Improved Model-based, Platform-independent Feature Extraction for Mass Spectrometry

Karin Noy a,b, Daniel Fasulo a

aIntegrated Data System Department, Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540, USA, bLife Sciences Department, Ben Gurion University of the Negev, Beer Sheva, Israel 84105

ABSTRACT
Motivation: Mass spectrometry (MS) is increasingly being used for biomedical research. The typical analysis of MS data consists of several steps. Feature extraction is a crucial step since subsequent analyses are performed only on the detected features. Current methodologies applied to low-resolution MS, in which features are peaks or wavelet functions, are parameter-sensitive and inaccurate in the sense that peaks and wavelet functions do not directly correspond to the underlying molecules under observation. In high resolution MS, the model-based approach is more appealing as it can provide a better representation of the MS signals by incorporating information about peak shapes and isotopic distributions. Current model-based techniques are computationally expensive; various algorithms have been proposed to improve the computational efficiency of this paradigm. However, these methods cannot deal well with overlapping features, especially when they are merged to create one broad peak. In addition, no method has been proven to perform well across different MS platforms.

Results: We suggest a new model-based approach to feature extraction in which spectra are decomposed into a mixture of distributions derived from peptide models. By incorporating kernel-based smoothing and perceptual similarity for matching distributions, our statistical framework improves existing methodologies in terms of computational efficiency and the accuracy of the results. Our model is parameterized by physical properties and is therefore applicable to different MS instruments and settings. We validate our approach on simulated data, and show that the performance is higher than commonly used tools on real high- and low-resolution MS, and MS/MS data sets.

Contact: daniel.fasulo@siemens.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Mass spectrometry (MS) is increasingly being used to discover proteomic markers and patterns in complex biological mixtures derived from tissues, or in more easily obtained fluids such as serum and urine. These markers can potentially be used as diagnostic and therapeutic targets, and for prognosis and therapy selection. Although existing gene expression microarray technologies have similar applications, proteomic analysis has the potential to provide additional understanding of the biological processes. It can provide not only the identities of proteins, but also information on their variant isoforms, splice variants, post-translational modifications, and how these variants differ in different cell types and tissues under different conditions and diseases (Kearney et al., 2003; Tyers et al., 2003).

The MS technologies most commonly applied to clinical and biological use are Matrix-Assisted Laser Desorption and Ionization (MALDI) (Baggerly et al., 2003; Tibshirani et al., 2004; Morris et al., 2005) and Surface Enhanced Laser Desorption and Ionization (SELDI) (Petricoin et al., 2002; Yasui et al., 2003; Coombes et al., 2005). Both technologies identify proteomic markers in one-dimensional MS data by their mass-to-charge ratios \((m/z)\). Liquid chromatography coupled with MS (LC-MS) and tandem MS technology provides a two-dimensional approach to proteomic profiling (Corthals et al., 1999; Gygi et al., 1999, 2002; Washburn et al., 2001; Zhou et al., 2002). Biological mixtures are physically resolved by chromatographic separation prior to MS to give additional information such as the LC retention time, the \(m/z\), and tandem MS associated with the markers. The additional level of separation may have several advantages over one-dimensional approaches such as better sensitivity and resolution. However, there are several additional computational challenges that must be addressed (Listgarten et al., 2005). For example, molecules in electrospray ionization (ESI) MS experiments are more likely to obtain multiple charges than molecules in MALDI and SELDI, which must be ascertained in order to compute their mass.

A typical MS data set contains tens to thousands of spectra, with each spectrum containing tens of thousands of intensity measurements representing an unknown number of molecules to be analyzed. The analysis of MS data is complicated by high levels of noise and the inherent high dimensionality of the data. The common approach for analyzing such data involves four major steps. First, low-level signal processing attempts to remove various types of noise from the spectra. Second, feature extraction aims to identify components associated with the underlying molecules in the spectra. Third, feature matching compares feature sets from the spectra in order to match recurring components. Lastly, statistical analysis and pattern detection aim to classify different conditions and diseases.

