Clin Proteom (2009) 5:15–22
DOI 10.1007/s12014-009-9025-4

Quality Assessment of Tandem Mass Spectra by Using a Weighted K-Means

Jiarui Ding · Jinhong Shi · Fang-Xiang Wu

Published online: 12 March 2009

Abstract
Introduction The tandem mass spectrometer is a powerful tool with which to generate peptide (tandem) mass spectrum data for the analysis of complex biological protein mixtures in genomic-related disease cell lines. However, the majority of experimental tandem mass spectra cannot be interpreted by any database search engines. One of the main reasons this happens is that majority of experimental spectra are of quality too poor to be interpretable. Interpreting these "un-interpretable" spectra is a waste of time. Therefore, it is worthwhile to determine the quality of mass spectra before any interpretation.
Objectives This paper proposes an approach to classifying tandem spectra into two groups: one with high quality and one with poor quality.
Methods The proposed approach has two steps. First, each spectrum is mapped to a feature vector which describes the quality of the spectrum. Then, a weighted K-means clustering method is applied in order to classify the tandem mass spectra.
Results and Conclusion Computational experiments illustrate that one cluster contains the majority of the high-quality spectra, while the other contains the majority of the poor-quality spectra. This result indicates that if we just search the spectra in the high-quality cluster, we can save the time for searching the majority of poor-quality spectra while losing a minimal amount of high-quality spectra. The software created for this work is available upon request.

Keywords Tandem mass spectrum · Quality assessment · Weighted k-means · Feature vector · Peptide · Protein

J. Ding · F.-X. Wu (✉)
Department of Mechanical Engineering,
University of Saskatchewan,
Saskatoon, SK S7N 5A9, Canada
e-mail: fangxiang.wu@usask.ca

J. Shi · F.-X. Wu
Division of Biomedical Engineering, University of Saskatchewan,
Saskatoon, SK S7N 5A9, Canada

Introduction

One of the most important goals in early detection of genomic-related disease such as cancer or obesity is to identify and characterize the proteins and protein complexes present in related cell lines. High-performance liquid chromatography (HPLC) coupled with a tandem mass spectrometer provides an automated, high-throughput approach widely used to generate peptide (tandem) mass spectral data for the analysis of complex biological protein mixtures [1]. Most frequently, peptide identifications are made by comparing tandem mass spectra with a sequence database to find the significantly matching peptides in the database. Through the assignment of peptides to spectra, the original proteins present in the sample are inferred. Over the past decade, many automated database search engines have been developed for assigning peptides to tandem mass spectra, for example, SEQUEST [2], Mascot [3], and Sonar [4]. These search engines, as well as de novo sequencing methods [5, 6], have been successfully applied to peptide mass spectrum assignments in many proteomics projects. However, the majority of tandem mass spectra cannot be interpreted by these and other automatic methods, even after filtering poor-quality spectra using some simple filters such as "most intensive peak selection" criterion [2–4]. There are several reasons that the automatic methods fail to interpret the mass spectra. However, one of the main reasons is that these spectra are of quality too poor to be

interpretable. In general, a tandem mass spectrum is considered to be of high quality if it is produced from peptides; otherwise, it is considered to be of poor quality. Hence, it is worthwhile to develop an automatic quality assessment algorithm to discriminate high-quality from poor-quality spectra before interpretation by any method.

In the past, several supervised machine learning algorithms have been proposed to assess the quality of tandem mass spectra, which means a labeled training dataset is needed to train a classifier, and the trained classifier is used to classify spectra as high-quality or poor-quality [7–11]. Ideally, the training set should be identified by some peptide identification algorithms and manually validated, i.e., the set should be correctly labeled without or with very few falsely labeled spectra. However, these sets are hard to obtain in most cases. Worse still, tandem mass spectrometers may produce different spectra even for the same peptide under different experimental conditions. Thus, the performance of classifiers can be improved by training a classifier for each experiment. Clustering algorithms, which do not need a labeled training set, may be alterative choices for the quality assessment of tandem mass spectra.

In this paper, we propose a clustering algorithm-weighted k-means (WKM) method to classify the experimental spectra into two clusters, one with high-quality and the other with poor-quality spectra without using any prior information about the spectra dataset from search engines. The remainder of the paper is organized as follows. The "Feature Extraction" section studies the properties of theoretical spectra and introduces a means of mapping a spectrum to a feature vector. The "WKM" section introduces the WKM method. In the "Experiments and Results" section, one dataset is used to investigate the performance of the proposed method. The "Conclusions and Future Work" section concludes this study.

