Charge state determination of peptide tandem mass spectra using support vector machine (SVM)

An-Min Zou, Jiarui Ding, Jin-Hong Shi, and Fang-Xiang Wu*

***Abstract*— A single mass spectrometry experiment could produce hundreds of thousands of tandem mass spectra. Several search engines have been developed to interpret tandem mass spectra. All search engines need to determine the masses of peptide ions from mass/charge ratios of ions. Unfortunately, mass spectrometers do not detect the charges of ions. A current strategy is to search candidate peptides multiply times, once for each possible charge state (typically $+2$ or $+3$). However, this strategy not only wastes the search time but also increases the risk of false positive peptide identification.**

This paper aims at discriminating doubly charged spectra from triply charged ones. 28 features are introduced to describe the discriminant characteristics of doubly charged and triply charged spectra. The support vector machine (SVM) technique is used to train the classifier on these 28 features. To verify the proposed method, computational experiments are conducted on two types of datasets: ISB dataset generated from the low-resolution ion-trap instrument and TOV dataset generated from the high-resolution quadrupole-time-of-flight (Q-TOF) instrument. For each type of dataset, the SVM-based classifiers are trained and tested on 20 randomly sampled sub-datasets. The results show that the proposed method reaches averagely 95% and 93% of correct rates to discriminate doubly charged spectra from triply charged ones for the low-resolution ISB dataset and the high-resolution TOV dataset, respectively.

I. Introduction

With the development of proteomics, tandem mass spectrometry (MS/MS) has been used for the rapid identification and characterization of protein components of complex biological mixtures. By using an enzyme, e.g. trypsin, proteins are digested into peptides. Tandem mass spectra are employed to analyze these peptides in view of their identifications by database search or *de novo* sequencing. Database search programs, such as SEQUEST [1] and MASCOT [2], identify peptides by comparing tandem mass spectra with theoretically predicted spectra derived from protein databases. *De novo* sequencing algorithms try to reconstruct original peptide sequences using tandem mass spectra. A review of several common *de novo* sequencing algorithms is given by [3].

Depending on the mass spectrometry ionization technology, peptide ions can carry more than one charge, especially the case with electrospray ionization (ESI) instruments [4]. In such a case, the instrument does not measure the masses of peptide ions but the mass/charge (m/z) ratios. However, most search algorithms need to calculate the mass of a peptide ion from the mass spectrum before comparing an experimental spectrum with a theoretical spectrum. Hence, if the charge state of a peptide ion is not known, search algorithms have to search candidate peptides multiple times, once for each possible charge state. However, this may increase both the search time and the risk of false positive peptide identification. It is, therefore, of great interest to assign reliable charge states to peptide ions.

The common technique used for the determination of the charge state is based on isotopic peaks of a peptide spectrum that are mostly caused by the ^{13}C isotope [5]. However, determination of charge states by isotopic peaks is possible only when high-resolution instruments such as quadrupole-time-of-flight (Q-TOF) are used. Many common mass spectrometers like the ion trap and triple quadrupole instruments generally have the limited resolution. Therefore, the isotopic peak based technique is not suitable to determine charge states of peptide tandem mass spectra produced by low-resolution instruments.

Several algorithms have been proposed to determine the charge state of a peptide ion from a mass spectrum. In general, singly charged spectra can be reliably distinguished from multiply charged spectra with currently existing algorithms [6] while it is rare to obtain high quality spectra with charges greater than three. Therefore, most existing methods aim at distinguishing doubly charged and triply charged spectra. In [7], a program named 2to3 was proposed to estimate the charge state of a precursor ion by counting the number of complementary fragment peak pairs. In [8], a support vector machine (SVM) based approach was proposed to determine charge states of multiply charged peptide tandem mass spectra. 34 spectral features were introduced to discriminate between doubly charged and triply charged peptide tandem mass spectra. Out of these 34 features, 19 features were derived from the *a priori* differences according to the charge state of the peptide mass spectrum, and other 15 features were the sum of intensities over each of 15 different m/z ranges divided by the total intensity of the spectrum. The multiply charged spectra were classified into three groups: "+2", "+3", and "+2OR+3" to minimize the losses of peptide identifications in [8]. More recently, Na *et al.* developed an SVM-based approach to differentiate doubly charged spectra