The feature extraction step is crucial since subsequent analyses are performed only on the detected features. Parameter-sensitive, inadequate, or incorrect methods can result in biased data sets which make it difficult to reach meaningful biological conclusions (Sorace et al., 2003; Baggerly et al., 2003, 2004). In low resolution MS data analysis, peak detection is the traditional method for extracting features; numerous techniques that try to identify peaks among the background noise have been proposed (Coombes et al., 2003; Yasui et al., 2003; Tibshirani et al., 2004). In contrast, multiscale wavelet decomposition approaches define peaks according to their
wavelet basis function coefficients (Coombes et al., 2005; Randolf et al., 2005; Morris et al., 2005; Du et al., 2006). However, peaks and wavelet functions do not directly correspond to the underlying molecules under investigation. In high-resolution MS data analysis, the model-based approach is more appealing as it can provide a better representation of the MS signals by incorporating information about peak shapes and isotopic distributions. There has been some prior work in model-based feature extraction specifically for MALDI peptide mass fingerprint identification (Gras et al., 1999; Berndt et al., 1999) and for MALDI ESI/FTMS peptide identification (Horn et al., 2000). While the former centers on the analysis of singly charged ions, the later handles multiply charged ions as well and targets the analysis of large biomolecules. Both approaches include additional processing steps for handling overlapping features. A key issue with these methods, however, is that they perform a huge number of regressions to fit signal models to spectra. In contrast, various algorithms that first perform peak detection and then group the detected peaks into isotope clusters have been proposed in order to speed up this paradigm (Breen et al., 2000; Wohlfarsky et al., 2001). Some of these methods have been applied to quantitative analysis of LC-MS data as well (Li et al., 2005; Bellew et al., 2006). However, by their nature, these methods cannot deal well with complex, overlapping features, in particular where the peak shape is degenerate due to overlapping features and resolution limitations.

In this paper, we suggest to return to the model-based approach in a different framework that improves existing methodologies in terms of computational efficiency and the accuracy of the results. In particular, we show that our method applies equally to different MS instruments and settings and therefore, will be valuable for both MS-based identification and quantification experiments.

In a proteomics experiment, MS is used to observe masses derived from a large, heterogeneous population of peptides or proteins; for brevity, we use the term “peptide” to refer to the molecules under observation although we recognize that some protocols observe undigested proteins, metabolites, or other biological molecules and our methods can apply to these as well. The observed \( m/z \) distribution from a given peptide is a function of the isotopic distribution of the elements comprising the peptide, the ionization source and the charge(s) it imparts, fragmentation and other modifications, and the resolution of the mass spectrometer determined by its separation and detection capabilities. We construct a closed-form model of this distribution for a given peptide by taking into consideration some of these phenomena and we show that the model fits real data quite well. We then consider a spectrum to be a histogram of observed masses from an unknown population of peptides. Since our model of the observed masses of peptides distributions provides us with a good estimate of the scale of the features in the histogram, we are able to employ kernel density estimation on the spectrum in order to obtain a better representation of the distribution of masses. In principle, our feature extraction method attempts to decompose the resulting density function into a mixture of distributions derived from the peptide models. Each component of the mixture is an application of a peptide model that we term a feature.

There are a number of advantages to our approach. First, most parameters in our model are based on the physical properties of the MS system and can be selected based on the system specifications or derived from empirical experiments; in contrast, many proposed approaches to smoothing and feature extraction require the selection of numerous non-intuitive parameters such as thresholds, wavelet functions, smoothing functions, coefficients, etc. Second, instead of performing a huge number of regressions to fit signal models to spectra, we perform regression only when the model fits the spectrum; we assess the fit by using cross-bin perceptual similarity for distribution matching. Third, our method can separate overlapping features in both high- and low-resolution MS even when they are partially convolved. Finally, our method applies equally to both the low-resolution spectrometry typically seen in MALDI and SELDI proteomics and high-resolution spectrometry seen in typical ESI and LC-MS systems; all prior work of which we are aware is targeted at either one application or the other.