Feature Extraction

This subsection describes a means of mapping a tandem mass spectrum to a feature vector which describes the quality of the spectrum. To do this, the properties of theoretical spectra are discussed first.

Properties of Peptide Theoretical Spectra

Many algorithms such as SEQUEST [2], Mascot [3], and Sonar [4] have been used to assign experimental MS/MS spectra to peptides in a protein/peptide database. A key component of these algorithms is the score function, which evaluates the similarity between each experimental MS/MS spectrum and the predicted (theoretical) spectrum of a given peptide in the database. A peptide whose theoretical spectrum has the maximum similarity to the experimental spectrum is a likely candidate for the solution of the peptide identification problem. An experimental peptide mass spectrum is often expressed by a peak list, i.e., $S = \{(x_i, h_i) | 1 \le i \le m\}$, where ($x_i$, h_i) denotes the fragment ion i with m/z value x_i and intensity h_i. Since ion intensities are the results of many unknown factors and are yet difficult to utilize for spectral quality assessment, this study does not take into account intensity values of ions after the original spectra are pre-preprocessed by filtering out the noise peaks. Therefore, the peptide mass spectra in this study are reduced into a set of m/z values and are denoted by S_E.

On the other hand, the perfect MS/MS spectrum of a peptide is the theoretical spectrum. In practice, no mass spectrometers can produce perfect MS/MS spectra. However, investigating the peptide theoretical spectrum is extremely helpful for understanding the high-quality spectra which could potentially be assigned to a peptide. Let P be a peptide consisting of n amino acids a_1, a_2, ..., a_n with respective mass $m(a_i)$. The mass of peptide P is calculated by

$$m(P) = m(H) + m(OH) + \sum_{i=1}^{n} m(a_i) \quad (1)$$

where $m(H)$ and $m(OH)$ are the additional masses of the peptide's N- and C-terminals. Hereafter, we will use $m(X)$ to express the mass of a molecule or a group of atoms X.

In a tandem mass spectrometry experiment, a protein is fragmented into a series of peptide ions (sometime also called precursor ions or parent ions) at the first stage. For ion trap spectrometers, the produced precursor ions are mostly doubly or triply charged [12]. In the second stage, a series of selected precursor ions are fragmented further into fragment ions. For a doubly charged precursor ion, most of its fragment ions are singly charged, whereas a triply charged precursor ion, is likely to fragment at backbone bonds to form a series of singly charged and doubly charged fragment ions. Therefore, in this study, we consider both doubly charged and triply charged precursor ions, but only singly and doubly charged fragment ions.

As peptide P fragments at backbone bond between the i-th and $i+1$-th amino acids counting from the N-terminal, several types of ions could be produced as shown in the Fig. 1. The singly charged ion with N-terminal is denoted by b_i^+, and its m/z value is computed by

$$m\left(b_i^+\right) = m(H) + \sum_{j=1}^{i} m\left(a_j\right) \quad (2)$$

The doubly charged ion with N-terminal is denoted by b_i^{++}, and its m/z value is computed by

