additional features into our classifier that take into account data from neighboring mass spectra scans.

Once we have an algorithm that performs well annotating the distributions, we can build feature vectors from the output. In those cases where the organism's proteome is known, we propose to further map these distribution annotations to peptide annotations. This will require MS/MS data or peptide fingerprinting, using a variety of widely available tools for searching peptide databases (e.g. SEQUEST (Eng *et al*., 1994), MS-Fit (Clauser *et al*., 1999), MASCOT (Perkins *et al*., 1999)) and evaluating the significance of their assignments (e.g. PeptideProphet (Keller *et al*., 2002).

7 CONCLUSION

This article has presented a naïve Bayes model for assigning probabilities to potential isotopic distributions in mass spectra. It has shown how performance of this model can be further improved by a dynamic programming algorithm to map a spectrum into its isotopic distributions. The algorithm also performs further removal of noise peaks—those not removed during initial pre-processing—while constructing the isotopic peak distributions.

ACKNOWLEDGEMENTS

We thank Neil Kelleher, Craig Wenger and Richard LeDuc for running THRASH on our data for us. This work is supported in part by the NLM training grant 5T15LM00739, NSF grant 0534908 and NIH training grant 5-T32-GM08349. We would also like to acknowledge Dr Amy Harms and the UW-Madison Biotechnology Center Mass Spectrometry Facility for technical advice and assistance.

Conflict of Interest: none declared.

REFERENCES

Baggerly,K.A. *et al*. (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, **20**, 777–785.

Beynon,R.J. and Pratt,J.M. (2005) Metabolic labeling of proteins for proteomics. *Mol. Cell. Proteomics*, **4**, 857–872.

Chen,L. *et al*. (2006) Automated intensity descent algorithm for interpretation of complex high-resolution mass spectra. *Anal. Chem*., **78**, 5006–5018. (http://www.bii.a-star.edu.sg/paper/chenli/)

Clauser,K.R. *et al*. (1999) Role of accurate mass measurement (+/−10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem*., **71**, 2871–2782.

Coombes,K.R. *et al*. (2005) Serum proteomics profiling—a young technology begins to mature. *Nat. Biotechnol*., **23**, 291–292.

Craven,M. *et al*. (2000) A probablistic learning approach to whole-genome operon prediction. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, pp. 116–127.

Davis,J. and Goadrich,M. (2006) The relationship between Precision-Recall and ROC curves. In *ICML 06: Proceedings of the 23rd International Conference on Machine Learning*, ACM Press, New York, USA, pp. 233–240.

Dekker,L.J. *et al*. (2005) A new method to analyze matrix-assisted laser desorption/ionization time-of-flight peptide profiling mass spectra. *Rapid Commun. in Mass Spectrom*., **19**, 865–870.

Desiere,F. *et al*. (2004) Integration with the human genome of peptide sequences obtained by highthroughput mass spectrometry. *Genome Biol*., **6**, R9.1–R9.12.

Eng,J.K. *et al*. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom*., **5**, 976–989.

Goldberg,D. *et al*. (2005) Automatic annotation of matrix-assisted laser desorption/ionization n-glycan spectra. *Proteomics*, **5**, 865–875.

Hilario,M. *et al*. (2003) Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics*, **3**, 1716–1719.

Horn,D.M. *et al*. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom*., **11**, 320–332.

Huttlin,E.L. *et al*. (2007) Comparison of full versus partial metabolic labeling for quantitiative proteomic analysis in arabidopsis thaliana. *Mol. Cell Proteomics*, **6**, 860–891.

Keller,A. *et al*. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal. Chem*., **74**, 5383–5392.

Krijgsveld,J. *et al*. (2003) Metabolic labeling of C. *elegans* and D. *melanogaster* for quantitative proteomics. *Nat. Biotechol*, **21**, 927–931.

Li,J. *et al*. (2002) Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin. Chem*., **48**, 1296–1304.

Nelson,C.J. *et al*. (2007) Implications of ^{15}N metabolic labeling for automated peptide identification using multiple search engine in arabidopsis thaliana. *Proteomics*, **7**, 1279–92.

Ong,S.-E. *et al*. (2002) Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics*, **1**, 376–386.

Perkins,D.N. *et al*. (1999) Probability-based protein identification by searching sequence data bases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Qu,Y. *et al*. (2002) Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clini. Chem*., **48**, 1835–1843.

Rai,A.J. *et al*. (2002) Proteomic approaches to tumor marker discovery. *Arch. Pathol. Lab. Med*., **126**, 1518–1526.

Rockwood,A.L. and Haimi,P. (2006) Efficient calculation of accurate masses of isotopic peaks. *J. Am. Soc. Mass Spectrom*., **17**, 415–419.

Rockwood,A.L. *et al*. (1996) Ultrahigh resolution isotope distribution calculations. *Rapid Commun. Mass Spectrom*., **10**, 54–59.

Schwegler,E.E. *et al*. (2005) SELDI-TOF MS profiling of serum for detection of the progression of chronic hepatitis c to hepatocellular carcinoma. *Hepatology*, **41**, 634–642.

Senko,M.W. *et al*. (1995a) Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *J. AM. Soc. Mass Spectrom*., **6**, 52–56.

Senko,M.W. *et al*. (1995b) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom*., **6**, 229–233.

Soltys,S.G. *et al*. (2004) The use of plasma surface-enhanced laser desorption/ionization time-of-flight mass spectrometry proteomic patterns for detection of head and neck squamous cell cancers. *Clini. Cancer Res*., **10**, pp. 4806–4812.

Tibshirani,R. *et al*. (2003) Class predictaion by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Science*, **18**, 104–117.

Tibshirani, R *et al*., (2004) Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, **20**, 3034–3044.

Whitelegge,J.P. *et al*. (2004) Subtle modification of isotope ratio proteomics; an integrated strategy for expression proteomics. *Phytochemistry*, **65**, 1507–1515.

Wu,B. *et al*. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**, 1636–1643.

Yao,X. *et al*. (2001) Proteolytic ^{18}O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal. Chem*., **73**, 2836–2842.

Zlatkis,A. *et al*. (1979) Capillary column gas chromatographic profile analysis of volatile compounds in sera of normal and virus-infected patients. *J. Chromatogr*., **163**, 125–133.