*Corresponding author.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

A.-M. Zou and J. Ding are with Department of Mechanical Engineering, University of Saskatchewan, 57 Campus Dr., Saskatoon, SK, S7N 59A, Canada {anz572, jid505}@mail.usask.ca

J.-H. Shi is with Division of Biomedical Engineering, University of Saskatchewan, 57 Campus Dr., Saskatoon, SK, S7N 59A, Canada jis958@mail.usask.ca

F.-X. Wu is with Department of Mechanical Engineering, and Division of Biomedical Engineering, University of Saskatchewan, 57 Campus Dr., Saskatoon, SK, S7N 59A, Canada fangxiang.wu@usask.ca

from triply charged ones [9]. They proposed 9 features to describe the discriminant characteristics of doubly charged and triply charged spectra. In these 9 features, one feature was the difference between the number of complementary peak pairs in doubly charged case and triply charged case, 5 features were calculated by the Good-Diff Fraction (GDFR) [10] for singly charged and doubly charged fragment ions at different m/z ranges, and other 3 features were the sum of intensities over each of 3 different m/z ranges divided by the total intensity of the spectrum. Similar to [8], the multiply charged spectra are classified into three groups.

Instead of classifying multiply charged spectra into three groups ("+2", "+3", and "+2OR+3") in the previous methods, in this paper we develop a new method which classifies multiply charged spectra into only two groups ("+2" and "+3") with high accuracy. To do this, 28 spectral features are proposed to distinguish between doubly charged and triply charged spectra. Out of the proposed 28 spectral features, 24 features are computed by combining the properties of peptide tandem mass spectra with peak intensities, and other 4 ones are the sum of relative intensities over each of 4 different m/z ranges in the spectrum. With the proposed 28 features, the SVM is applied to constructing the classifier. The proposed method can reduce the number of database searches for multiply charged spectra by 50%. Computational experimental results from the low-resolution ISB and high-resolution TOV datasets have shown the performance of the proposed method.

II. Methods and Materials

A. Properties of peptide tandem mass spectra

A peptide P is a sequence of n amino acids. $P = p_1 p_2 \cdots p_n$ in an alphabet of 20 amino acids, each amino acid having a mass $m(p_i)$. The mass of the peptide P is calculated by

$$m(P) = \sum_{i=1}^{n} m(p_i) + M_w \tag{1}$$

where M_w is the mass of a water molecule. Generally, in mass spectrometry experiments the cleavage along the peptide's backbone in the collision-induced dissociation (CID) stage results in an N-terminal fragment ion b_i and a C-terminal fragment ion y_{n-i}. The m/z-values of singly charged b_i and b_{i-1} ions are, respectively, computed by

$$m(b_i) = \sum_{j=1}^{i} m(p_j) + m(H) \tag{2}$$

$$m(b_{i-1}) = m(b_i) - m(p_i) \tag{3}$$

where $m(H)$ is the mass of a hydrogen atom, and the m/z-values of doubly charged b_i and b_{i-1} ions are, respectively, computed by

$$m(b_i^{2+}) = (m(b_i) + m(H))/2 \tag{4}$$

$$m(b_{i-1}^{2+}) = m(b_i^{2+}) - m(p_i)/2 \tag{5}$$

The m/z-values of singly charged y_{n-i} and y_{n-i+1} ions are respectively computed by

$$m(y_{n-i}) = \sum_{j=i+1}^{n} m(p_j) + m(H) + M_w \tag{6}$$

$$m(y_{n-i+1}) = m(y_{n-i}) + m(p_i) \tag{7}$$

and the m/z-values of doubly charged y_{n-i} and y_{n-i+1} ions are, respectively, computed by

$$m(y_{n-i}^{2+}) = (m(y_{n-i}) + m(H))/2 \tag{8}$$

$$m(y_{n-i+1}^{2+}) = m(y_{n-i}^{2+}) + m(p_i)/2 \tag{9}$$

From (1) through (9) the following identities can be obtained for the relations between b_i and y_{n-i} ions