We first use simulated data sets for an initial validation of our approach. Then, we have tested our method on both low- and high-resolution time-of-flight (TOF) data, and compared with previously proposed methods according to the type of data for which they were designed. For the low-resolution SELDI data, we compared our method with the continuous wavelet transform-based algorithm which has proven to perform better than commonly used techniques (Du et al., 2006). For the high-resolution ESI MS data, we compared our method with msInspect (Bellew et al., 2006), a well-known suite for the analysis of LC-MS data. For the high-resolution MALDI TOF/TOF data with MS/MS acquisitions, we show that our method identifies more peptides than by using the Applied Biosystems GPS software. In addition, we demonstrate the flexibility of our approach by analyzing the same sample acquired using different MALDI MS techniques, showing that we detect a substantial number of matching features.

2 METHODS

2.1 Modeling Features

A proteomics MS experiment can be modeled as a process in which an enormous population \( \mathbb{P} \) of ionized peptides represented by the multiset \( \mathbb{P} = \{ m_1, m_2, \ldots, m_N \} \), where each \( m_i \) is actual mass of molecule \( i \), is separated and detected. It is important to note that by peptide, we mean a set of molecules whose amino acid sequences are identical; when we say molecule we refer explicitly to a single molecule. The MS process can be thought of as mapping each \( m_i \) to an observed \( m/z \) ratio \( \tilde{m}_i = (m_i + z_i \cdot m_a + X)/z_i \), where \( z_i \) is the imparted charge, \( m_a \) is the mass of the charge particle (which may be negligible in the case of negatively charged ions) and \( X \) is a random variable representing error in the observation process. For simplicity of exposition, we ignore charge and consider \( z_i = 1 \), thus allowing the simpler model \( \tilde{m}_i = (m_i + X) \). However, we will describe later how we account for the charge.

2.1.1 Modeling the Distribution of Observed Masses from a Single Actual Mass. In practice, \( X \) follows some non-trivial distribution. This distribution is not generally part of the instrument specification, but detailed modeling of the MS process (Coombes et al., 2005) or empirical observations from controlled experiments can be utilized to generate a good estimate. What is generally available is information from the manufacturer on the spectrometer’s resolution. The standard MS definition of resolution is given by \( M/\Delta M \), where \( \Delta M \) is the minimum distance between the apexes of separable peaks. It is also common to specify \( \Delta M \) to be the Full Width at Half Maximum (FWHM) of a peak with \( m/z \) ratio \( M \), and to state the ideal FWHM value for one or more representative masses. The FWHM resolution typically varies from 600 for a low-resolution
MALDI-TOF spectrometer to 15,000 for a very high-resolution TOF analyzer, to 100,000 or more for an FT-ICR MS analyzer.

Because the best distribution model to use is specific to every instrument, we simply recommend some possibilities here. We have found that standard, easily-computed distributions approximate the peak functions well enough to obtain good results even if they are not precisely accurate (see Section 3.6). The Gaussian distribution is an obvious choice, and the standard deviation $\sigma$ at mass $m$ can be calculated using its relation to FWHM $^1$.

The Cauchy distribution (also known as the Lorentz distribution) is also appropriate to some instruments, and is directly parameterized by the Half Width at Half Maximum (HWHM), $\text{HWHM} = 0.5 \cdot \text{FWHM}$.

In the remainder of the paper, we let $p(m, R)$ be the peak function; that is, the density function of the observed masses from a single actual mass $m$ given the resolution model $R$.

2.1.2 Modeling the Distribution of Masses Associated with a Peptide. In the previous section, we described how the mass of a single molecule is subject to observation error, leading to a distribution of observed masses from a population of identical molecules. It is important to note, however, that peptides with identical sequences do not necessarily have identical masses. This is due to the isotope masses of the peptide elements. Since the ratios of the different isotopes occurring in nature is known, the set of masses associated with each peptide can be predicted, as can their expected relative abundances, from the peptide’s molecular formula (Yergey et al., 1983).