$$m\left(b_i^{++}\right) = \left[m\left(b_i^+\right) + m(H)\right]/2 \quad (3)$$

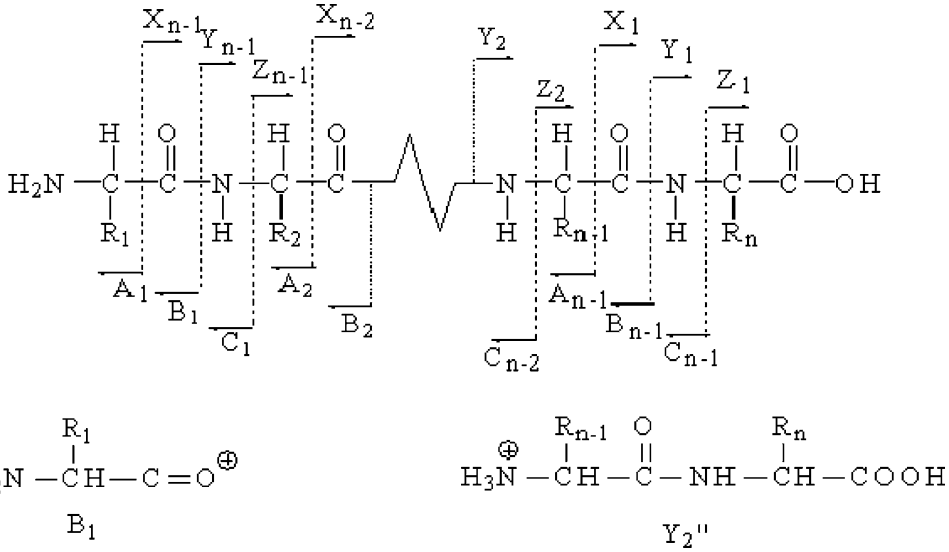

Fig. 1. The schematic of the common peptide fragment ions

The singly charged ion with C-terminal is denoted by y^{+}_{n-i}, and its *m/z* value is computed by

$$m\left(y^{+}_{n-i}\right) = 2 \times m(H) + m(OH) + \sum_{j=i+1}^{n} m\left(a_j\right) \tag{4}$$

The doubly charged ion with C-terminal is denoted by y^{++}_{n-i}, and its *m/z* value is computed by

$$m\left(y^{++}_{n-i}\right) = \left[m\left(y^{+}_{n-i}\right) + m(H)\right]/2 \tag{5}$$

From Eqs. 1 through 5, the following complementary equations

$$m(P)/2 \times m(H) = m\left(b^{+}_{n-i}\right) + m\left(y^{+}_{n-i}\right) \tag{6}$$

$$m(P)/2 + 2 \times m(H) = m\left(b^{++}_{i}\right) + \left(m\left(y^{+}_{n-i}\right) + m(H)\right)/2 \tag{7a}$$

$$m(P)/2 + 2 \times m(H) = \left(m\left(b^{+}_{i}\right) + m(H)\right)/2 + m\left(y^{++}_{n-i}\right) \tag{7b}$$

$$m(P)/2 + 2 \times m(H) = \left(m\left(b^{++}_{i}\right) + m(H)\right)/2 + m\left(y^{++}_{n-i}\right) \tag{8}$$

hold for a theoretical peptide spectrum. Therefore, Eqs. 6 through 8 indicate that high-quality spectra should have more complementary pairs of *m/z* values than poor-quality spectra.

According to the principle of peptide fragmentation in tandem mass spectrometry [13], these ions could lose a molecule of water or ammonia. Therefore, high-quality spectra should also have pairs of *m/z* values with differences of (half) a water molecular mass or an ammonia molecular mass for (doubly) singly charged ions, in contrast with poor-quality spectra. In addition, the N-terminal ions could lose a CO group, while C-terminal could lose an NH group, Therefore, high-quality spectra could have more pairs of *m/z* values with differences of (half) a CO mass or (half) an NH mass for (doubly) singly charged ions compared with poor-quality spectra.

In addition, for a theoretical spectrum, the difference between two consecutive singly charged N-terminal (C-terminal) ions is one of 20 amino acid mass weights. The difference between two consecutive doubly charged N-terminal (C-terminal) ions is half a mass weight of one of 20 amino acids. Therefore, high-quality spectra should also have more pairs of *m/z* values with difference of (half) an amino acid mass weight for (doubly) singly charged ions than poor-quality spectra.

Features of Peptide Mass Spectra

According to the properties of the theoretical spectra, we introduce 12 discriminatory features to describe the quality of peptide mass spectra. These features may be classified into four categories: amino acid distances, complements, water or ammonia losses, and supportive ions. To do this, we first define four variables for a given peptide mass spectrum S_E. For a peak in S_E with *m/z* value x, this peak is also denoted by x for simplicity. In the following, x and y are the *m/z* values of peaks x and y, respectively.

