Peptide charge state determination for low-resolution tandem mass spectra

Aaron A. Klammer
Department of Genome Sciences, Box 357730
1705 NE Pacific Street #J205
Seattle, WA 98195-7730, USA
aklammer@u.washington.edu

Christine C. Wu
Department of Pharmacology, UCHSC
Fitzsimons RC1 South L18-6117
P.O.Box 6511, Mail Stop 8303
Aurora, CO 80045, USA
christine.wu@uchsc.edu

Michael J. MacCoss
Department of Genome Sciences, Box 357730
1705 NE Pacific Street #K328A
Seattle, WA 98195-7730, USA
maccoss@gs.washington.edu

William Stafford Noble
Department of Genome Sciences, Box 357730
1705 NE Pacific Street #J205
Seattle, WA 98195-7730, USA
noble@gs.washington.edu

Abstract

Mass spectrometry is a particularly useful technology for the rapid and robust identification of peptides and proteins in complex mixtures. Peptide sequences can be identified by correlating their observed tandem mass spectra (MS/MS) with theoretical spectra of peptides from a sequence database. Unfortunately, to perform this search the charge of the peptide must be known, and current charge-state-determination algorithms only discriminate singly- from multiply-charged spectra: distinguishing +2 from +3, for example, is unreliable. Thus, search software is forced to search multiply-charged spectra multiple times.

To minimize this inefficiency, we present a support vector machine (SVM) that quickly and reliably classifies multiply-charged spectra as having either a +2 or +3 precursor peptide ion. By classifying multiply-charged spectra, we obtain a 40% reduction in search time while maintaining an average of 99% of peptide and 99% of protein identifications originally obtained from these spectra.

Availability: Supplementary data at `http://noble.gs.washington.edu/papers/klammer_peptide.html`. Binary executables available from authors upon request.

Keywords: mass spectrometry, proteomics, charge state, machine learning, support vector machine

1 Introduction

A major goal in modern biology is the identification and characterization of the cell's entire protein complement, or proteome. Towards this end, mass spectrometry-based technologies offer the ability to rapidly identify proteins in complex mixtures (9; 16). In a common approach, the cell's entire unfractionated protein mixture is digested to peptides and the peptides are then separated by micro-capillary liquid chromatography followed by electrospray-ionization collision-induced-dissociation (ESI-CID) tandem mass spectrometry (μLC/MS/MS). The peptide sequences are then identified by correlating their respective MS/MS fragmentation spectra against predicted spectra of peptide sequences obtained from a protein sequence database.

Modern mass spectrometers can acquire more than five spectra per second, resulting in over 400,000 MS/MS spectra per day per instrument. Although searching each spectrum is fast, the sheer amount of data makes the total search step slow, especially for large sequence databases. This situation is aggravated by the need to search some spectra multiple times: mass spectrometry measures mass-to-charge ratios (m/z), but most search algorithms identify candidate peptides based on mass. Hence, if a peptide charge is ambiguous, as with low-resolution multiply-charged spectra, search algorithms are forced to search candidate peptides at multiple masses, one mass for each possible charge. Unfortunately, searching spectra multiple times has become an unmanagable computational burden because of the rapid expansion of genomic sequence information, the increas-

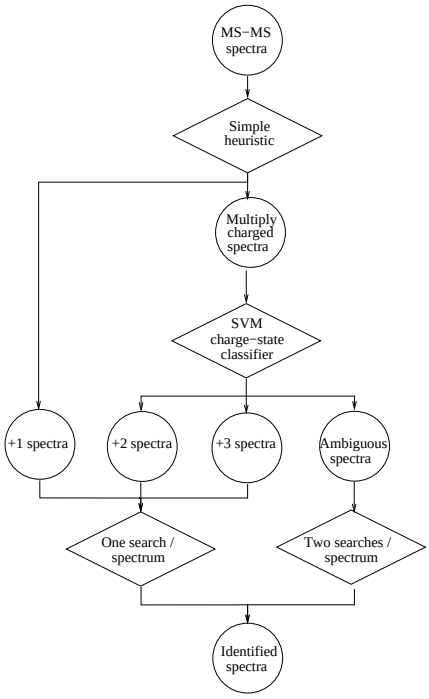


Figure 1. Improved spectrum search strategy. Current search methods require multiply-charged spectra to be searched against a sequence database at least twice, once assuming a precursor-ion charge of $+2$ and once assuming a charge of $+3$. A trained SVM classifier allows searching multiply-charged spectra only once, as either $+2$ or $+3$, and only a small remaining fraction of ambiguous spectra twice, reducing search time for multiply-charged spectra time by nearly 50%.

ing speed of tandem mass spectrometers, and the increased complexity of protein samples being characterized.

This burden can be alleviated with a spectrum search method that incorporates a rapid charge-state-determination step (Figure 1), avoiding redundant searching for multiple-charged spectra. Singly-charged spectra can be reliably distinguished from multiply-charged spectra with currently existing algorithms (14). There have been two previous reports of charge state determination for multiply charged low-resolution ESI-CID MS/MS fragmentation spectra of peptides. Though these approaches provide some discriminatory power among different charge states, the results still leave significant room for improvement. One simple approach reported by (13) counts the number of fragment pairs present in the spectrum that when summed equal the peptide mass derived from each respective charge state. The charge state that has the greatest number of pairs that sum to the respective peptide mass is chosen. This method works well for sparse spectra; however, classification of spectra with signal at nearly every m/z is less effective, possibly because the intensity of the fragment ion pairs is not factored into the approach.

In another effort, (3) use several methods to categorize spectra into $+1$, $+2$, $+2/+3$, $+3$ or $+4$ classes. Their first method calculates a posterior probability for each charge state based on a multinomial distribution which models the probability of fragment ions in different m/z ranges. The group's second method also calculates a posterior probability for each charge state, but uses complementary ion pairs as described in (5). Their third method combines these two probabilities, into a final, superior, posterior probability. All methods, however, have difficulty discriminating between the two most common multiply charged peptides, $+2$ and $+3$.