More specifically, let $m_0(P)$ denote the mass of peptide $P$ with all atoms in their lightest configuration, also referred to as the monoisotopic mass. Let $m_k(P)$ refer to the mass of the species of $P$ with $k$ additional neutrons, $0 \leq k < K_P$. The isotopic distribution of $P$ refers to the relative probabilities of observing the masses associated with the different values of $k$ given the peptide’s atomic composition. We denote this $\Pr[k|P]$. A number of algorithms have been proposed to quickly estimate this probability efficiently (Kubinyi et al., 1991; Rockwood et al., 2004; Yergey et al., 1983).

2.1.3 Modeling the Observed Mass Distribution from a Peptide. Given a peptide sequence, we first calculate the relative frequencies of each mass in the isotopic distribution associated with the peptide’s atomic composition. Then, we can calculate the distribution of observed masses for each isotopic mass. This yields a preliminary density estimate of the observed masses from the peptide.

We also need to account for potential charges. Unfortunately, we are not aware of any method to predict the likely distribution of charge states imparted to a peptide sequence from a given ion source. Thus, we use a more simple model: we select the maximum charge that is likely to be observed in a given experiment using our experience with MS. We then can parameterize our model by the charge state; to do this, we adjust the masses in the isotopic distribution appropriately by dividing by the charge and adding the positive ion mass if necessary.

We thus model the density function of the observed mass distribution for a peptide $P$ given a charge value $z$ and resolution model $R$ as follows:

$$ T(P, R, z) = \sum_{k=0}^{K_P} \Pr[k|P] \cdot p((m_k(P) + z \cdot m_0)/z, R) $$

We call this the template for peptide $P$ with charge $z$.

2.1.4 Mass Spectra as Complex Density Functions. Kernel density estimation is a method for estimating the density function of a continuous random variable (Wand et al., 1995); we present this method as an intuitive way for smoothing mass spectra. We start by considering a mass spectrum to be a histogram of the observed masses from a representative population of the peptides present in some sample. The histogram has the form $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The histogram bin widths and positions are a function of the nature of the mass spectrometer and its detector. The problems associated with using a histogram as a population density estimate have been well documented in the statistical literature (Wand et al., 1995). In the case where the scale and nature of subpopulations in the data is known, kernel density estimation provides a superior representation of the population density function. The observed mass distribution model gives us the required data scale and error model, making this technique ideal.

The kernel density estimator, $\hat{f}$, creates an estimate of the density by replacing each observation with a kernel function $K(x)$. The generic definition of $\hat{f}$ on a set of $n$ observations $x_i$ is:

$$ \hat{f}(x) = \frac{1}{hn} \sum_{i=1}^{n} K \left( \frac{x - x_i}{h} \right) $$

The smoothing parameter, $h$, is determined by the FWHM of the kernel according to the kernel function, which corresponds to the observation error distribution discussed in Section 2.1.1. Since we do not have information on the individual masses behind the spectrum, we could compute the density estimate from the spectrum as follows:

$$ \hat{f}(x) = \frac{1}{N} \sum_{i=1}^{n} K_{h}(x - x_i) \cdot y_i, $$

where $N = \sum y_i$ and $K_{h}(u) = (1/h)K(u/h)$.

2.2 A Practical Heuristic for Feature Extraction

In the previous section, we showed that a mass spectrum can be modeled as a histogram of observed masses containing many components. Each component is the observed mass distribution associated with a peptide in a particular charge configuration, which we can also model. In this section, we describe how we put our model to practical use for feature extraction.

We attempt to identify the peptides in the spectrum by first identifying peptide-containing regions, then detecting features via template generation and matching. Here, by “identify” we mean determine the monoisotopic mass, charge state, area, and intensity; determining the exact amino acid sequence from the mass value alone is unlikely to be possible in general.