$$\mathrm{dif1}(x,y) = x - y, \quad x,y \in S_E \tag{9}$$

$$\mathrm{dif2}(x,y) = x - (y+1)/2, \quad x,y \in S_E \tag{10}$$

$$\mathrm{sum1}(x,y) = x + y, \quad x,y \in S_E \tag{11}$$

$$\mathrm{sum2}(x,y) = x + (y+1)/2, \quad x,y \in S_E \tag{12}$$

1. *Amino acid distances*: These features measure how likely two components in a peptide mass spectrum S_E are to differ by one of 20 amino acids. Let

$$\mathrm{DIF}_1 = \{(x,y)|\mathrm{dif1}(x,y) \approx M_i, i = 1,\cdots,17\}$$

$$\mathrm{DIF}_2 = \{(x,y)|\mathrm{dif1}(x,y) \approx M_i/2, i = 1,\cdots,17\}$$

$$\mathrm{DIF}_3 = \{(x,y)|\mathrm{dif2}(x,y) \approx M_i/2, i = 1,\cdots,17\}$$

where M_1,$\cdots$, M_{17} stand for the 17 mass weights of all 20 amino acids. In this study, we consider all methionine amino acids to be sulfoxidized and do not distinguish three pairs of amino acids in their masses: isoleucine vs. leucine, glutamine vs. lysine, and sulfoxidized methionine vs. phenylalanine. This is because the masses of each of these three pairs are very close. The comparison implied by $\approx$ uses a tolerance which is set to 0.5 Thompson in this study, but can be changed by the user. The set DIF_1 collects all pairs of singly charged ions in the spectrum S_E that are different from one amino acid. The set DIF_2 collects all pairs of doubly charged ions in the spectrum S_E that are different from one amino acid. The set DIF_3 collects all pairs of one doubly charged and the other singly charged ions that are different from one amino acid. Let

$$F_i = |\mathrm{DIF}_i|, \quad i = 1,2,3$$

where $|\bullet|$ represents the cardinality of a set. If a tandem mass spectrum is produced from a peptide with well fragmentation, one expects that values F_i $(i = 1,2,3)$ calculated from this spectrum should be much higher than those from a spectrum produced randomly.

2. *Complements*: These features measure how likely an N-terminus ion and a C-terminus ion in the peptide mass spectra S_E are to be produced as the peptide fragments at the same peptide bond. Let

$$\mathrm{SUM}_1 = \left\{(x,y)|\mathrm{sum1}(x,y) \approx M_{\mathrm{parent}} + 2 \times m(H)\right\}$$

$$\mathrm{SUM}_2 = \left\{(x,y)|\mathrm{sum1}(x,y) \approx M_{\mathrm{parent}}/2 + 2 \times m(H)\right\}$$

$$\mathrm{SUM}_3 = \left\{(x,y)|\mathrm{sum2}(x,y) \approx M_{\mathrm{parent}}/2 + 2 \times m(H)\right\}$$

where M_{parent} is the mass of the precursor ion of spectrum S_E. The set SUM_1 collects the complementary pairs of singly charged ions. The set SUM_2 collects the complementary pairs of doubly charged ions. The set SUM_3 collects the complementary pairs of one doubly charged ion and the other singly charged ion. For the same reason given for the first three features, we define another three features as the cardinalities of these three sets, i.e.,

$$F_{3+i} = |\mathrm{SUM}_i|, \quad i = 1,2,3$$

3. *Water or ammonia losses*: These features measure how likely one ion in a peptide mass spectrum S_E is to be produced by losing a water or an ammonia molecule from other ions. Let

$$\mathrm{WAD}_1 = \{(x,y)|\mathrm{dif1}(x,y) \approx M_{\mathrm{water}}\mathrm{or}M_{\mathrm{ammonia}}\}$$

$$\mathrm{WAD}_2 = \{(x,y)|\mathrm{dif1}(x,y) \approx M_{\mathrm{water}}\mathrm{or}M_{\mathrm{ammonia}}/2\}$$

$$\mathrm{WAD}_3 = \{(x,y)|\mathrm{dif2}(x,y) \approx M_{\mathrm{water}}\mathrm{or}M_{\mathrm{ammonia}}/2\}$$

where M_{water} and M_{ammonia} are the mass of a water molecule and an ammonia molecule, respectively. The set WAD_1 collects the pairs of singly charged ions with a difference of a water or an ammonia molecule. The set WAD_2 collects the pairs of doubly charged ions with a difference of a water or an ammonia molecule. The set WAD_3 collects the pairs of one doubly charged ion and the other singly charged ion with a difference of a water or an ammonia molecule. Similarly, we define the next three features as the cardinalities of these three sets, i.e.,

$$F_{6+i} = |\mathrm{WAD}_i|, \quad i = 1,2,3$$

One can consider the water losses and the ammonia losses separate features, but the resulting feature vector will have more components. In the classification problem, more features do not mean a better classifier. The reverse is often true, as the insignificant features could degrade the discriminatory power of other significant features [14].