$$m(P) + 2m(H) = m(b_i) + m(y_{n-i}) \tag{10}$$

$$m(P)/2 + 2m(H) = m(b_i^{2+}) + (m(y_{n-i}) + m(H))/2 \tag{11}$$

$$m(P)/2 + 2m(H) = (m(b_i) + m(H))/2 + m(y_{n-i}^{2+}) \tag{12}$$

$$m(P)/2 + 2m(H) = m(b_i^{2+}) + m(y_{n-i}^{2+}) \tag{13}$$

and for the relations between b_{i-1} and y_{n-i} ions

$$m(P) + 2m(H) - m(p_i) = m(b_{i-1}) + m(y_{n-i}) \tag{14}$$

$$m(P)/2 + 2m(H) - m(p_i)/2 = (m(y_{n-i}) + m(H))/2 + m(b_{i-1}^{2+}) \tag{15}$$

$$m(P)/2 + 2m(H) - m(p_i)/2 = (m(b_{i-1}) + m(H))/2 + m(y_{n-i}^{2+}) \tag{16}$$

$$m(P)/2 + 2m(H) - m(p_i)/2 = m(b_{i-1}^{2+}) + m(y_{n-i}^{2+}) \tag{17}$$

According to the CID fragmentation principle, a peptide mass spectrum may include several other types of fragment ions such as neutral loss (H_2O, NH_3), a, c, x, z, and so on. That is, there exist some pairs of m/z-values with difference of the mass of certain molecular (e.g., water, ammonia) in the peptide tandem mass spectra.

B. Spectral features

A mass spectrum usually contains tens to hundreds of m/z-values on the x-axis, each with corresponding signal intensity on the y-axis. In this study, after removing the noisy peaks by use of the morphological reconstruction method [11], 28 spectral features are introduced to distinguish between doubly charged and triply charged spectra as follows.

In the ESI process, a peptide ion can be multiply charged. This study is concerned with the doubly charged and triply charged peptide tandem mass spectra. If the precursor ion with $m/z = M_z$ is doubly charged, the m/z-values of the peaks in a spectrum S should be in the interval $(0, 2M_z]$ and the m/z-values of doubly charged fragment ions should be in the interval $(0, M_z]$. If the precursor ion is triply charged, the m/z-values of the peaks in the spectrum should be in the interval $(0, 3M_z]$ and the m/z-values of doubly charged fragment ions should be in the interval $(0, 3/2M_z]$. Based on this observation, the entire m/z range of a spectrum is divided into four bands, $(0, M_z], (M_z, 3/2M_z], (3/2M_z, 2M_z]$, and $(2M_z, 3M_z]$. The intensity density in each of these four ranges is expected to be different for doubly charged and triply charged spectra. Thus we propose the following four features for a given peptide mass spectrum S

$$F_1 = \sum_{m(x) \leq M_z} I_r(x) \tag{18}$$

$$F_2 = \sum_{M_z < m(x) \leq \frac{3}{2}M_z} I_r(x) \tag{19}$$

$$F_3 = \sum_{\frac{3}{2}M_z < m(x) \leq 2M_z} I_r(x) \tag{20}$$

$$F_4 = \sum_{2M_z < m(x) \leq 3M_z} I_r(x) \tag{21}$$

where $m(x)$ and $I_r(x)$ denote the m/z-value and the relative intensity of the peak x in the spectrum S, respectively. In the existing literature, the relative intensity is usually defined as the raw intensity normalized by the intensity of the most abundant peak or the sum of peak intensities in a spectrum. This study employs the raw intensity normalized by the sum of peak intensities in a spectrum as the relative intensity since better performance of the SVM classifier can be obtained.

A spectrum obtained from a precursor ion with $m/z = M_z$ can have different peptide's mass according to its charge state. For example, if the precursor ion is doubly charged, its mass can be calculated by $M_{p2} = 2M_z - 2m(H)$. If the precursor ion is triply charged, its mass can be calculated by $M_{p3} = 3M_z - 3m(H)$. Based on the properties of peptide tandem mass spectra and the different mass of the precursor ion according to its charge state, we introduce other 24 features as follows. To do this, we first define two variables for a given peptide mass spectrum S

$$sum_1(x,y) = m(x) + m(y) \tag{22}$$

$$sum_2(x,y) = m(x) + (m(y) + m(H))/2 \tag{23}$$

where $m(x)$ and $m(y)$ denote the m/z-values of peaks x and y in the spectrum S, respectively, and a weighting factor

$$W(x,y) = \frac{I_r(x) + I_r(y)}{2} \tag{24}$$

where $I_r(x)$ and $I_r(y)$ represent the relative intensities of peaks x and y in the spectrum S, respectively.

