Mono-isotope Prediction for Mass Spectra Using Bayes Network

Hui Li  
*Department of Systems and Computer Science, Howard University, Washington, DC 20059, USA.*

Chunmei Liu  
*Department of Systems and Computer Science, Howard University, Washington, DC 20059, USA.*

Mugizi Robert Rwebangira  
*Department of Systems and Computer Science, Howard University, Washington, DC 20059, USA.*

Legand Burge  
*Department of Systems and Computer Science, Howard University, Washington, DC 20059, USA.*

Follow this and additional works at: https://tsinghuauniversitypress.researchcommons.org/tsinghua-science-and-technology

Part of the Computer Sciences Commons, and the Electrical and Computer Engineering Commons

**Recommended Citation**
Mono-isotope Prediction for Mass Spectra Using Bayes Network

Hui Li, Chunmei Liu*, Mugizi Robert Rwebangira, and Legand Burge

Abstract: Mass spectrometry is one of the widely utilized important methods to study protein functions and components. The challenge of mono-isotope pattern recognition from large scale protein mass spectral data needs computational algorithms and tools to speed up the analysis and improve the analytic results. We utilized naïve Bayes network as the classifier with the assumption that the selected features are independent to predict mono-isotope pattern from mass spectrometry. Mono-isotopes detected from validated theoretical spectra were used as prior information in the Bayes method. Three main features extracted from the dataset were employed as independent variables in our model. The application of the proposed algorithm to publicMo dataset demonstrates that our naïve Bayes classifier is advantageous over existing methods in both accuracy and sensitivity.

Key words: Bayes network; tandem mass spectrum; mono-isotope prediction

1 Introduction

Tandem mass spectrometry (MS/MS) is an analytic technique that uses the degree of deflection of charged particles by a magnetic field to find the relative masses of molecular ions and fragments. The generated ions are then separated by electrostatic or magnetic fields or by a combination of both. It has become one of the critical technologies for biological molecular identification including metabolites, peptide, and protein. It has been widely used in various fields such as determining molecular mass, finding the structure of an unknown substance, verifying the identity and purity of a known substance, diagnosing disease process, and providing sensitive and specific means for pharmaceutical quality control[1, 2].

MS/MS data contains the intensity information for tens to hundreds of thousands of mass channels from a set of different mass analysis technologies. The isotope pattern of a peptide is determined by the elemental formula of the peptide and the abundance of heavy isotopes, usually the natural abundance, and therefore known. Collections of isotope peaks form isotopic distributions which give us multiple peaks, and hence multiple evidence sources for specific peptides[3-6]. The general process of MS/MS data analysis includes chemical separation procedures, electrospray ionization, matrix-assisted laser desorption/ionization, and mass analyzers. There are limitations of MS/MS technology such as the availability of high resolution instrument, quality of samples, less accurate detection for low abundance and short proteins[7, 8], which make it extremely urgent to automatically decompose the raw spectrum into a list of peptide masses. Such a step is needed for a variety of purposes, for example, determining isotopic clusters and their isotope patterns. Due to natural abundance of heavy isotopes, several peaks exist for each peptide mass at a given charge state. There are many issues to consider when analyzing proteomic mass spectrometry data for isotopic pattern recognition, such as mass-to-charge (m/z) shifts of common peaks between spectra, noise peaks caused by chemical and electronic noise during the sample acquisition within the spectra,
redundant features caused by isotopic distributions, various adducts, multiple charge states and peptide fragments occurring from proteolysis, overlap between isotope series in a peptide spectrum due to the presence of isotope series and multiple charge states for each peptide, and the fact that peptide masses tend to be clustered around certain values. All these limitations have left us with a set of challenges for improving the quality, efficiency, and interpretation of the biological molecules in the specific context.

Various methods have been proposed for iso
topic pattern recognition from MS/MS spectra. Most of these methods compare the experimental spectral pattern to a set of theoretical mono-isotope patterns. Comparison of theoretical and experimental isotope distributions is typically accomplished based on subtractive fitting and peak selection algorithms, attempting to sequentially detect the dominant components in a mixture spectrum. These subset selection methods attempt to determine a small set of basis functions capable of approximating the observed signal well. Facing the infeasibility of an exhaustive search over all possible subsets of explanatory basis functions, they apply greedy search strategies. The calculation of mono-isotope patterns is based on the estimation of average stoichiometries for a particular molecular mass or on relative isotope abundance estimation or on protein database driven mean isotope distribution calculation\[^9\]-\[^11\].

