View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Saskatchewan's Research Archive

Filtering Methods for Mass Spectrometry-based Peptide Identification Processes

A Thesis Submitted to the

College of Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the Degree of Master of Science

in the Department of Biomedical Engineering

University of Saskatchewan

Saskatoon

By

Wenjun Lin

© Wenjun Lin, October 2013. All rights reserved.

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from the University of Saskatchewan, I agree that the libraries of this university may make it freely available for inspection. I further agree that permission to copy this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the head of the department or the dean of the college in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use that may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Division of Biomedical Engineering

University of Saskatchewan

57 Campus Drive

Saskatoon, Saskatchewan

S7N 5A9

Canada

Abstract

Tandem mass spectrometry (MS/MS) is a powerful tool for identifying peptide sequences. In a typical experiment, incorrect peptide identifications may result due to noise contained in the MS/MS spectra and to the low quality of the spectra. Filtering methods are widely used to remove the noise and improve the quality of the spectra before the subsequent spectra identification process. However, existing filtering methods often use features and empirically assigned weights. These weights may not reflect the reality that the contribution (reflected by weight) of each feature may vary from dataset to dataset. Therefore, filtering methods that can adapt to different datasets have the potential to improve peptide identification results.

This thesis proposes two adaptive filtering methods; denoising and quality assessment, both of which improve efficiency and effectiveness of peptide identification. First, the denoising approach employs an adaptive method for picking signal peaks that is more suitable for the datasets of interest. By applying the approach to two tandem mass spectra datasets, about 66% of peaks (likely noise peaks) can be removed. The number of peptides identified later by peptide identification on those datasets increased by 14% and 23%, respectively, compared to previous work (Ding et al., 2009a). Second, the quality assessment method estimates the probabilities of spectra being high quality based on quality assessments of the individual features. The probabilities are estimated by solving

a constraint optimization problem. Experimental results on two datasets illustrate that searching only the high-quality tandem spectra determined using this method saves about 56% and 62% of database searching time and loses 9% of high-quality spectra.

Finally, the thesis suggests future research directions including feature selection and clustering of peptides.

Key words: Tandem mass spectrometry, peptide identification, denoise, quality assessment.

Acknowledgements

To the members of my advisory committee for advice and supervision:

Dr. Fang-Xiang Wu, Ph.D. (Co-supervisor)

Dr. Chris Zhang, Ph.D. (Co-supervisor)

Dr. Saadat Mehr Aryan, Ph.D.

Dr. Tony Kusalik, Ph.D.

Dr. Randall Purves, Ph.D.

For scientific advice:

Mr. Steve Ambrose

Dr. Haixia Zhang, Ph.D.

To the members and friends of my research group:

Bolin Chen

Weiwei Fan

Jianwei Li

Lizhi Liu

Jinhong Shi

Yan Yan

To the following organizations for financial support:

Natural Sciences and Engineering Research Council of Canada (NSERC)

University of Saskatchewan (UofS)

Contents

List of Tables

List of Figures

List of Abbreviations

CID	Collision-induced Dissociation
LC	Liquid Chromatography
LDA	Linear Discriminant Analysis
MS/MS	Tandem Mass Spectrometry
SVM	Support Vector Machine
ISB	Institute for Systems Biology
TOV	Tumours of the Ovary

Chapter 1. Introduction

1.1 Background

1.1.1 Definition of peptides

Proteins are the main components of living cells and organisms. They are made of amino acids that are arranged in linear chains and usually folded into a three-dimensional form (Maton, 1993). Within the protein, a peptide is a short sequence of amino acids that does not have a three-dimensional structure. Since the structure and behaviour of peptides are highly related to those of proteins, understanding peptides is an important part of research in proteomics (Anderson & Anderson, 1999).

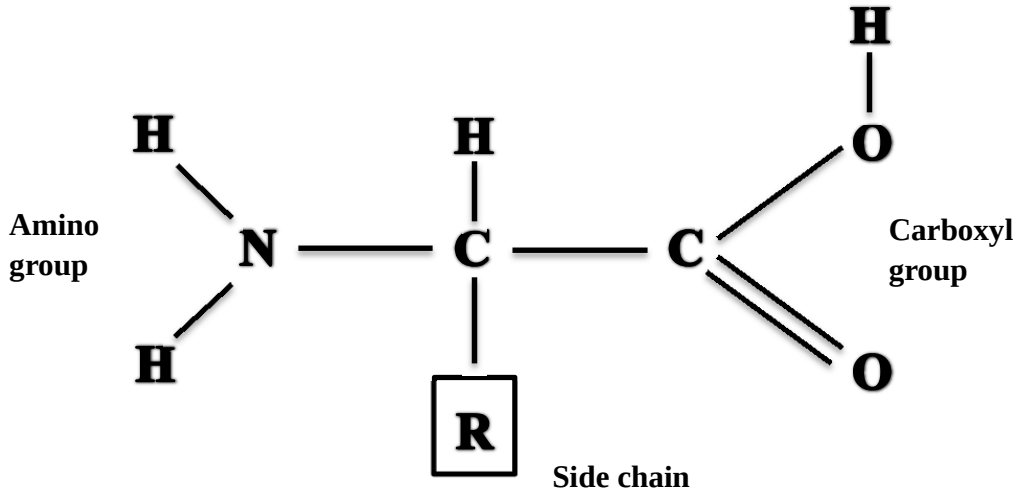


Figure 1.1 Structure of amino acids.