$F_5 - F_{10}$: Perfect complements. These features measure how likely an N-terminus ion and a C-terminus ion in the spectrum S are produced as the peptide fragments at the same peptide bond. The doubly charged features are defined as

$$F_5 = \sum\{W(x,y)|sum_1(x,y) \approx M_{p2} + 2m(H))\} \tag{25}$$

$$F_6 = \sum\{W(x,y)|sum_1(x,y) \approx M_{p2}/2 + 2m(H)\} \tag{26}$$

$$F_7 = \sum\{W(x,y)|sum_2(x,y) \approx M_{p2}/2 + 2m(H)\} \tag{27}$$

The feature F_5 measures the presence of complementary peak pairs of singly charged ions in the spectrum S; the feature F_6 measures the presence of complementary peak pairs of doubly charged ions in the spectrum S, and the feature F_7 measures the presence of complementary peak pairs of one doubly charged and the other singly charged ions in the spectrum S. These three features are expected to be greater for doubly charged spectra than for triply charged spectra. The comparison implied by ≈ employs a tolerance, which was set to 1 Thompson for the low-resolution ISB dataset and 0.1 Thompson for the high-resolution TOV dataset in this paper. These values for the tolerance were obtained by trial and error. The use of the weighting factors in the features in this paper is to account the increased likelihood of more intense peaks being true fragment ions.

Similarly, we can define three triply charged features F_8, F_9 and F_{10}. The formulas defining these three features are analogous to (25)-(27), except that the precursor ion mass term is replaced with M_{p3}. These three features are expected to be greater for triply charged spectra than for doubly charged spectra.

$F_{11} - F_{16}$: Complements with an amino acid difference. These features measure how likely a fragment ion and the complementary fragment ion of another ion in the spectrum S differ by one of the twenty amino acids. The doubly charged features are defined as

$$\begin{aligned} F_{11} = \sum\{W(x,y)|sum_1(x,y) \approx M_{p2} + 2m(H) - M_i, \\ i = 1, \cdots, 17\} \end{aligned} \tag{28}$$

$$\begin{aligned} F_{12} = \sum\{W(x,y)|sum_1(x,y) \approx M_{p2}/2 + 2m(H) - M_i/2, \\ i = 1, \cdots, 17\} \end{aligned} \tag{29}$$

$$\begin{aligned} F_{13} = \sum\{W(x,y)|sum_2(x,y) \approx M_{p2}/2 + 2m(H) - M_i/2, \\ i = 1, \cdots, 17\} \end{aligned} \tag{30}$$

where $M_i (i = 1, 2, \cdots, 17)$ are the 17 different masses of all 20 amino acids. This study considers all Methionine amino acids to be sulfoxidized and does not distinguish three pairs of amino acids in their masses: Isoleucine vs. Leucine, Glutamine vs. Lysine, and sulfoxidized Methionine vs. Phenylalanine since the masses of each pair are very close. The feature F_{11} measures the presence of peak pairs of one singly charged ion and the complementary fragment ion of the other singly charged ion corresponding to an amino acid mass difference in the spectrum S; the feature F_{12} measures the presence of peak pairs of one doubly charged ion and the complementary fragment ion of the other doubly charged ion corresponding to an amino acid mass difference in the spectrum S, and the feature F_{13} measures the presence of peak pairs of one doubly charged ion and the complementary fragment ion of the other singly charged ion corresponding to an amino acid mass difference in the spectrum S.

Similarly, we can define three triply charged features F_{14}, F_{15} and F_{16}. The formulas defining these three features are analogous to (28)-(30), except that the precursor ion mass term is replaced with M_{p3}.

$F_{17} - F_{22}$: Complements with a water or ammonia difference. These features measure how likely one ion in the spectrum S is produced by losing a water or ammonia molecule from the complementary ion of a b-ion or y-ion.