2.2.1 Step 1: Isotopic Distribution Database. Release 50.6 of UniProtKB/Swiss-Prot contains 231,234 sequence entries (Boeckmann et al., 2003). It is impractical to compare each of these to each spectrum; in addition, such an approach inappropriately ignores sequences that are not represented in the database. Instead, we construct a model approximating the isotopic composition of peptides

---

of various masses, and compare the model to the spectrum. Specifically, we construct a database where we first calculate the isotopic distribution for each entry (Yergey et al., 1983). Then, the average distribution of all peptides within non-overlapping mass intervals of 500 Daltons are stored. A similar approach for the database construction was presented elsewhere (Horn et al., 2000; Gras et al., 1999). We found that within these intervals the isotopic distributions are virtually indistinguishable (data not shown), thus allowing reduction to a single representative distribution for each mass interval. Thus, in our model, we may replace $Pr[K|P]$ with $Pr[K|m]$, where $m$ is the mass of $P$, and still gain an acceptably accurate representation of the mass distribution of $P$. Our revised template model becomes:

$$T(m, R, z) = \sum_{k=0}^{K_m} Pr[k|m] \cdot p((m + k \cdot m_n + z \cdot m_a)/z, R) \quad (4)$$

where $m_n$ is the mass of a neutron and $K_m$ is set to a reasonable upper bound. It is important to observe that the maximum density does not always occur at the monoisotopic mass, since the probability of the monoisotopic composition decreases with peptide mass. Moreover, for low resolution values and high masses, the distinct monoisotopic peaks cannot be observed (see Supplemental Material).

2.2.2 Step 2: Spectrum Preprocessing. The purpose of this step is to minimize contributions from various types of noise. We first remove the baseline of each spectrum, which is a side effect of the MS sensor technology. For this purpose, we use a modified version of the top-hat operator that has been applied previously to MS (Sauve et al., 2004). In our modified version, we allow the width of the structuring element to increase with the $m/z$, as occurs in typical spectra.

We then smooth the spectrum by using a technique based on kernel density estimation as described in Section 2.1.4. For the sake of computational efficiency, we generate an interpolated approximation of the smoothed function. We calculate the smoothed spectrum at a set of $n'$ sample observations $X' = \{x'_1, x'_2, \ldots, x'_{n'}\}$. The set $X'$ is generated based on the resolution model. Given the resolution, we define the minimum distance between two separable masses in terms of the depth of the valley separating the peak functions. We convert the depth of the valley to a minimum difference in mass, and from there to a recommended sampling rate.

More specifically, to determine the local sampling rate in the area of $x$, we assume two Gaussian peaks of equal height representing minimally-separable masses, and derive a sampling rate that will allow us to separate these masses in the interpolated function. The separability of the masses is parameterized by the valley coefficient $v \in [0, 1]$, which is a fraction of the peak height at the local minimum separating the masses. The sampling rate $s(x)$ is expressed in terms of $\sigma(x)$, the standard deviation of the peak function, as follows:

$$s(x) = \sqrt{-2 \ln \left( (1 - v) \sigma(x) \sqrt{2\pi} \right)} \cdot \sigma(x) \quad (5)$$

Additional information about the derivation of this equation can be found in the supplemental material. Thus, to construct $X'$ from the original spectrum $\{x_1, x_2, \ldots, x_n\}$, we define $x'_1 = x_1$, and each successive $x'_{i+1} = x'_i + s(x'_i)$ until $x'_{n'} \geq x_n$. The values of the smoothed function $y'_i$ are calculated at each corresponding $x'_i$ using the Nadaraya-Watson estimator $\hat{M}(x)$, which is a direct extension of kernel density estimation to non-parametric regression (Wand et al., 1995):

$$y'_i = \hat{M}(x'_i) = \frac{\sum_{j=1}^{n} K_h(x'_i - x_j) \cdot y_j}{\sum_{j=1}^{n} K_h(x'_i - x_j)} \quad (6)$$

For practical purposes, we limit the domain of the kernel function to the interval where the function is non-negligible, for example $\pm 4\sigma$ for the Gaussian kernel.