Here we present a support vector machine (SVM) (2; 15; 4) classifier that improves on existing methods in its ability to classify large numbers of spectra while maintaining virtually all true positive peptide and protein IDs. An SVM allows separation of two classes of data; thus, it is well-suited to the problem of separating the two most common forms of multiply-charged peptides, $+2$ and $+3$. In this case, we use the SVM to classify spectra into three categories: high-confidence $+2$, high-confidence $+3$, with the remaining spectra considered ambiguous (Figure 1). Spectra classified into either $+2$ and $+3$ charge states are searched only once, thus decreasing search time for multiply-charged spectra by up to 50%.

To train our SVM, we extract a set of spectra with high-confidence charge-state assignments from various data sets, where the charge state was derived by database search using a normalized version of SEQUEST (8). Each spectrum is summarized as a vector of features. Some features are expected to discriminate *a priori*, such as those that measure paired ions resulting from either a $+2$ or $+3$ fragmentation (Figure 2). Other features exploit empirically observed differences in $+2$ and $+3$ spectra (Figure 3).

The SVM's performance is evaluated by examining the trade-off between spectra classified (and thus time saved) and peptide or protein identifications. We find that a classifier trained on one data set can reduce the number of database searches for multiply-charged spectra by 40% while maintaining an average of 99% all protein and peptide identifications. We see this classifier as being useful both in conjunction with other filtering methods, such as the spectrum quality filter proposed by (1), as well as a stand-alone pre-filtering step for sequence database search algorithms.

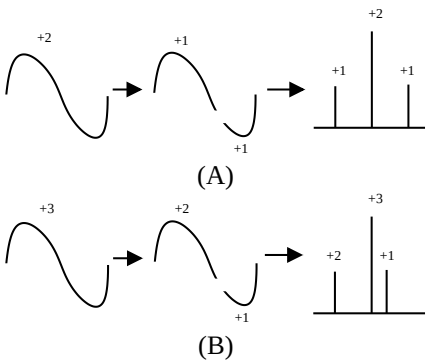


Figure 2. Fragmentation patterns of charge $+2$ and $+3$ precursor peptides. (A) Peptides of charge $+2$ tend to fragment into two $+1$ fragment ions with m/z values that are symmetric about the precursor m/z, and with predicted masses that sum to twice the original peptide m/z (7). (B) In contrast, peptides of charge $+3$ fragment predominantly into $+1$ and $+2$ fragment ion pairs with m/z values distributed asymmetrically, and with predicted masses that sum to three times the precursor m/z.

2 Algorithms

For the SVM to classify spectra according to their charge state, each spectrum is converted into a vector representation. A good representation makes the SVM's job easy by including features that differ between $+2$ and $+3$ charged spectra. This vectorization step thus corresponds to the encoding of our prior knowledge about how charge state affects a given spectrum. Below, we describe 19 distinct features, each of which we expect *a priori* to differ according to the charge state of the spectrum; in addition, we present 15 features derived from empirical differences noticed by the authors (Figure 3). Some of these features are more informative than others, but the combination of all 34 features allows the SVM to accurately predict a spectrum's charge state.

The first type of feature relies upon the intuition portrayed in Figure 2 that precursor ions usually fragment into a pair of ions, the masses of which sum to the original precursor ion mass. For a $+2$ precursor ion, the fragmentation into pairs of $+1$ ions is more common than into a $+2$ and a $+0$ ion due to the repulsion of like charges (7). Hence, the m/z values of the fragments of a $+2$ precursor will usually sum to twice the m/z of the original precursor ion, and assuming no secondary fragmentation, these ion pairs will be measurable. Thus, our first feature measures the extent to which such ion pairs occur in the spectrum using a form of correlation. Assuming we have a precursor ion with m/z m_p, and a spectrum $S = S_0 \ldots S_{m_{max}}$, where S_i is the sum of all spectrum peaks within 0.5 m/z of the m/z value i, the $+2$ correlation feature is

$$X_{+2} = \sum_{i=0}^{m_{max}} S_i S_{2m_p - i}. \tag{1}$$

This feature is expected to be greater for $+2$ ions than for $+3$ ions. An analogous correlation can be used to test for fragment ion pairs generated from a $+3$ precursor ion:

$$X_{+3} = \sum_{i=0}^{m_{max}} S_i S_{\frac{3m_p - i}{2}} \tag{2}$$

Finally, because these two features are complementary to each other, the ratio of the first two features is used as a third feature:

$$X_{+3/+2} = \frac{X_{+3}}{X_{+2}} \tag{3}$$

For a $+2$ ion, the numerator of this fraction will be small and the denominator will be large, yielding a small value, and vice versa for a $+3$ ion.

During fragmentation, many precursor peptides lose small groups of molecules with no charge. Common losses are of carbon monoxide, water and ammonia (7). Therefore, three additional triplets of features are defined corresponding to these three types of common losses. The formulas defining these features are analogous to Equations 1–3, except that the precursor mass term is replaced with some other fixed mass. For example, after a loss of carbon monoxide, the first of the three features is

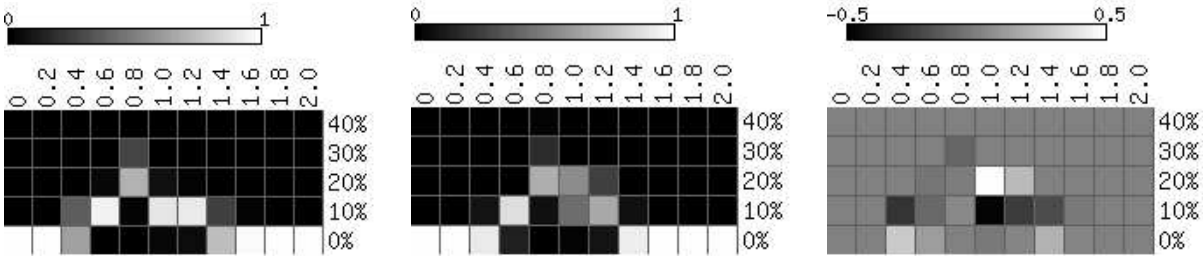
$$X_{+2}^{CO} = \sum_{i=0}^{m_{max}} S_i S_{2m_p - i - m_{CO}}, \tag{4}$$

where m_{CO} is the mass of carbon monoxide.