The doubly charged features are defined as

$$F_{17} = \sum\{W(x,y)|sum_1(x,y) \approx M_{p2} + 2m(H) - M_w \text{ or } M_{p2} + 2m(H) - M_a\} \quad (31)$$

$$F_{18} = \sum\{W(x,y)|sum_1(x,y) \approx M_{p2}/2 + 2m(H) - M_w/2 \text{ or } M_{p2}/2 + 2m(H) - M_a/2\} \quad (32)$$

$$F_{19} = \sum\{W(x,y)|sum_2(x,y) \approx M_{p2}/2 + 2m(H) - M_w/2 \text{ or } M_{p2}/2 + 2m(H) - M_a/2\} \quad (33)$$

where M_a is the mass of an ammonia molecule. The feature F_{17} measures the presence of peak pairs of one singly charged ion and the complementary fragment ion of the other singly charged ion with a difference of a water or ammonia molecule in the spectrum S; the feature F_{18} measures the presence of peak pairs of one doubly charged ion and the complementary fragment ion of the other doubly charged ion with a difference of a water or ammonia molecule in the spectrum S, and the feature F_{19} measures the presence of peak pairs of one doubly charged ion and the complementary fragment ion of the other singly charged ion with a difference of a water or ammonia molecule in the spectrum S.

Similarly, we can define three triply charged features F_{20}, F_{21} and F_{22}. The formulas defining these three features are analogous to (31)-(33), except that the precursor ion mass term is replaced with M_{p3}.

$F_{23} - F_{28}$: Complements with a CO or NH difference. These features measure how likely one ion in the spectrum S is a supportive ion of the complementary ion of a b-ion or y-ion. The doubly charged features are defined as

$$F_{23} = \sum\{W(x,y)|sum_1(x,y) \approx M_{p2} + 2m(H) - M_C \text{ or } M_{p2} + 2m(H) - M_N\} \quad (34)$$

$$F_{24} = \sum\{W(x,y)|sum_1(x,y) \approx M_{p2}/2 + 2m(H) - M_C/2 \text{ or } M_{p2}/2 + 2m(H) - M_N/2\} \quad (35)$$

$$F_{25} = \sum\{W(x,y)|sum_2(x,y) \approx M_{p2}/2 + 2m(H) - M_C/2 \text{ or } M_{p2}/2 + 2m(H) - M_N/2\} \quad (36)$$

where M_C and M_N are the masses of a CO group and an NH group, respectively. The feature F_{23} measures the presence of peak pairs of one singly charged ion and the complementary fragment ion of the other singly charged ion with a difference of a CO or NH group in the spectrum S; the feature F_{24} measures the presence of peak pairs of one doubly charged ion and the complementary fragment ion of the other doubly charged ion with a difference of a CO or NH group in the spectrum S, and the feature F_{25} measures the presence of peak pairs of one doubly charged ion and the complementary fragment ion of the other singly charged ion with a difference of a CO or NH group in the spectrum S.

Similarly, we can define three triply charged features F_{26}, F_{27} and F_{28}. The formulas defining these three features are analogous to (34)-(36), except that the precursor ion mass term is replaced with M_{p3}.

C. Normalization

After the feature extraction, each spectrum is mapped into a 28-dimensional feature vector, and then each value of these spectral features is normalized as $F_i/\max(F_i), i = 1, 2, \cdots, 28$, where $\max(F_i)$ is the maximum value of the ith spectral feature across the samples in the training set. This normalization is also applied to the spectral features in the test data, using the maximum value from the training data.

D. Classification method

In this paper, the SVM is applied to determine the charge states of multiply charged spectra because of its good generalization ability. The SVM is proposed by Vapnik based on the statistical learning theory [12]. An important characteristic of the SVM is that "while most classical neural network algorithms require an *ad hoc* choice of system's generalization ability, the SVM approach proposes a learning algorithm to control the generalization ability of the system automatically" [13]. In this study, the sequential minimal optimization (SMO) algorithm [14] is employed to train the SVM.