2.2.3 Step 3: Identifying Peptide-containing Regions. The local maxima of mass spectra ideally correspond to observed masses, thus helping to localize regions where peptides may be present. Instead of evaluating every position in the spectrum as a potential peptide, we find the set of local maxima $\tilde{m}_1, \tilde{m}_2, \ldots, \tilde{m}_L$ in the smoothed spectrum $\hat{M}(x)$. These local maxima serve as the starting points for the search for peptide signals. These local maxima can optionally be filtered by a local signal-to-noise ratio threshold as is common in traditional peak detection methods; this sacrifices some sensitivity for speed. In our own work, we rarely use such a cutoff.

2.2.4 Step 4: Feature Detection via Template Matching. We process the local maxima in increasing order. At each $\tilde{m}_i$, for each possible $z$ we calculate the associated de-charged mass $m' = z(\tilde{m}_i - m_a)$ and the template $T(m', R, z)$ is generated. It is aligned to the spectrum by placing the first non-negligible local maximum in the template in correspondence with $\tilde{m}_i$. For each of these templates, we assess the fit to the observed spectrum via metrics for matching shapes or distributions. Thus we avoid the need to calculate the best least-squares fits of every template to the spectrum prior to deciding if there is a match, which increases the efficiency of our method.

Our matching procedure is as follows: the Pearson correlation coefficient is first used to assess the fit between the template $T(m', R, z)$ and the spectrum. We set two correlation coefficient cutoffs, $c_1$ and $c_2$ where $c_1 > c_2$. These parameters are somewhat subjective; we set $c_1 = 0.90$ and $c_2 = 0.75$ and have observed good behavior across the wide range of data sets presented in our results. If one or more templates exceed $c_1$, or if just a single template exceeds $c_2$, we take the strongest scoring template as a match. Otherwise, if multiple templates exceed $c_2$, an additional voting scheme is applied to determine the most appropriate match. The scheme includes three voters: the Pearson correlation, a convolution-based score, and the Earth Mover’s Distance (EMD) that has been used successfully for perceptual similarity measures in imaging (Rubner et al., 2000). We first calculate the EMD between the template and the region of the spectrum where it is aligned. We then convolve the template with a slightly broader region of the spectrum around the potential matching area, and take the maximum convolved value. If a majority of these voters select the same template as the best match, then that template is considered to be the match for the region. Otherwise, no match is made.

Once a match is found, the template is shifted in a small mass window around its original location. At each mass position, linear regression is used to fit the template intensities to the actual spectrum. The best fit among these candidates is taken, and the monoisotopic mass, charge, area, and peak intensity are reported. The scaled and translated template is then subtracted from the spectrum, the list

4
of local maxima is updated to reflect the modified spectrum, and the process is restarted.

3 RESULTS
We have tested our method on both simulated data and real data from various MS instruments and protocols.

3.1 Validation on Simulated Data Set
It is difficult to evaluate the performance of a given method without knowing the true molecular composition of the samples used in the experiments (Coombs et al., 2005). In order to provide robust quantitative performance metrics, we have developed a simulation tool with which we validate the correctness of our methods. The simulation procedure is comprised of the following steps. First, we provide peptide sequences as input, and calculate the specific expected isotopic distribution associated with each peptide, yielding a list of masses with relative frequencies. Second, we generate observed mass distributions (peaks) for each mass at the desired resolution. Finally, we scale the distributions according to the desired peptide signal intensities, sum the results to generate a simulated spectrum, and add Gaussian noise.

We tested our methods on both high and low resolution simulated data sets. The first simulated data set is comprised of tryptically digested β-galactosidase peptides taken from the Uniprot database (Boeckmann et al., 2003) in both 12,000 (Fig. 1(a)) and 2,000 FWHM values (Fig. 1(b)). Another set contains only Gaussian noise. These data sets contain many cases where feature overlap due to low resolution and closely spaced masses. For example, the β-galactosidase data set contains masses at 2,442 and 2,445 which merge to form two overlapping broad peaks (Fig. 1(b)). Nevertheless, in both the simulated data sets, we detect the correct features in these experiments with 100% sensitivity and a False Discovery Rate (FDR) of 0%, and we detect no features in the Gaussian noise.