The next feature also relies upon the observation that $+2$ precursor ions tend to fragment into pairs of $+1$ ions whose masses sum to the precursor mass. Furthermore, the intensities of the paired ions generated by a given fragmentation event will likely have similar intensities, because secondary fragmentation is rare in radio-frequency only ion-trap mass spectrometers ((7)) (e.g. the LCQ and LTQ). Therefore, the sum of the intensities of the peaks above and below the precursor ion mass should be approximately equal:

$$\sum_{i=0}^{m_p - 1} S_i = \sum_{i=m_p+1}^{2m_p} S_i,$$

where $2m_p$ is the maximum possible fragment ion generated from a $+2$ precursor. Based upon these observations,



(A) $+2$ spectra (B) $+3$ spectra (C) Difference

Figure 3. The difference between the shape of the distributions of fragment ions from the MS/MS spectra of peptide precursor ions of $+2$ and $+3$ charge states. In a given set of spectra, the X-axis of each spectrum is rescaled so that the precursor m/z is equal to 1 and the peaks sum to 1. Peaks are then binned along the m/z axis, with each bin containing the sum of the peaks within the bin. In the heat maps (A) and (B), each column is a distribution across all spectra in the collection. Thus, the intensity of the square in row X, column Y is proportional to the percentage of spectra for which the sum of the peaks in the Xth m/z bin is Y. (A) Distribution of peak distributions for 128 $+2$ charged spectra. (B) Distribution of peak distributions for 353 $+3$ charged spectra. (C) The difference between (A) and (B). All spectra are from the E-LTQ-1 data set.

we define the "$+2$ balance" feature as the normalized difference between the total ion intensities above and below the precursor ion m/z:

$$B_{+2} = \frac{|\sum_{i=0}^{m_p-1} S_i - \sum_{i=m_p+1}^{2m_p} S_i|}{\sum_{i=0}^{2m_p} S_i} \quad (5)$$

If the spectrum was generated from a $+2$ precursor ion, then we hypothesize that the spectrum will be approximately balanced around the precursor, and this feature will have a small value.

In an analogous fashion, we can define a "$+3$ balance" feature. In this case, we assume that most fragmentations will yield a $+2$ and a $+1$ ion (rather than an uncharged ion and a $+3$ ion), and we again assume that secondary fragmentation is rare. Consequently, if we consider the m/z range up to $3m_p$ (i.e., the mass of the precursor), then we expect the sum of the peak intensities in the first one-third of this range to equal the sum of the intensities in the remaining two-thirds:

$$B_{+3} = \frac{|\sum_{i=0}^{m_p-1} S_i - \sum_{i=m_p+1}^{3m_p} S_i|}{\sum_{i=0}^{3m_p} S_i} \quad (6)$$

Note that a similar formula could be written for $+4$ ions, including terms for both $+3/+1$ and $+2/+2$ ion pairs.

Another easy-to-identify difference between spectra generated from $+2$ and $+3$ precursor ions appears near the top of the m/z range. In general, $+2$ precursor ions should never generate fragment ions with m/z greater than $2m_p$, whereas $+3$ precursor ions can do so, albeit rarely. Therefore, a "high m/z" feature can be defined as the percentage of the total spectrum intensity that occurs between $2m_p$ and $3m_p$:

$$H_{m/z} = \frac{\sum_{i=2m_p+1}^{3m_p} S_i}{\sum_{i=0}^{3m_p} S_i}. \quad (7)$$

This feature should be close to zero for $+2$ ions and positive for $+3$ ions.

All 15 features described above are derived from MS/MS spectra. However, the precursor ion's relationship to other ions in the initial MS scan can also be used to determine its charge state. Specifically, a $+2$ ion is often observed with a corresponding $+3$ ion, and vice versa. Thus, a $+2$ precursor ion at m_p in the MS "survey" scan might have a corresponding $+3$ ion at $\frac{2m_p+m_{H+}}{3}$. Conversely, a $+3$ precursor ion at m_p may have a corresponding $+2$ ion at $\frac{3m_p-m_{H+}}{2}$. Thus three additional features (two features and their ratio) can be extracted from an MS spectrum, $P = P_0 \ldots P_{m_{max}}$:

$$P_{+2\to+3} = P_{\frac{2m_p+1}{3}} \quad (8)$$

$$P_{+3\to+2} = P_{\frac{3m_p-m_{H+}}{2}} \quad (9)$$

$$P_{+3/+2} = \frac{P_{+3\to+2}}{P_{+2\to+3}} \quad (10)$$

In an MS/MS scan resulting from a $+3$ precursor ion, one might observe $+2$ and $+1$ ion species of the same fragment. Thus, one might distinguish $+3$ precursors from $+2$ precursors by a correlation in which each putative $+1$ ion is matched to its $+2$ counterpart. We define a feature that

computes such a correlation using only ions that are likely to be $+1$, that is, those that have m/z values greater than m_p:

$$X_{+1,+2} = \sum_{i=m_p+1}^{m_{max}} S_i S_{\frac{i+m_{H+}}{2}} \quad (11)$$

The final set of features is inspired by the observation, illustrated in Figure 3, that the distribution of fragment m/z values across a spectrum relative to the precursor m/z are likely to be different for $+2$ precursor ions versus $+3$ precursor ions. Many of the features described above attempt to encode intuitions about how these two distributions should differ. The final set of features, in contrast, assumes nothing about these distributions, except that they are different. A vector is defined of $3n$ elements, where the ith element is the fraction of total ion intensity contained within the m/z range $(\frac{i-1}{n}m_p, \frac{i}{n}m_p]$. In the experiments reported here, we use $n = 5$, resulting in 15 features.