4. *Supportive ions*: These features measure how likely one ion in a peptide mass spectrum S_E is to be a supportive ion. In this paper, we consider two kinds of supportive ions: a-ions and z-ions. Although a-ions and x-ions are complementary if a peptide fragments at the specific bond shown in Fig. 1, the a-ions are often generated by losing a CO group from b-ions [13], but not by fragmenting at the specific bond. For the same reason, we take z-ions into account but not c-ions

$$\mathrm{AZD}_1 = \{(x,y)|\mathrm{dif1}(x,y) \approx M_{\mathrm{CO}}\mathrm{or}M_{\mathrm{NH}}\}$$

$$\mathrm{AZD}_2 = \{(x,y)|\mathrm{dif1}(x,y) \approx M_{\mathrm{CO}}\mathrm{or}M_{\mathrm{NH}}/2\}$$

$$\mathrm{AZD}_3 = \{(x,y)|\mathrm{dif2}(x,y) \approx M_{\mathrm{CO}}\mathrm{or}M_{\mathrm{NH}}/2\}$$

where M_{CO} and M_{NH} are the mass of a CO group and an NH group, respectively. The set AZD_1 collects the pairs of singly charged ions with a difference of a CO or an NH group. The set AZD_2 collects the pairs of doubly charged ions with a difference of a CO or an NH group. The set AZD_3 collects the pairs of one doubly charged ion and the other singly charged ion with a difference of a CO or an NH group. Finally, we define the next three features as the cardinalities of these three sets, i.e.,

$$F_{9+i} = |\mathrm{AZD}_i|, \quad i = 1, 2, 3$$

At this point, we have introduced 12 features with physical meaning to describe the quality of peptide spectra. The four features F_j (j= 1, 4, 7, 10) are evidence of the existence of singly charged ions, called singly charged features. The other eight features are evidence of the existence of doubly charged ions. In principle, the high-quality spectra are expected to have larger feature values than the poor-quality spectra. However, the longer the peptide, the larger the feature values are. The classifier used for quality assessment may have low sensitivity, as the high-quality spectra produced from a shorter peptide would have smaller feature values. To alleviate these effects, we normalize the feature values by formula $F_i/\log(L_E)$, where L_E is the estimated peptide length of a peptide ion. L_E is computed by dividing the peptide ion mass by an average amino acid mass of 110 Da.

WKM

Let (x_i,i= 1,⋯,n) be a dataset of n objects (spectra) with the dimensionality of d. Let x_{ij} denote the jth feature of object x_i. X=(x_{ij}) is called a feature matrix of object set D. For a given partition Δ with K clusters, the cost function for a weighted K-means clustering technique [15] is defined by

$$J_G(\Delta) = \sum_{k=1}^{K} \sum_{x_i \in D_k} (x_i - \overline{m}_k) G (x_i - \overline{m}_k)' \tag{13}$$

where $\overline{m}_k = \frac{1}{n_k} \sum_{x_i \in D_k} x_i$, n_k are the mean and the number of objects in D_k, respectively, and G is an arbitrary symmetrical positive matrix whose determinant is 1, i.e.,$(\det(G)) = 1$.

The objective of a weighted K-means algorithm is to find an optimal partition expressed by Δ* and a symmetrical positive matrix G* with the determinant of 1 which minimize $J_G(\Delta)$, i.e.,

$$J_{G^*}(\Delta^*) = \min_{\Delta} \{ J_{G^*}(\Delta) \} \tag{14}$$

The problem is a constraint optimization problem. By the use of Lagrange multiplier, it can prove that a given partition Δ with K clusters

$$G = W^{-1} (\det(W))^{1/d} \tag{15}$$

where $W = \sum_{k=1}^{K} W_k$ and $W_k = \sum_{x_i \in D_k} (x_i - \overline{m}_k)'(x_i - \overline{m}_k)$ is the within-group variance of cluster k (k=1,⋯,K). Obviously, W is dependent on partition Δ. To avoid ambiguousness, denote W induced by Δ as $W(\Delta)$. Substituting Eq. 15 into 13 leads to $J(\Delta) = d \det(W(\Delta))$. Since d is constant, the cost function of a weighted K-mean algorithm can be reduced to

$$J(\Delta) = (\det(W(\Delta)))^{1/d} \tag{16}$$

Weighted K-means algorithm (WKMA)