For example, Senko et al.\[^11\] introduced a notion of an “average” amino acid called averagine and suggested a computational method to determine mono-isotopic masses using it. ZSCORE\[^12\] is a fast and automated isotopic cluster identification algorithm that is based on a charge scoring scheme. Many other algorithms including THRASH\[^10\] employ the Fourier transform/Patterson method for charge determination and least squares fitting to compare a peak cluster with an average isotopic distribution. LASSO\[^13\] uses least squares fitting and/or average isotopic distribution, which often leads to an inaccurate mono isotopic mass that is 1-2 Da different from the correct value. Renard et al.\[^14\] proposed NITPICK which overcomes Kaur’s greedy iterative weighted least squares fitting approach to support the calculation of accurate mono-isotopic peak lists from raw mass spectra. It is specifically tailored to cases where the raw spectra stem from unknown, possibly overlapping experimental isotope patterns of multiple charge states. MATCHING\[^15\] performs deconvolution of overlapping isotope series only if the peptide sequences and therefore masses are known as a priori. Furthermore, PepList\[^16\], ESI-ISOCONV\[^17\], and AID-MS\[^18\] implement a top-down peak selection approach supplemented with novel charge state determination and other features to reduce false positive rates. These methods compare the observed isotope pattern to the expected isotope pattern.

In this paper, we propose a naïve Bayes classifier model to calculate the isotope distribution which regards peak intensities in an isotopic distribution as the existential probabilities of isotope compositions. Naïve Bayes Classifier (NBC) is generally known as a simple probabilistic graph classifier which assumes the independence of the features. This assumption reduces the complexity of the development of the classifier. NBC is one of the most effective and efficient inductive learning algorithms for data mining along with machine learning. NBC has proven its effectiveness in various domains such as text classification, image processing, fault prediction, and biomarker discovery\[^19\]-\[^21\]. It is of high computational efficiency as compared to other wrapper machine learning methods. It has low variance due to less searching, incremental learning due to working from approximation of low order probabilities deduced from training data, and high capability of handling noise and missing values in the dataset\[^22\]. The major pitfall of NBC is the assumption of feature independence\[^23\].

Compared with other existing methods, our model has the following features: (1) We introduce prior knowledge to predict the potential mono-isotope, and (2) Three features we extracted from the observed dataset characterize the structure of the network behind the observed data. Experimental results show that our algorithm outperforms THRASH in both accuracy and efficiency. In the following sessions, the first is overviewing the model. Then the experimental and simulation study are presented in detail. Finally, we give the conclusions and discussions.

2 Methods

We first preprocess the observed spectrum to remove the baseline, filter the noise, and generate a list of peptide candidates. Our preprocess method is to remove the negative intensities and then the whole spectrum via a fast Fourier transformation algorithm. The maxima
in the Fourier spectrum are eliminated and the mass spectrum is regenerated by a reverse fast Fourier. We estimate the parameters in the Bayes model from the manually annotated spectra using THRASH on a single spectrum. The overview flowchart of the method is shown in Fig. 1.

We start from data extraction from the dataset. We aim to build a full isotopic distribution to increase the precision. Our method is based on the statistical probabilities from distribution features including charge, ratio of intensity, and fitness score. The Bayes model is used to calculate the isotopic distribution probability to recognize their patterns from tandem mass spectrometry data. First, we extract the mass to charge value of the theoretical and observed mass spectra from the dataset using the THRASH\cite{10}. Automated reduction and interpretation of high resolution mass spectra of large molecules are performed by setting a threshold.

We then calculate the intensity ratio of the theoretical intensity to the observed intensity and charge value for each spectrum. The fitness score is the standard deviation error of the theoretical and observed intensities and charge values of each spectrum. We choose the mass to charge value determined by the Fourier transformation algorithm as one feature. Then the standard deviation error of the distance between two adjacent consecutive peaks forward within the distribution is used as another feature to measure the isotopic distribution. The hit scores for each spectrum are utilized as the third feature in our model.

The data set we used was downloaded from http://omics.pnl.gov. After manual downloading, we use the MSFileReader lib which was downloaded from the provider (http://sjsupport.thermofinnigan.com/) to read and obtain MS/MS fragmentation results to identify isotopic distribution. For further analysis, we program an in-house C++ tool to extract mass spectrum, intensity, and noise information for each \(m/z\) measurement from a collection of scans of samples in the format of raw data. Then we use known annotated mass spectrum as prior knowledge to train the Bayes network.