In this study, the SVM classifier is employed to classify multiply charged spectra into two groups in order to save the spectral searching time. To assess the performance of the SVM classifier, three correct rates are calculated in this study: correct rate for doubly charged spectra (CRD), correct rate for triply charged spectra (CRT), and accuracy

$$\text{CRD} = \frac{\text{\# of correctly classified doubly charged spectra}}{\text{\# of doubly charged spectra}}$$

$$\text{CRT} = \frac{\text{\# of correctly classified triply charged spectra}}{\text{\# of triply charged spectra}}$$

$$\text{accuracy} = \frac{\text{\# of correctly classified multiply charged spectra}}{\text{\# of multiply charged spectra}}$$

E. Experimental data

This study used two different proteome datasets: ISB and TOV. The first dataset was acquired on a low-resolution ion trap mass spectrometer, and the second one was generated by a high-resolution Q-TOF instrument.

1) ISB dataset: The ISB dataset used in this study was acquired on an LC-ESI ion trap (Thermo Finnigan LCQ instrument) and was provided by the Institute of Systems Biology (ISB, Seattle, USA). This dataset consists of 37044 peptide CID mass spectra. The samples analyzed were generated by the tryptic digestion of a control mixture of standard 18 proteins (not of human origin) [15]. The peptide and charge-state assignments were annotated for 125 singly charged spectra, 1242 doubly charged spectra and 573 triply charged spectra, using SEQUEST search program.

2) TOV dataset: This dataset consists of 83224 peptide tandem mass spectra which were acquired on a QSTAR Plusar (MDS Sciex Corp.) in Eastern Quebec Proteomic Center in Laval University Medical Research Center in Canada. The samples analyzed were generated by the tryptic digestion of a whole-cell lysate from the 36 fractions of TOV-112. These spectra were searched against a subset of the IPI database including 67971 human protein sequences using MASCOT and X! Tandem [16], respectively. The assignments of 2765 doubly charged and 3145

TABLE I

THE NUMBER OF THE SAMPLES IN THE TRAINING AND TEST SETS

SVM classify	Training set (D : T)	Test set (D : T)
for ISB dataset	300 : 300	942 : 273
for TOV dataset	400 : 400	2365 : 2745

TABLE II

THE RESULTS IN THE ISB TEST DATASET

	CRD (%)	CRT (%)	Accuracy (%)
Range	92.6∼95.3	92.7∼97.8	93.3∼95.6
Ave.	94.1	95.7	94.5
SD	0.71	1.36	0.62

triply charged spectra were verified to be correct by Scaffold (http://www.proteomesoftware.com/index.html) with the minimum identified probability of 0.95.

III. Results and Discussion

In this study, the SVM classifiers were respectively trained and tested on the identified multiply charged spectra in two different datasets: ISB and TOV. Doubly charged spectra were labeled as "$+1$", and triply charged spectra were labeled as "-1". This study employed radial basis functions (RBF) whose width parameter was set equal to 0.5 as the kernel functions of the SVMs. The penalty term for training set errors was set equal to 10. We selected these values for the parameters of SVM classifiers since better classification performance was obtained by using these values. The number of the samples in the training and test sets for the SVM classifiers is shown in Table I. 'D' represents the number of doubly charged spectra, and 'T' represents the number of triply charged spectra.

In this study we repeated to train and test each SVM classifier on 20 randomly sampled datasets to investigate the performance of the proposed method. The results are shown in Tables II and III. In these tables, 'Ave.' stands for the average, and 'SD' for the standard deviation. For the low-resolution ISB dataset, Table II shows that the proposed method can reach the accuracy of about 96% at the best case. This means that while classifying multiply charged spectra into two groups the proposed method can maintain about 96% of peptide identifications. On average it can maintain about 95% of peptide identifications for the ISB dataset. Though the essential interest of the proposed approach is for the low-resolution spectra, we tested our algorithm on the high-resolution TOV dataset. From Table III, it can be seen that the proposed method can classify multiply charged spectra into two groups while losing an average of about 7% of peptide identifications for the high-resolution TOV dataset. This indicates that the proposed method can be employed to determine charge states of peptide tandem mass spectra produced by both low-resolution and high-resolution instruments. Table IV shows the average correct rates over 20 randomly sampled datasets using the proposed method without normalization. Comparing Table IV with Tables II and III, it indicates that the normalization of spectral features can improve the performance of the SVM classifiers.