Amino acids, the basic elements of peptides and proteins, are molecules containing an amino group, a carboxyl group, and a side chain (Maton, 1993), as shown in Figure 1.1.

One nitrogen atom and two hydrogen atoms form the amino group ($-NH_2$), and one carbon atom, one oxygen atom and one hydroxyl (-OH) group form the carboxyl group (-COOH). The composition of the side-chain varies between different amino acids. There are 20 standard amino acids in nature: each one has been assigned a letter code for simplicity in use; e.g., A, R, N, D (Maton, 1993).

1.1.2 The process of tandem mass spectrometry of peptides

Peptide sequencing, which aims to determine the order of amino acids in a peptide, is a very important task in the process of identifying proteins and their primary structures. Currently, tandem mass spectrometry (MS/MS) is one of the most popular experimental methods for peptide sequencing. The process of an MS/MS experiment happens in three steps, as shown in Figure 1.2: 1) sample preparation and peptide separation, 2) tandem mass spectrometer analysis, and 3) peptide identification (Nesvizhskii, 2010). A typical tandem mass spectrometer has two mass analyzers. The first analyzer measures the *m/z* (mass over charge) values of peptide ions and selects the desired ions called precursor ions. The precursor ions are fragmented into smaller ions called fragment ions. The second mass analyzer measures the *m/z* values and intensities of fragment ions, which yields an MS/MS spectrum.

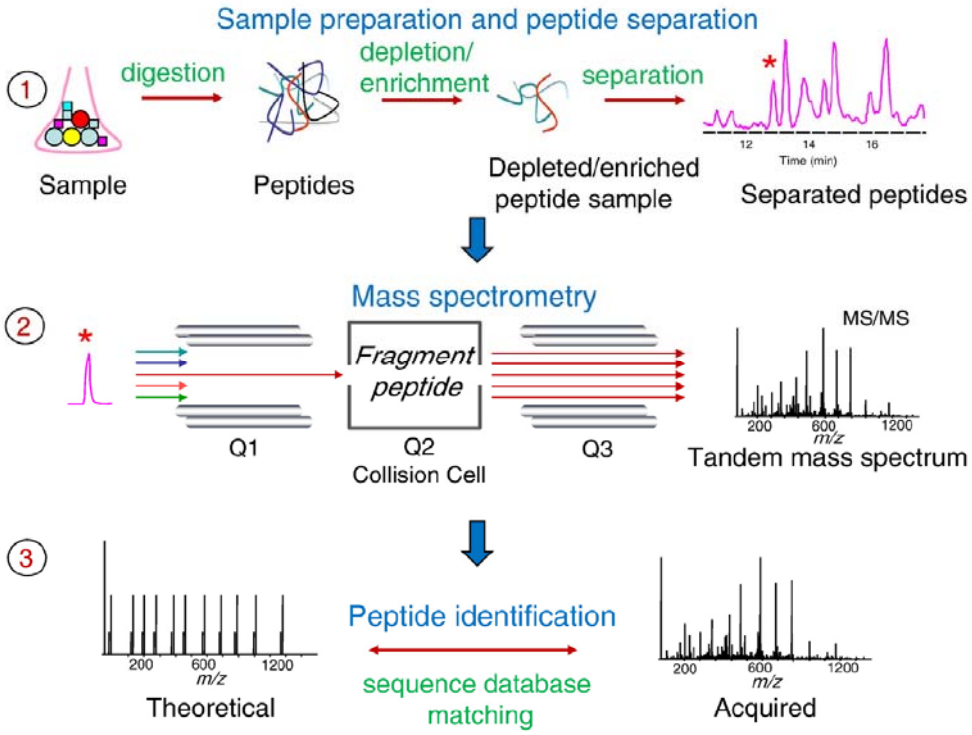


Figure 1.2 Overview of shotgun proteomics (Nesvizhskii, 2010). ① Sample proteins are digested into peptides using enzymes such as trypsin. Resulting peptide mixtures are optionally processed to capture a particular class of peptides, and then separated using a liquid chromatography (LC) system coupled online to a mass spectrometer. ② Peptides are subjected to MS/MS analysis that results in the acquisition of MS/MS spectra. ③ MS/MS spectra are assigned to peptide sequences through database search.