$[\Delta_o, J(\Delta_o)] = WKMA(\Delta, X, K)$

Input: an initial partition, Δ; the feature matrix, X; the number of clusters, K

Output: the optimal partition, Δ_0 and its cost function value, $J(\Delta_0)$

1. **Compute** $\overline{m}_k$ and $J(\Delta)$ for an initial partition Δ.
2. **repeat**
3. **for** $i = 1$ to n
4. $k \leftarrow$ the index of the cluster where x_i belongs
5. **if** $n_k \neq 1$ **then** compute

$$\rho_l = \begin{cases} \dfrac{m_l}{m_l + 1} (x - \overline{x}_l)[W(\Delta')]^{-1} (x - \overline{x}_l)', & l \neq k \\ \dfrac{m_l}{m_l - 1} (x - \overline{x}_l)[W(\Delta')]^{-1} (x - \overline{x}_l)', & l = k \end{cases}$$

6. **if** $\rho_j = \min_l \rho_l < \rho_k$ **then** move x_i to D_j, adjust Δ, and re-compute $J, \overline{m}_j, \overline{m}_k$
7. **end for**
8. **until** no significant change of J in n consecutive attempts
13. **return** Δ_0 and $J(\Delta_0)$

Fig. 2 Iterative optimal weighted K-means algorithm

Table 1 The mean centers for SMP2

Cluster 1 (11,356)	12.20	14.65	2.99	0.28	0.01	0.06
	2.10	1.68	0.14	1.95	2.03	0.18
Cluster 2 (7,131)	59.64	68.57	18.00	1.66	0.01	0.35
	9.57	7.55	0.95	7.75	7.82	1.03

The number of spectra in cluster 1 is 11,356, while that in cluster 2 is 7,131. Because each spectrum is represented by a 12-dimensional feature vector, the cluster center of each cluster is also a 12-dimensional vector. The meaning of each value of the vector is the corresponding introduced feature in the "Features of Peptide Mass Spectra" section

The objective of a weighted K-mean algorithm becomes to find an optimal partition expressed by Δ_o which minimizes

$$J(\Delta_o) = \min_{\Delta}(\det(W(\Delta)))^{1/d} \tag{17}$$

Now consider how the cost function J changes when an object x currently in cluster D_i tentatively moves to a different cluster D_j. Let $\Delta = (D_1, \cdots, D_K), \Delta' = (D_1, \cdots, D_i \backslash \{x\}, \cdots D_K)$, and $\Delta'' = (D_1, \cdots, D_i \backslash \{x\}, \cdots, D_j \cup \{x\}, \cdots, D_K)(i \neq j)$. Obviously the condition for successfully moving x from D_i into D_j is

$$\det(W(\Delta'')) < \det(W(\Delta)) \tag{18}$$

The following two equations can be derived from the definitions

$$W(\Delta) = W(\Delta') + \frac{m_i}{m_i - 1}(x - \bar{x}_i)'(x - \bar{x}_i) \tag{19}$$

Table 3 The mean centers for SMP 3

Cluster 1 (7,739)	19.21	21.12	9.26	0.28	0.13	1.42
	3.12	2.61	0.57	2.32	2.33	0.61
Cluster 2 (10,305)	2.78	3.06	1.20	0.03	0.02	0.21
	0.40	0.35	0.08	0.35	0.36	0.09

The number of spectra in cluster 1 is 7,739, while that in cluster 2 is 10,305. Because each spectrum is represented by a 12-dimensional feature vector, the cluster center of each cluster is also a 12-dimensional vector. The meaning of each value of the vector is the corresponding introduced feature in the "Features of Peptide Mass Spectra" section

$$W(\Delta'') = W(\Delta') + \frac{m_j}{m_j + 1}(x - \bar{x}_j)'(x - \bar{x}_j) \tag{20}$$

Condition 18 is reduced to

$$\frac{m_j}{m_j + 1}(x - \bar{x}_j)[W(\Delta')]^{-1}(x - \bar{x}_j)' < \frac{m_i}{m_i - 1}(x - \bar{x}_i)[W(\Delta')]^{-1}(x - \bar{x}_i)' \tag{21}$$

since $\det(A + \beta y' y) = \det(A)(1 + \beta y A^{-1} y')$ for any $d \times d$ invertible matrix A, any d-dimensional row vector y, and any number β.