Prior to calculation of any of these features, all peak intensities are replaced with their square root to reduce dynamic range and then normalized so that their sum is 1.

3 Methods

In this section we first describe the raw data sets used and a method for extracting high-confidence $+2$ and $+3$ training examples from these data sets. Next, we outline the method for training the SVM from this data. Finally, we present the method for estimating the number of true positive and false positive peptide and protein IDs in the test data sets, values which are necessary for evaluating the success of the SVM classifier.

3.1 Data sets

We analyze nine separate data sets (Table 1) and one hybrid data set (see Supplementary Data for generation of the hybrid data set). Four replicate data sets (E-LTQ-1, E-LTQ-2, E-LTQ-3 and E-LTQ-4) were generated using an LTQ ion-trap mass spectrometer (ThermoElectron, San Jose, CA) from a single *E. coli* protein digest. An additional four replicate data sets (E-LCQ-1, E-LCQ-2, E-LCQ-3 and E-LCQ-4) were generated from separate but identically prepared protein digests using an LCQ-XP Max mass spectrometer (ThermoElectron, San Jose, CA). The ninth data set (S-LCQ) is publically available and described by (6). The S-LCQ data set was also generated on an LCQ, but from a protein sample containing a mixture of 18 commercially available proteins.

For clarity of exposition, we group these data sets into three phases, I, II and III. In addition, five of the nine data sets are randomly sampled from to produce a hybrid training data set (see Supplementary Data) that we use to train an additional classifier.

Table 1. Ten data sets used to test and train the SVM. We use each of the first three data sets listed below (Phase I) to train a separate SVM classifier. We test each classifier on four other data sets: the remaining two sets from Phase I and both Phase II data sets. We train a final classifier on a hybrid data set generated from random samples of each of these first five data sets, and test it on four additional testing-only data sets (Phase III). Each column below lists the total number of spectra (Total), the number of singly-charged spectra (+1), the number of multiply-charged spectra (Multi) and the number of high-confidence $+2$ and $+3$ charge spectra for each data set.

	Data set	Total	+1	Multi	$+2$	$+3$
I	S-LCQ	19000	504	18496	1640	992
I	E-LCQ-1	4333	2823	1510	75	124
I	E-LTQ-1	29823	7904	21919	353	128
II	E-LCQ-2	4234	2928	1306	53	95
II	E-LTQ-2	32222	8638	23584	492	131
I+II	Hybrid	—	—	—	577	624
III	E-LCQ-3	3780	2479	1301	74	153
III	E-LTQ-3	32757	8600	24157	489	115
III	E-LCQ-4	4468	2734	1734	119	278
III	E-LTQ-4	32459	8746	23713	496	139

E-LCQ and E-LTQ data sets To generate the E-LCQ and E-LTQ data sets, aqueous soluble proteins from an *E. coli* lysate were reduced, carbamidomethylated and digested with trypsin in the presence of an acid labile detergent (RapiGest, Waters Corp) as recommended by the manufacturer. The resulting peptides were analyzed by μLC/MS/MS using data-dependent acquisition. The resulting spectra were searched against the 2004-May-02 *E. coli* Refseq protein database using a normalized version of SEQUEST (8). The specific SEQUEST search details can be found in the Supplementary Data.

S-LCQ data set The S-LCQ data set was downloaded from `http://www.systemsbiology.org/protein_mixture.html` (6). Only spectra that matched a peptide sequence of one of the proteins expected to be in the sample were used for this analysis.

3.2 Charge state assignment

From each of the four training data sets, we extract a set of multiply-charged spectra for which the charge state can be determined with high confidence. Charge states are assigned using the following protocol: spectra are assigned a specific charge state ($+2$ or $+3$) if the difference between the highest $Xcorr$ values for one charge state is more than 0.01 greater than the other charge state, and the highest $Xcorr$ score was itself greater than 0.4. Spectra that have ambiguous or low $Xcorr$ values (that is, a maximum $Xcorr$ less than 0.4 or a difference less than 0.01) are ignored. These two thresholds are set to obtain the highest charge-filter curve area (CFCA see Subsection 4.3) for the H-LTQ-1 and S-LCQ data sets.

3.3 SVM training

We train SVM classifiers using the publicly available PyML software (`pyml.sourceforge.net`) to find the maximum-margin hyperplane between our $+3$ and $+2$ training examples in several train-test data set pairs. Each data set is summarized in a matrix with n rows and 34 columns, where n is the number of spectra in the data set, and each row summarizes an individual spectrum with the 34 features described in Section 2. Prior to training, the matrix for each training set is normalized by subtracting the column mean from each entry and then dividing each entry by the column's standard deviation. For consistency, the mean and standard deviation of the training set are used to standardize the test data set in each case.

We experimented with two standard classes of kernel functions, polynomial kernels and radial-basis kernels (4). A polynomial kernel of degree d is defined as $K(X,Y) = (X \cdot Y + 1)^d$. A higher-degree polynomial provides more flexibility in the SVM decision boundary by including separate features for all d-way correlations among the original features. The radial-basis kernel of width σ is $K(X,Y) = exp(\|X - Y\|)/2\sigma^2)$.