3.2 A Reference Low-resolution MS Data Set
Next, we tested our method on a reference SELDI MS data set of known polypeptide compositions. The data is from the All-in-one Protein Standard (Ciphergen Biosystems Inc.) acquired using a Ciphergen NP20 chip, downloaded from the CAMDA2006 forum (CAMDA, 2006 http://www.camda.duke.edu/). There are seven polypeptides in the sample with the \( m/z \) values of 7,034, 12,230, 14,520, 16,952, 29,023, 46,671, 66,433 and 147,300; we detected all the seven polypeptides with both one and two charges, and several with higher charges as well (Fig. 2(a)). We compared our results with the CWT-based peak detection algorithm that was previously shown to perform better than other commonly-used techniques on this data set (Du et al., 2006). Interestingly, our results are comparable to the CWT-based method except for one true signal with \( m/z \) value of 14,520 which this method seems to miss (Fig. 2(b)). Also, we noted that both method and the CWT-based method agreed that there were several features present that are not part of the standard mixture, thus illustrating the difficulty of using real data as a benchmark in MS.

3.3 A High-resolution LC-MS Data Set
We also tested our methods on high resolution LC-MS data from Applied Biosystems QStar XL. In this data set, an isotope labeling method for quantification is used (Flory et al., 2002). The samples are comprised of isotopically-labeled peptides incorporating the isotopes \( ^{18}O \) and \( ^{16}O \) from water molecules under different conditions. This labeling protocol results in partially overlapping features with a mass shift of 4 to be detected and analyzed. These partially overlapping features are referred to as feature pairs. We compared our results with msInspect, a well-known suite for the analysis of LC-MS data (Bellew et al., 2006) where the detected features are two-dimensional (both LC and MS). With our method, we detected features separately in each scan and then used a simple matching
method unify results from consecutive scans. We formed feature pairs using the same simple approach for both programs. Overall, we found more feature pairs by using our methods; according to msInspect, there are 122 feature pairs while by using our methods, we were able to detect 151. Figure 4(a) shows an example of a feature pair at the m/z values 543.63, 544.97 with charge state 3 that was detected by our methods and not by msInspect. We visually inspected all feature pairs that were found by our program but not by msInspect, and found no false positives.

3.4 Application to Protein ID via MALDI-TOF MS/MS

In order to evaluate whether our method makes a difference in identifying proteins via tandem MS, we tested our methods on the Aurum data set (Falkner et al., 2007). This high-resolution data set of known purified and trypsin-digested protein samples was generated on an ABI 4700 MALDI TOF/TOF with MS/MS acquisitions. We randomly picked 20 proteins from this study and analyzed their corresponding set of tandem mass spectra. The features detected by using our methods were submitted to the X!Tandem search engine for peptide identification (Craig et al., 2004). We compared our results with X!Tandem results for the peak lists provided in this study, generated using the Applied Biosystems GPS software (Falkner et al., 2007). Using our feature extraction method, 135 peptides were identified by X!Tandem with an average p-value of −53, while 124 peptides with an average p-value of −51 were identified using peak lists from the GPS software. The average number of peptides identified per protein was higher with our method than with GPS (8.57 versus 7.76, p = 0.02), as was the number of unique peptides identified per protein (6.43 versus 5.90, p = 0.004). Although both methods found the correct proteins, we found some evidence of increased sensitivity and specificity. In terms of sensitivity, 3 additional contaminant proteins—Trypsin precursor, KERATIN 1, KERATIN 10—that are usually seen in MS experiments were detected by our methods. With the GPS software, Only KERATIN 1 was found. In terms of specificity, 6 apparent false positive proteins were also found with the GPS software, and only 2 with our methods. More detailed information and search engine results can be found in the supplemental material.