If reassignment is profitable, the greatest decrease in the cost function is obtained by selecting the cluster for which $\frac{m_j}{m_j+1}(x - \bar{x}_j)[W(\Delta')]^{-1}(x - \bar{x}_j)'$ is minimal. According to the above discussion, an iterative optimal weighted K-algorithm is designed and shown in Fig. 2.

Table 2 Number of spectra in two clusters with respect to SEQUEST score

SEQUEST score	No. of spectra in cluster 1	No. of spectra in cluster 2
≥2.0	496	2,050
≥2.2	315	1,471
≥2.4	221	1,181
≥2.6	164	969
≥2.8	127	822
≥3.0	102	689
≥3.2	80	578
≥3.4	58	486
≥3.6	47	408
≥3.8	39	351
≥4.0	30	290

Columns 2 and 3 list the number of spectra in clusters 1 and 2 whose SEQUEST scores are bigger than a threshold listed in column 1, respectively

Table 4 Number of spectra in two clusters with respect to SEQUEST score

SEQUEST score	No. of spectra in cluster 1	No. of spectra in cluster 2
≥2.0	5,776	2,448
≥2.2	4,402	1,494
≥2.4	3,045	869
≥2.6	2,055	445
≥2.8	1,378	217
≥3.0	978	94
≥3.2	787	39
≥3.4	614	21
≥3.6	499	12
≥3.8	414	9
≥4.0	339	4

Columns 2 and 3 lists the number of spectra in clusters 1 and 2 whose SEQUEST scores are bigger than a threshold listed in column 1, respectively

Experiments and Results

Dataset

This study employs the standard protein mixture (SPM) dataset acquired on an ion trap mass spectrometer [16, 17] to investigate the performance of the proposed method. This dataset consists of 37,044 peptide tandem spectra collected in 22 HPLC/MS/MS runs. The samples analyzed were generated by the tryptic digestion of a control mixture of standard 18 proteins (not of human origin) [16, 17]. The MS/MS spectra were searched using SEQUEST against a human protein database appended with the sequences of the 18 standard proteins and other common contaminants. The SEQUEST will be used to verify the clustering results.

The spectra with different charges have significant different properties. This study applies the proposed method to two subsets of the SPM dataset: one subset consisting of all 18,496 doubly charged spectra (denoted by SPM2) and the other consisting of all 18,044 triply charged spectra (denoted by SPM3). All singly charged spectra are ignored in this study.

Results

Using the proposed method, SMP2 is divided into two clusters: cluster one consisting of 11,365 spectra and cluster two consisting of 7,131 spectra. Table 1 lists the mean centers of two clusters. Obviously, the spectra in cluster 2 are of high quality, while those in cluster 1 are of poor quality because the mean center of cluster 2 is much larger than that of cluster 1.

Table 2 shows the number of spectra with the SEQUEST scores greater than a variety of threshold values. It indicates that the majority of the spectra with higher SEQUEST scores are in cluster 2. Generally, if the SEQUEST score of a doubly charged spectrum is greater than 2.5, this spectrum is considered to be identified (well interpreted). If we used the SEQUEST score of 2.6 as the cutoff value, $85.53\%(= 969/(164 + 969))$ of the interpretable spectra are in cluster 2. In other words, if we just search spectra in cluster 2 using a database, we can save $61.45\%(= 11,365/(11,365 + 7,131))$ of the time while only losing 14.47% (=1–85.53%) of the interpretable spectra.

Using the proposed method, SMP3 is also divided into two clusters: cluster one consisting of 7,739 spectra and cluster two consisting of 10,305 spectra. Table 3 lists the means centers of two clusters. Obviously, the spectra in cluster 1 are of high quality, while those in cluster 2 are of poor quality, as the mean center of cluster 1 is much larger than that of cluster 2.