Our SVM algorithm has two parameters that are not learned from the data: the soft-margin penalty C, and the kernel parameter (d or σ). We set these parameters using leave-one-out cross-validation within the training set. For the Gaussian kernel, we varied the width parameter over the values (0.01, 0.1, 1, 10); for the polynomial kernel, we varied the degree of the polynomial over the values (1, 2, 3, 4, 5, 6). In both cases, we varied C over the values (0.01, 0.1, 1, 10, 100). For each pair of parameters, we perform leave-one-out cross-validation, and then select the pair of parameters that yields the best performance, as measured by the receiver operating characteristic (ROC) curve area (described in Section 4.1 and in (10)).

3.4 Estimation of True and False Positive Rates

We seek to evaluate our SVM classifiers based on the number of true-positive peptide and protein hits that each maintains after classification. This requires an estimate of the rate of false positive assignment of peptides to spectra by SEQUEST, which we obtain with the following procedure. Spectra are searched against a customized database consisting of the 2004-May-02 *E. coli* RefSeq protein database concatenated with a database of common contaminant proteins and randomized sequences with the same length and amino acid distributions as the original RefSeq database. The number of protein hits to the contaminant and randomized databases is used as a numerical estimate of the number of false positive protein hits to the RefSeq database (similar to the reversed database false positive estimate in (11)). The number of true positive protein hits is then determined by subtracting twice this estimated number of false positive hits from the number of hits to the RefSeq database. A similar procedure is used to estimate the number of true positive and false positive peptide matches.

4 Results

The trained SVM classifies multiply-charged spectra as $+2$ or $+3$ with a high degree of accuracy. By classifying 80% of multiply-charged spectra with an SVM trained on a hybrid data set, we attain a decrease in total spectrum search time of 40% while maintaining an average of 99% of protein IDs and 99% of peptide IDs. This level of performance is robust across mass spectrometry platforms (ESI-CID ion-trap mass spectrometers: ThermoFinnigan LCQ and LTQ) and for assorted mammalian and *E. coli* proteins.

4.1 Discriminative power of each feature

We measure the ability of each of the 34 features described in Section 2 to discriminate between charge states prior to performing any SVM analysis. Most features demonstrate some ability to discriminate positive from negative training examples, but none sufficiently well to be used alone.

We quantify discriminative ability using the area under receiver operating characteristic (ROC) curves. An ROC curve relates the number of false positive classifications versus the number of true positive classifications for a varying classification threshold. A perfect classifier will classify all true positives above all false positives, yielding an ROC area of one. A random classifier will lack any ability to distinguish between true and false positives, yielding an ROC area of approximately 0.5. By calculating the area under the ROC curve for each feature, we measured the ability of that feature to distinguish between $+3$ and $+2$ charged spectra.

Table 2. ROC areas for 19 of 34 features measuring discrimination between $+2$ and $+3$ training examples across the four Phase I data sets. Rows correspond to the features listed in Section 2 and columns correspond to the data sets listed in Table 1. ROC areas of less than 0.5 indicate that the feature is higher for $+2$ than $+3$ training examples. Dashes indicate missing data. The ROC areas for the remaining 15 of the 34 features are shown graphically in Figure 4.

ID	E-LTQ-1	E-LCQ-1	S-LCQ	Hybrid
X_{+3}	0.764	0.324	0.926	0.637
X_{+2}	0.220	0.017	0.148	0.156
$X_{+3/+2}$	0.913	0.400	0.971	0.750
X_{+3}^{CO}	0.562	0.152	0.697	0.463
X_{+2}^{CO}	0.416	0.166	0.543	0.426
$X_{+3/+2}^{CO}$	0.534	0.199	0.629	0.437
$X_{+3}^{H_2O}$	0.739	0.263	0.816	0.580
$X_{+2}^{H_2O}$	0.351	0.058	0.288	0.255
$X_{+3/+2}^{H_2O}$	0.874	0.368	0.896	0.693
$X_{+3}^{NH_3}$	0.690	0.253	0.770	0.541
$X_{+2}^{NH_3}$	0.422	0.136	0.411	0.339
$X_{+3/+2}^{NH_3}$	0.781	0.281	0.798	0.606
B_{+3}	0.678	0.829	0.583	0.699
B_{+2}	0.678	0.829	0.577	0.697
$H_{m/z}$	0.531	0.500	0.500	0.502
P_{+3}	0.475	0.293	—	0.459
P_{+2}	0.433	0.634	—	0.470
$P_{+3/+2}$	0.466	0.335	—	0.463
$X_{+1,+2}$	0.714	0.228	0.719	0.559
I_i	See Figure 4			

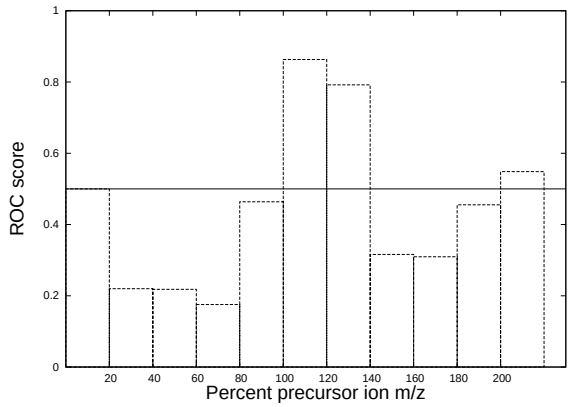


Figure 4. ROC areas for spectrum bin features. The figure plots the ROC area associated with each of the intensity histogram features 20-34 from Section 2. To compute these features, the axes of each spectrum are first rescaled so that the precursor m/z is 1 and the peaks sum to 1. Peaks are then binned along the m/z axis according to their percent of the precursor m/z, with each bin containing the fraction of total spectrum peak intensity within that bin. The figure plots ROC area calculated with $+3$ charge in the positive class. Thus, the fraction of total spectrum intensity in bin 100%-120% is relatively higher in $+3$ ions in the E-LCQ-3 data set, reflected in an ROC area greater than 0.5. The trends in other data sets are largely qualitatively similar.