TABLE III

THE RESULTS IN THE TOV TEST DATASET

	CRD (%)	CRT (%)	Accuracy (%)
Range	90.1∼94.4	91.3∼94.9	92.1∼93.1
Ave.	92.1	92.9	92.5
SD	1.23	1.17	0.27

TABLE IV

THE RESULTS USING THE PROPOSED METHOD WITHOUT NORMALIZATION

Dataset	CRD (%)	CRT (%)	Accuracy (%)
ISB	93.6	93.8	93.7
TOV	91.2	92.9	92.1

In addition, the last row in each of Tables II and III gives the standard deviations of our proposed methods over twenty randomly sampled datasets. All the standard deviations for CRD, CRT and accuracy are very small (from 0.27%-1.36%). This indicates that the proposed method is insensitive to the variations of the training and test sets in the same dataset.

In [8] and [9], the authors classified multiply charged spectra into three groups: "+2", "+3", and "+2OR+3" in order to minimize the losses of peptide identifications, and about 40% reduction in the search time was obtained. However, this study classifies multiply charged spectra into only two groups: "+2" and "+3". Thus the proposed method can reduce the number of database searches for multiply charged spectra by 50%, and data reduction improvement of the proposed approach is about 10% compared with [8] and [9]. In addition, the methods in [8] and [9] have not been evaluated on high-resolution datasets.

Although the percentage of misclassified identified spectra by the proposed method is higher than those in [8] and [9], the multiply charged spectra are intentionally classified into three groups by their methods. For the low-resolution ISB dataset, the method proposed in [9] retains 93.1% of peptide identifications while classifying multiply charged spectra into two groups. This indicates that the proposed approach outperforms this existing algorithm while classifying the multiply charged spectra into two groups. One major reason for this is that the proposed spectral features combining the properties of peptide tandem mass spectra with peak intensities have good discriminant characteristics of doubly charged and triply charged spectra. In addition, the misclassified identified spectra may be false positive identifications since the search results by one search engine may also have false positives. In this study, to further illustrate the bias raised from the search results by SEQUEST, we randomly selected 30 misclassified spectra from the ISB dataset, which were classified as doubly charged or triply charged spectra by the proposed method, yet were determined by the SEQUEST search program as triply charged or doubly charged spectra. These 30 spectra were re-searched by on-line MASCOT against the SwissProt database (http://www.matrixscience.com/). The parent mass tolerance was set at ±2 Da, and the fragment ion mass tolerance was set at ±0.6 Da. The enzyme parameter was set as

TABLE V

Peptide-spectrum match scores by SEQUEST and MASCOT. [a]The precursor ion charge state was determined by the SEQUEST search result; [b]Boldface letter represents that the MASCOT ion score is over the significantly identified cutoff (p value<0.05)

Spectrum	Precursor ion charge state[a]	SEQUEST Xcorr score	MASCOT ion score
1	+3	4.1491	2
2	+3	3.7519	5
3	+2	3.3490	**49**[b]
4	+2	3.0861	2
5	+2	3.0358	27
6	+2	3.0143	13
7	+2	2.7514	28
8	+2	2.5227	23
9	+2	3.0628	32
10	+2	2.9468	17
11	+2	2.8124	9
12	+3	3.6251	7
13	+2	3.4835	8
14	+2	4.7507	**53**
15	+3	4.6131	20
16	+3	3.7381	5
17	+2	3.4835	8
18	+3	4.6891	**42**
19	+2	2.9700	33
20	+2	2.5432	13
21	+3	3.5476	15
22	+3	4.7086	38
23	+2	3.7191	7
24	+3	3.5661	6
25	+2	2.6438	9
26	+3	3.6299	16
27	+2	3.4664	26
28	+2	2.8471	14
29	+2	2.8032	11
30	+2	3.4664	26

tryptic sequences, and the maximum of missed cleavage site was 1. All 30 peptide-spectrum match scores by SEQUEST and MASCOT are shown in Table V. We found that only 3 spectra were significantly identified (p value<0.05) by MASCOT. This indicates that the other 27 spectra out of these 30 misclassified spectra may be false positive identifications. That is, 90% of the misclassified spectra may be false positive peptide identifications. Thus, the correct rates in this study should be higher, and manual verification by a mass spectrometry expert is also required to confirm this indication.