We use the scan number of the instrument to construct a function to obtain the mono-isotope and charge state information on a given scan number as shown in Fig. 2.

Isotope peaks in the spectrum of high resolution are able to be distinguished among the local maximum location. In a spectrum of low-resolution, they cannot be distinguished from the same location. In an isotope peak cluster, the most abundant mono-isotope peak is called a mono-isotopic mass. In a spectrum of high precision level, if there are no noise peaks, the mono isotopic parent mass can be obtained directly from the maximum peak value among the mass spectrum.

3 Algorithms

The algorithm starts from the head information which is extracted from the raw file. Then the correlation
between two peaks is calculated to obtain the charge of each spectrum. By ranking the entire spectrum, we can identify the local maximum peak value of the spectrum. The structure of Bayes network is shown as Fig. 3.

Next, we check the theoretical peak intensity distribution from a given mass spectrum. First, we get the experimental and theoretical peak intensity distributions with the THRASH\(^{10}\). Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules are used. Because only one distance value cannot give us the accurate result, we take the minimum distance corresponding to the formula and the ratio of the intensity distribution of the theoretical isotopic peaks into consideration. We use the three features to improve the accuracy. Figure 4 depicts the detailed process of extracting features from the observed peaks.

- Set a window of size +/-5 Da around the selected peak as shown in Fig. 4.
- Read the head information and get the actual mono peak value which is used as the training set for Bayes network model.
- Select the local maximum peak in the window.
- Compute the charge state using the correlation between observed intensity and the peak value.
- Use the average molecular mass to guess the molecular formula of the detected compound based on the mass and composition of the average amino acid determined from the Protein Informatics Resource (PIR) database.
- Use Mercury algorithm\(^{24}\) to generate the theoretical spectrum of the detected compound from the predicted molecular formula\(^{25}\).
- Extract the observed isotopic distribution from the theoretical isotopic distribution.
- Get the theoretical mono peak value.
- Calculate the difference between the theoretical mono peak value and the observed mono peak value and mark the positive or negative sample.
- Compare the theoretical and observed isotopic distributions to calculate an isotopic fitness value as feature 1 for the naïve Bayes model.
- Use the ratio of the observed isotopic distribution as feature 2 for the Bayes model.
- Use the charge state as feature 3.
- Output the file of three features and the positive or negative samples.
- Estimate the parameters from hand annotated spectra using THRASH on a single spectrum for the naïve Bayes classifier model.

We extracted the three features such as charge, fitness score of the peak value, and the ratio of intensity for each mass spectrum to estimate the isotope distribution. We estimate the parameters of the naïve Bayes model using the training data which were derived from annotated theoretical spectra.

Our naïve Bayes classifier works as follows: let \( X \) be a vector of random variables denoting the observed attribute values in the training set \( X = [x_1, x_2, \cdots, x_n] \) to class \( c \). For each mass spectrum, NBC produces a binary class \( C(0, 1) \) where 1 denotes that the spectrum is predicted as mono-isotope and 0 denotes non mono-isotope. The NBC was trained using a set of labeled training data set \( (X, C) \). In the binary classification, the class for the target residue was determined by comparing two posteriors with Eq. (1). The probability of each class given the vector of the observed values for the predictive attributes can be computed using the following formula:

\[
P(Y_j | X) = \frac{p(Y_j) p(X|Y_j)}{\sum_{i=1}^{c} p(Y_i) p(X|Y_i)}
\]

where \( p(Y_j) \) is the probability of the occurrence of \( Y_j \). \( P(Y_j | X) \) is the probability of observing attribute values being in class \( Y_j \). \( P(X|Y_j) \) is the probability of generating \( X \) given class \( Y_j \), which can be
computed with Eq. (2). The conditional independence assumption assumes that each variable in the data set is conditionally independent of each other.

\[ P(X|Y_j) = \prod_{i=1}^{n} p(x_i|Y_j), \ j = 1, \cdots, c \]  

(2)

where \( x_i \) is the value of the \( i \)-th attribute in \( X \) and \( n \) is the number of attributes. Let \( k \) be the number of classes, and \( c_i \) is the \( i \)-th class; the probability distribution over the set of features is calculated using the following equation:

\[ p(X) = \prod_{i=1}^{k} p(c_i) p(X|c_i) \]  