ROC areas for the first 19 features are listed in Table 2 and shown graphically for the remaining 15 features in Figure 4. For each of the four training data sets, the most discriminative feature is in the (X_2, X_3, $X_{+3/+2}$) triplet, which measures matching ions that sum to the predicted precursor ion mass. Briefly, the best feature for the E-LCQ-1 and Hybrid data sets is X_{+2}; the best feature for the E-LTQ-1 and S-LCQ is $X_{+3/+2}$. The ROC areas for other features are broadly consistent between data sets; the main exception is the E-LCQ-1 data set, in which many features designed to increase for positive ($+3$) training examples instead show ROC areas less than 0.500. This is an indication of the difficulty of finding a classifier that works across several platforms.

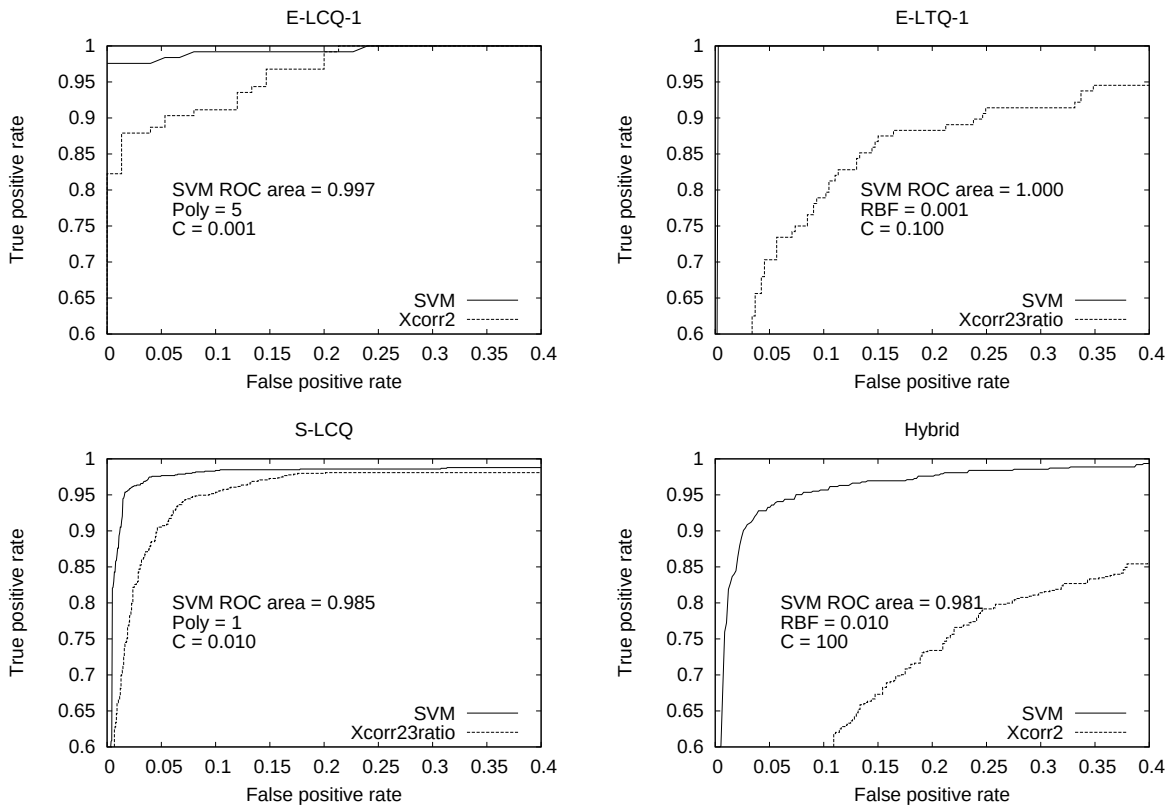


Figure 5. The SVM discriminates better than any single feature. Each plot shows an ROC curve for an SVM trained using leave-one-out cross-validation, as well as the corresponding ROC curve for the best-performing single feature in the given data set. The SVM parameters and ROC area are listed above each plot. The best feature the E-LCQ-1 and hybrid data sets is X_{+2}; the best feature for the E-LTQ-1 and S-LCQ data sets is $X_{+3/+2}$.

4.2 SVM training

We trained an SVM on each of the "phase I" data sets listed in Table 1, as well as on the hybrid (I+II) data set. During the selection of SVM parameters, each model's discriminative power is evaluated using leave-one-out cross-validation, as described in Section 3.3. The resulting ROC curves are shown in Figure 5. No single combination of SVM parameters performed optimally across all data sets; however, most ROC areas are close (within 0.010) to the best ROC area, indicating that the kernel and soft-margin parameters do not strongly affect the SVM's classification ability. Figure 5 also shows ROC curves for the single best-performing feature in each data set. In every case, the SVM successfully combines information from multiple features into a single classifier that out-performs the best-performing single feature. In the case of the H-LTQ-1 data, the SVM classifies the data perfectly, with an ROC area of 1.000.

4.3 Evaluating the SVM classifier

Ideally, a charge-state classifier would work generally across a broad range of platforms and species. Although both the LCQ and the LTQ use frequency-based activation to acquire a tandem mass spectrum, the data is much more rich in the LTQ because of the significantly increased ion capacity. Thus, any precursor-ion classifier should have the ability to handle data from either of these two instrument types. We seek to assess the ability of each SVM to classify two kinds of test examples: spectra generated from similar sources—such as training on the E-LCQ-1 data set and testing on E-LCQ-2—and spectra generated from dissimilar sources—such as training on the E-LTQ-1 and testing on a S-LCQ.