IV. Conclusions

In this paper, an SVM-based approach is proposed to determine charge states of multiply charged peptide tandem mass spectra. 28 spectral features are introduced to distinguish between doubly charged and triply charged spectra. Each spectrum is mapped into a 28-dimensional feature vector. The SVM is applied to construct the classifier in the feature space that distinguishes between doubly charged and triply charged spectra. The SVM classifiers are trained and tested on the low-resolution ISB and high-resolution TOV datasets, respectively. Computational experimental results have demonstrated the effectiveness of the proposed method.

The significance of the proposed method is two-fold. First, the proposed method provides a reliable algorithm to determine charge states of peptide tandem mass spectra before database search. The proposed method can reduce 50% of the spectral searching time for multiply charged spectra and reduce the risk of false positive peptide identification. Second, the proposed method can be employed to evaluate the database search results from one search engine while incorporating with different identification methods. For example, by both re-searching misclassified spectra with MASCOT and manual verification, we can confirm that the assignments of some of these spectra by SEQUEST are actually false.

Acknowledgements

The authors would like to thank Dr. Andrew Keller from Institute for Systems Biology for generously providing spectral data and protein databases for the ISB dataset and Dr. Guy G. Poirier from Laval University for providing the TOV dataset.

References

[1] J.K. Eng, A.L. McCormack, and J.R.III. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, *American Society for Mass Spectrometry*, vol. 5, 1994, pp. 976-989.

[2] D.N. Perkins, D.J.C. Pappin, D.M. Creasy, and J.S. Cottrell, Probability-based protein identification by searching sequence database using mass spectrometry data, *Electrophoresis*, vol. 20, 1999, pp. 3551-3567.

[3] B. Lu and T. Chen, Algorithms for de novo peptide sequencing using tandem mass spectrometry, *Drug Discovery Today: BioSilico*, vol. 2, 2004, pp. 85-90.

[4] P. James, ed., *Proteome Research: Mass Spectrometry*, Springer, Berlin, 2001.

[5] H. Steen and M. Mann, The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews. Molecular Cell Biology*, vol. 5, 2004, pp. 699-711.

[6] D.L. Tabb, J.K. Eng, and J.R.III. Yates, Protein idebtification by SEQUEST, in P. James, (ed.), *Proteome Research: Mass Spectrometry*, Springer, Berlin, 2001.

[7] R.G. Sadygov, et al., Code developments to improve the efficiency of automated MS/MS spectra interpretation. *Journal of Proteome Research*, vol. 1, 2002, pp. 211-215.

[8] A.A. Klammer, C.C. Wu, M.J. MacCoss, and W.S. Noble, Peptide charge state determination for low-resolution tandem mass spectra, *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05)*, 2005, pp. 175-185.

[9] S. Na, E. Paek, and C. Lee, CIFTER: Automated charge-state determination for peptide tandem mass spectra, *Anal. Chem.*, vol. 80, 2008, pp. 1520-1528.

[10] M. Bern, D. Goldberg, W.H. McDonald, and J.R.III. Yates, Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*, vol. 20, 2004, pp. i49-i54.

[11] L. Vincent, Morphological grayscale reconstruction in image analysis: application and efficient algorithm, *IEEE Transaction on Image Processing*, vol. 2, 1993, pp. 176-201.

[12] V. Vapnik, *Statistical Learning Theory*, John Willey & Sons, 1998.

[13] M. Pontil and A. Verri, *Properties of support vector machines*, Artificial Intelligence Laboratory, C.B.C.L., MIT Press, 1997.

[14] J. Platt, Fast training of support vector machines using sequential minimal optimization, In *Advances in Kernel Methods - Support Vector Learning*, MIT Press, 1999, pp. 42-65.

[15] A. Keller, et al., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, *Analytical Chemistry*, vol. 74, 2002, pp. 5383-5392.

[16] R. Craig and R.C. Beavis, A method for reducing the time required to match protein sequences with tandem mass spectra, *Rapid Communications in Mass Spectrometry*, vol. 17, 2003, pp. 2310-2316.