Table 4 shows the numbers of spectra with the SEQUEST scores greater than a variety of threshold values. It indicates that the majority of the spectra with higher SEQUEST scores are in cluster 1. Generally, if the SEQUEST score of a triply charged spectrum is greater than 3.5, this spectrum is considered to be identified (well interpreted). If we used the SEQUEST score of 3.6 as the cutoff value, $97.65\%(499/(12 + 499))$ of the interpretable spectra are in cluster 1. In other words, if we just search spectra in cluster 1 using a database, we can save $57.11\%(= 10,305/(10,305 + 7,739))$ of the time while losing only 2.35% (=1–97.65%) of the interpretable spectra.

In summary, considering the SMP2 and SMP3 as a whole dataset, if we just search spectra in the clusters with high quality by SEQUEST, we can save $59.3\%(= (10,305 + 11,365)/(11,365 + 7,131 + 10,305 + 7,739))$ of the time while losing only $10.71\%(= 1-(969 + 499)/(164 + 969 + 12 + 499))$ of the interpretable spectra in the cluster with poor quality.

Conclusions and Future Work

The evaluation of tandem mass spectra is important for the reduction of the database search time. This study has proposed a method of classifying tandem mass spectra into one group of mass spectra with high quality and one with poor quality. Computational experiments illustrate that if we just search the spectra in the high-quality group, we can save about 60% of searching time while losing only about 10% of high-quality spectra. This result indicates that the proposed method is useful in saving database search time because it ignores the spectra in the cluster with poor quality.

In this study, the proposed method has been applied to raw tandem mass spectra which were noise-contaminated. Recently, we have developed a method to denoise raw tandem mass spectra [18], which can improve the reliability of peptide identification. It could make more sense and improve the reliability of tandem mass spectral quality assessment by classifying denoised mass spectra. One direction of our future work is to combine the denoising method with quality assessment methods to improve the reliability of mass spectral quality assessment.

Acknowledgments This research is supported by Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to thank Dr Andrew Keller from Institute for Systems Biology for generously providing spectral data and protein databases for SPM dataset in this paper.

References

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207.

2. Eng KJ, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequence in a protein database. J Am Soc Mass Spectrom 1994;5:976–89.
3. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence database using mass spectrometry data. Electrophoresis 1999;20:3551–67.
4. Field HI, Fenyö D, Beavis RC. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relation database. Proteomics 2002;2:36–47.
5. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem 2005;77:964–73.
6. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 2003;17:337–2342.
7. Bern M, Goldberg D, McDonald W, Yates J. Automatic quality assessment of peptide tandem mass spectra. Bioinformatics 2004;20:i49–i54.
8. Salmi J, Moulder R, Filen J, Nevalainen O, Nyman T, Lahesmaa R, Aittokallio T. Quality classification of tandem mass spectrometry data. Bioinformatics 2006;22:400–6.
9. Flikka K, Martens L, Vandekerckhove J, Gevaert K, Eidhammer I. Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. Proteomics 2006;6:2086–94.
10. Na S, Paek E. Quality assessment of tandem mass spectra based on cumulative intensity normalization. J Proteome Res 2006;5:3241–8.
11. Nesvizhskii A, Roos F, Grossmann J, Vogelzang M, Eddes J, Gruissem W, Baginsky S, Aebersold R. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data. Mol Cell Proteomics 2006;5:652–70.
12. Wu FX, Gagne P, Droit A, Poirier GG. Quality assessment of peptide tandem mass spectra. BMC Bioinformatics 2008;9:S13.
13. Kinter M, Sherman NE. Protein sequencing and identification using tandem mass spectrometry. New York: Wiley; 2000.
14. Ding J, Shi JH, Zou AM, Wu FX. Feature selection for tandem mass spectrum quality assessment, Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, 2008, pp: 310–13.
15. Spath H. Cluster analysis algorithms for data reduction and classification of objects. West Sussex, UK: Ellis Horwood Limited; 1975.
16. Keller A, Purvine S, Nesvizhskii AI, Stolyar S, Goodlett DR, Kolker E. Experimental protein mixture for validating tandem mass spectral analysis. OMICS 2002;6:207–12.
17. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–92.
18. Ding J, Shi JH, Poirier GG, Wu FX. A novel approach to denoising tandem mass spectra, BMC Proteome Science, Accepted, 2009.