To evaluate the utility of our charge-state classifiers, we define a new figure of merit called the "charge-filter curve area" (CFCA). The goal of the charge-state classifier is to

save compute time by reducing the number of database searches that must be performed. If a spectrum's charge state is identified correctly, then the classifier will save time; if the spectrum is classified incorrectly, however, then there is a chance that we will miss a peptide ID (but only a chance, since some spectra have no ID). Thus, the CFCA measures the trade-off between time saved by classifying spectra versus the number of lost peptide or protein identifications. The CFCA is analogous to the ROC area, but is defined with respect to a curve that plots the number of peptides (or proteins) identified as a function of the number of database searches performed (Figure 6). Similar to an ROC area, a perfect charge-state filter would receive a score of 1, whereas a completely random classifier would receive a score of 0.5.

In calculating the CFCA we assume that each spectrum takes a roughly equal amount of time to search, and that the classification of each spectrum is negligible when compared to search time. In addition, search time and peptide IDs are calculated from multiply-charged spectra only; charge $+1$ spectra are excluded entirely from our CFCA calculations. In our data sets, anywhere between 2% and 65% of the spectra are singly charged.

Using the CFCA, we first evaluate the performance of the three SVMs trained from a single type of data. These results are summarized in Table 3. All three classifiers receive high CFCA scores when classifying examples that come from a data set generated on the same mass spectrometer and from the same organism. However, the performance generally deteriorates when classifying examples from a different data source. The classifier trained on the E-LTQ-1 demonstrates the most significant lack of portability: it performs slightly worse than random when tested on the E-LCQ data sets. In contrast, the classifier trained on the S-LCQ data set demonstrates the most robustness in classifying examples from data sets generated on different platforms. Furthermore, the S-LCQ classifier achieves the highest CFCA of all trained classifiers for all comparable test data sets.

An additional benefit of the charge-state classifier is its ability to eliminate some false positive peptide identifications. Occasionally, searching using a spectrum with an incorrect charge-state assignment nonetheless yields a peptide ID. These false-positive identifications can be eliminated by avoiding searching using the wrong charge-state assignment. In all data sets analyzed here, the estimated false positive rate after charge-state filtration either remains constant or is reduced substantially, sometimes by as much as 60%.

The classifier trained on the hybrid data set shows the most robustness across multiple platforms and organisms. Figure 6 shows charge-filter curves for two independent test sets, and Table 4 summarizes the results for these two and

Table 3. Charge-filter curve area (CFCA) for 12 training and testing data set pairs. The CFCA (Figure 6) measures the ability of the SVM classifier to classify spectra (and thus eliminate possible charge states) while still maintaining peptide or protein IDs. In analogy to an ROC area, a CFCA of 1 indicates a perfect classifier and a CFCA of 0.5 indicates a random classifier. Here we trained an SVM classifier on each of three non-hybrid training data sets and tested on four other data sets.

Testing data	Training data		
	E-LCQ-1	S-LCQ	E-LTQ-1
E-LCQ-1	–	0.992	0.444
E-LCQ-2	0.941	0.986	0.547
S-LCQ	0.907	–	0.949
E-LTQ-1	0.924	0.981	–
E-LTQ-2	0.950	0.978	0.948

their two replicate data sets numerically. The hybrid SVM achieves a CFCA greater than 0.970 on all data sets. Also, this classifier identifies more than 99% of peptide IDs and 99% of protein IDs when 40% of the multiply-charged spectra are eliminated.

To be useful in practice, the charge-state classifier needs to assign an interpretable confidence value to each of its predictions. The SVM algorithm assigns to each charge-state classification a discriminant score; this discriminant score can be converted into a probability using a straightforward sigmoid-curve fitting procedure (12). In the Supplementary Data, we demonstrate that this conversion is possible for our hybrid classifier. The resulting classifier can thus be thresholded only to predict charge state when the confidence is, for example, 99% or better.

5 Discussion

Using a trained SVM, we successfully determine the charge state of a substantial portion of mass spectra while maintaining virtually all true-positive protein and peptide IDs. At one threshold, we reduce estimated search time for multiply-charged spectra by 40% while maintaining an average of 99% of all peptide and protein IDs. This filtration step thus almost halves search time for multiply-charged spectra with minimal loss of identifications. In addition, the SVM trained on the hybrid data set shows comparable results across multiple platforms. As an added benefit, charge-state filtration reduces the rate of false positive identification in all analyses, in some cases by over 60%.

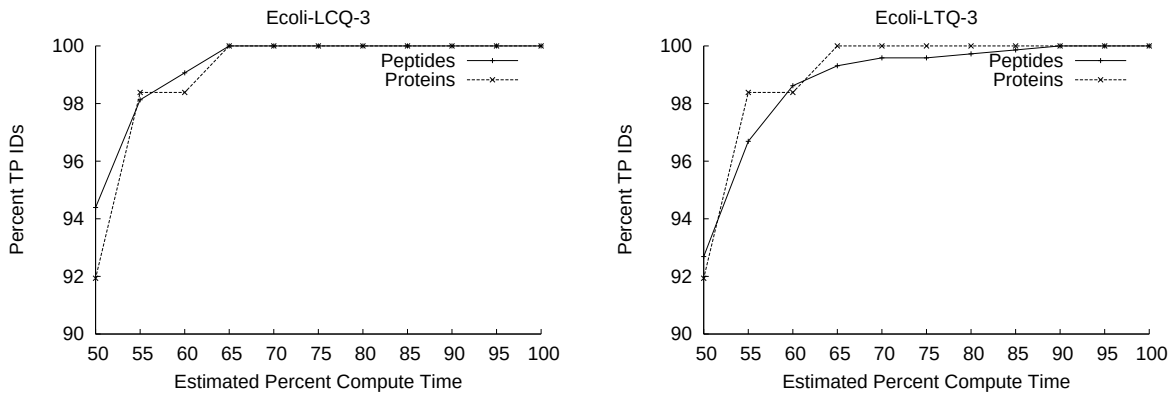


Figure 6. Charge-filter curves for the SVM classifier trained on the hybrid data set. The figures plot percent of peptides and proteins identified versus the estimated search time. Each plot corresponds to a different, independent test set. The areas under these curves are given in Table 4.