(3)

Then we have,

\[ \frac{p(c = 1|X = x_1, x_2, x_3)}{p(c = 0|X = x_1, x_2, x_3)} = \frac{p(c = 1) \prod_{i=1}^{n} p_i(x_i|c = 1)}{p(c = 0) \prod_{i=1}^{n} p_i(x_i|c = 0)} \]  

(4)

The mass spectrum is classified as 1 if

\[ \log\left( \frac{p(c = 1|X = x_1, x_2, x_3)}{p(c = 0|X = x_1, x_2, x_3)} \right) \geq \theta \]  

(5)

and 0 (non mono-isotope) otherwise, \( \theta \) is the threshold which was trained on the training set. The distance feature reflects the similarity between the theoretical isotope distribution and the observed distribution.

4 Results

We compare our method against THRASH for predicting the isotope distribution. We use four measurements to evaluate the performance:

- True positive if a mono-isotope peak is predicted in the isotope distribution.
- False positive if the false mono-isotope peak is predicted as positive.
- False negative if the mono-isotope is not found in the isotope distribution but it is actually a mono-isotope.
- True negative if the mono-isotope is not found in the isotope distribution.

We apply these measurements in our data analysis. Accuracy is computed as the ratio of the true mono-isotopes identified to the total number of identified mono-isotopes from the isotope distribution. Our method can get an accuracy of 89%, better than THRASH of 87%. In addition, we use 10-fold cross validation to generate ROC curves for our naïve Bayes classifier and THRASH as shown in Fig. 5. The plot of

![Fig. 5 ROC curves of our naïve Bayes Network and THRASH.](image)

our method lies above the other curve. This indicates that our method performs better than THRASH.

5 Discussion

We use the correlation method between two peaks to obtain the charge of each spectrum and select this value as one of the features to estimate isotopic distributions using average compounds around the mass of interest. The method we proposed allows the naïve Bayes model to produce better isotopic maps. Although results from our method demonstrate that our method is better than THRASH, we are trying other types of classifiers which may also improve the performance, i.e., Support Vector Machines (SVM), decision tree, and Artificial Neural Network (ANN). Also, including more beneficial features from other existing algorithms such as AID-MS to our current model may possibly improve the performance of the classifier.

The precision of identification of mono-isotope is affected by the overlapping distributions of different peptides. In this paper, we did not address fully this issue. In the future, we are planning to modify the algorithm to handle overlapping distributions.

Acknowledgements

This work was supported by an NSF Science and Technology Center, under Grant Agreement CCF-0939370 and 2 G12 RR003048 from the RCMI program, Division of Research Infrastructure, National Center for Research Resources, NIH.
References


Hui Li is currently a postdoctoral fellow at the Department of Systems and Computer Science in Howard University. He received his PhD degree in computer science from Beijing University of Technology in 2009. His research interests include computational biology, bioinformatics, pattern recognition, and algorithm.

Chunmei Liu received her BS and MS degrees in computer software from Anhui University in 1999 and 2002, respectively. She received her PhD in computer science from The University of Georgia in 2006. She became an assistant professor at the Department of Systems and Computer Science of Howard University in the same year. She has been a full professor since 2014. Her research interests include computational biology, graph algorithms, and theory of computation. Her recent research involves designing computationally efficient algorithms for protein identification, protein structure prediction, and protein-protein interactions. She is a member of IEEE and ACM.

Mugizi Robert Rwebangira received his BS degree in systems and computer science from Howard University in 2002 and his PhD degree in computer science from Carnegie Mellon University in 2008. He has been an assistant professor at Howard University since 2010. He has received grant funding from the Army Research Lab and the National Science Foundation and published in the areas of semi-supervised learning algorithms, computational biology, and voting theory. His current research interests are in transfer learning and computational sociolinguistics.

Legand Burge received his BS degree in computer and information science from Langston University in 1992 and his PhD degree in computer science from Oklahoma State University in 1998. He has been a full professor at Howard University since 2009. His current research interests lie in the field of distributed computing. The primary thrust of his current research is in global resource management in large-scale distributed systems. In particular, he is interested in middleware technology to support scalable infrastructures for pervasive environments capable of servicing a very large number of small (possibly mobile) distributed and embedded devices efficiently. He is also interested in the application of distributed high performance computing to solve computational science problems in Biology, Physics, and Chemistry.