3.5 Analysis of A Sample Acquired in Two MS Modes

Finally, we tested our methods on both low- and high-resolution MS analyses of the same biological sample. The sample was processed on a MALDI-TOF instrument in reflectron (high-resolution) mode, and then the same sample was reprocessed on the same instrument linear (low-resolution) mode. It is difficult to quantify the performance of our methods since the content of the samples is not known. Therefore, we use the high-resolution data set as a reference to determine whether the correct features in the low-resolution data set were identified. We first detected features in the high-resolution data using our method. Next, we detected features in the low-resolution data by using both our method and the peak detection method in the PROcess R library with the recommended parameters 2. Then, we calculated how many detected features in the low-resolution data matched the detected features in the high-resolution. We limited our analysis to the mass region of the spectrum where both MS techniques showed comparable sensitivity.

In the high resolution spectrum, we found 52 features. In the low resolution spectra, we found 40 and 45 features using our methods and PROcess respectively. Note that PROcess detects only peaks so we did not apply this method on the high-resolution data. We next matched these features to the features found in the reference high-resolution data. Overall, 32 features found by our method had matches, while 29 of the features found by PROcess matched. Unmatched features, which are apparently false positives, account for 20% and 36% of the low resolution features from our method and PROcess respectively.

In general, we found that when the peaks are clearly separated, the results are quite comparable (Fig. 3(a)). However, when peaks overlap, the peak detection method in the PROcess library can not distinguish between them, resulting in missed features (Figures 3(b) and 3(c)). PROcess did seem to detect a few low-intensity features which our method did not, at a cost of substantially more false positives. In general, given the obvious sensitivity and resolution differences between these two spectra, we were pleased to see such a high level of agreement.

3.6 Kernel-Based Smoothing and Performance

The distribution associated with a peptide or actual mass is specific to every MS instrument. In the previous analyses, we utilized the Cauchy distribution as the kernel function. To validate this choice,
The analysis of MS data is complicated by the high level of noise and the inherent high dimensionality of the data. We apply our methods to the analysis of SELDI-TOF MS data, which is particularly difficult due to the complexity of the spectrum and low resolution of the instrument, and we show that we are able to find the expected features in a standard reference protein mixture. In high resolution LC-MS data, where the isotopic distributions can be clearly seen, our methods perform better than the commonly-used software package msInspect. We also show that our methods can make a difference in the identification of proteins via tandem MS by identifying more peptides than the commonly-used software GPS. In addition, we employ simulated data to benchmark the performance of our methods in a context where the correct results are completely known. We have shown that our method performs well on various types of samples, resolutions, and complex cases where features overlap and therefore are not easily detected.

The metrics for matching distributions is based on local best-fit heuristic that comprised of cross-bin perceptual similarity measure rather than bin-to-bin similarity. We found that in some cases (in particular, in low-resolution data), bin-to-bin similarity measure indicates multiple matched templates (data not shown). In this case, additional score based on cross-bin similarity is used to select the best fit template. The combination of kernel smoothing and cross-bin perceptual similarity matching yields a highly efficient, yet accurate method for feature detection across a variety of MS platforms.

5 ACKNOWLEDGEMENTS

We thank Dr. Catherine C. Fenselau and Dr. Nathan Edwards from the University of Maryland for providing the LC-MS data. We also thank Chandan Reddy from Cornell and Dr. Kaz Okada from San Francisco State University for helpful discussions. Finally, we thank Nir Kalisman from Ben Gurion University and the anonymous reviewers for valuable suggestions and feedback.

REFERENCES


Figure 4. Feature pairs from high-resolution LC-MS isotopically-labeled data set. The squares indicate the monoisotopic peak. (a) Feature pair with charge state 3. (b) A kernel-based smoothed version of (a).

Figure 5. Running time under different scenarios. The size of the input spectrum is presented in terms of the number of sample points.
by denoising spectra with the undecimated discrete wavelet transform., Proteomics., 16, 4107-4117.