Table 4. Evaluation of an SVM trained on a hybrid data set. We trained an SVM classifier on a hybrid data set combining data from Phase I and II data sets and tested this classifier on the Phase III data sets. In Column 1, we measured this classifier's performance using the charge-filter curve area (CFCA), which measures the trade-off between spectra classified and peptide IDs maintained. The two remaining columns show the percent of true positive peptide and true positive protein IDs maintained when the SVM eliminates 40% (classifies 80%) of all multiply-charged spectra.

Testing Data	CFCA	% TP Peptide	% TP Protein
E-LCQ-3	0.984	99.1	98.4
E-LCQ-4	0.976	99.3	101.3
E-LTQ-3	0.972	98.6	99.5
E-LTQ-4	0.976	99.0	100.0

Our set of features could easily be extended to additional charge-state classes, such as $+4$ or greater. It appears, however, that these high charge-state ions are not common in many data sets—approximately 2–3% of all ions in (3)—and thus classifying them would offer only marginal search time improvement.

We use SEQUEST as the final arbiter of high-quality charge-state assignments in our training data; thus, any systematic inaccuracies in SEQUEST charge-state assignment will be replicated by our SVM classifier. A possible improvement to the algorithm would entail selecting training examples from different algorithm sources; however, given the strictness of our thresholds, it is likely that our training data errs on the side of caution, excluding true identifications rather than including false ones.

We compared our final hybrid SVM classifier with two other published methods for charge-state determination. Our method shows considerable improvement over the 2to3 algorithm presented by (13). When classifying spectra in the E-LTQ-3 data set, 2to3 classifies only 10% of all spectra (for 5% time savings) while maintaining 100% of both peptide and protein IDs. The performance of 2to3 on the E-LCQ-3 data set is difficult to compare with our method, because 2to3 eliminates some low-quality spectra entirely. Regardless, classification and elimination of spectra with 2to3 on the E-LCQ-3 data set yields a 52% reduction in search time while maintaining 95.3% of peptide and 98.4% of protein IDs from multiply-charged spectra. If one only considers those spectra that 2to3 does not eliminate entirely (that is, only the spectra that it classifies as $+2$, $+3$ or ambiguous, rather than eliminating entirely, in a matter consistent with our method) then it offers a 22% reduction in search time while again maintaining the same percentages of IDs as before.

We did not compare our method with that described in (3) directly, because the program was not made available for distribution. It is also unclear exactly how the authors calculate the reduction in search time reported in the paper. However, the authors report an estimated reduction in search time of 66% while maintaining 88.8% of peptide IDs when classifying spectra generated on the Bruker Esquire 3000 mass spectrometer. Their method has the advantage of detecting $+4$ charge peptides; these peptides are, as they

state, rare.

In sum, our charge-filter tool offers substantial savings in compute time with minimal cost in peptide and protein IDs. We envision the tool becoming an important part of the suite of mass spectrometry sequence database search tools.

References

[1] M. Bern, D. Goldberg, W. H. McDonald, and J. R. Yates, III. Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*, 20:I49–I54, 2004.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144–152. ACM Press, Pittsburgh, PA, 1992.

[3] J. Colinge, J. Magnin, T. Dessingy, M. Giron, and A. Masselot. Improved peptide charge state assignment. *Proteomics*, 3(8):1434–1440, 2003.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge UP, Cambridge, UK, 2000.

[5] V. Dancik, T. Addona, K. Clauser, J. Vath, and P. Pevzner. *De novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6(3-4):327–342, 1999.

[6] A. Keller, S. Purvine, A. I. Nezvizhskii, S. Stolyar, D. R. Goodlett, and E. Kolker. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS*, 6(2):207–212, 2002.

[7] M. Kinter and N. E. Sherman. *Protein sequencing and identification using tandem mass spectrometry*. Wiley-Interscience, 2000.

[8] M. J. MacCoss, C. C. Wu, and J. R. Yates, III. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Analytical Chemistry*, 74(21):5593–5599, 2002.

[9] A. L. McCormack, D. M. Schieltz, B. Goode, S. Yang, G. Barnes, D. Drubin, and J. R. Yates, III. Direct analysis and identification of proteins in mixtures by LC-MS/MS and database searching at the low-femtomole level. *Analytical Chemistry*, 69(4):767–776, 1997.

[10] C. E. Metz. Basic principles of ROC analysis. *Semin. Nucl. Med.*, 8:283–298, 1978.

[11] R. E. Moore, M. K. Young, and T. D. Lee. Qscore: An algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, 2002.

[12] J. C. Platt. Probabilities for support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[13] R. G. Sadygov, J. Eng, E. Durr, A. Saraf, H. McDonald, M. J. MacCoss, and J. R. Yates, III. Code developments to improve the efficiency of automated MS/MS spectra interpretation. *Journal of Proteome Research*, 1(3):211–215, 2002.

[14] D. L. Tabb, J. K. Eng, and J. R. Yates, III. *Proteome Research: Mass Spectrometry*, pages 125–142. Springer, New York, 2001.

[15] V. N. Vapnik. *Statistical Learning Theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York, 1998.

[16] J. R. Yates, III. Mass spectrometry and the age of the proteome. *Analytical Chemistry*, 33:1–19, 1998.