In the MS/MS experiment, peptide's *m/z* value is related to its mass. During ionization typically a proton(s) is added for each charge. Therefore, the *m/z* value of a peptide, denoted as *m/z(p)*, is calculated as:

$$m/z(p) = \frac{m(p)+z\times m(H)}{z} \quad (1.1)$$

where z is the charge number of the peptide, $m(p)$ is the mass of the peptide, and $m(H)$ is the mass of hydrogen. Assuming that a peptide $p = a_1 \ldots a_n$ consists of n amino acids, where a_i, $i = 1, \ldots, n$ is one of the 20 amino acids, $m(a_i)$ is the mass of a_i, and the

mass *m(p)* can be calculated as:

$$m(p) = m(H) + m(OH) + \sum_1^n m(a_i) \quad (1.2)$$

where $m(a_i)$ is the mass of the amino acids; and *m(OH)* is the mass of hydroxide.

In an MS/MS experiment, precursor ions are typically fragmented into six kinds of fragment ions (a-ion, b-ion, c-ion, x-ion, y-ion, and z-ion) along the peptide backbone. Their letter-names indicate peptide fragments that are fractured in different positions in the MS/MS spectrum. Figure 1.3 shows the different cleavage sites and ion types in detail. The N-terminal of a peptide refers to a peptide fragment that is terminated by an amino acid with a free amine group. The C-terminal refers to a peptide that is terminated by an amino acid with a free carboxyl group (Aebersold & Mann, 2003).

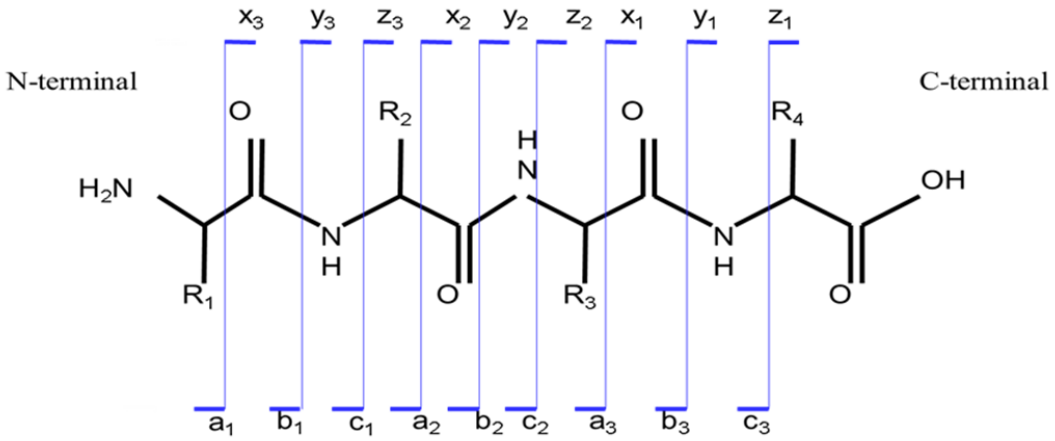


Figure 1.3 Fragmentation of a peptide (Mujezinovic et al., 2006). For example, the peptide GPFR may be broken into the N-terminal ions G, GP, GPF (donate as b_1, b_2, and b_3 for b- type ions), and C-terminal ions PFR, FR, R (donate as y_3, y_2, and y_1 for y- type ions)

Different fragmentation techniques used in MS/MS yield different dominating types of fragment ions. Collision-induced dissociation (CID) is the most commonly used fragmentation technique, and yields b-ions and y-ions as dominating ions. Given an ion generated by a partial peptide containing k amino acids (note: the peptide contains n amino acids in total), its m/z value can be calculated by:

$$m/z(b^{z+}) = (m(H) + \sum_1^k m(a_i))/z \quad (1.3)$$

$$m/z(y^{z+}) = (m(H) + m(H_2O) + \sum_{n-k}^n m(a_i))/z \quad (1.4)$$

where b^{z+} is the b-ion with charge z and y^{z+} is the y-ion with charge z. In an MS/MS experiment, fragment ions can lose some small molecules such as H_2O and NH_3.

Peptide sequences need to be inferred from the MS/MS spectra—a process called peptide identification. In the literature, two methods are used for peptide identification with MS/MS: database searching (Yates, 1998) and *de novo* sequencing (Ma et al., 2003).

The database searching approach compares an experimental spectrum with a theoretical spectrum constructed from the database to find a peptide whose theoretical spectrum best matches the experimental data. The successful matching thus identifies the peptide in the spectrum. Construction of the theoretical spectra requires two types of information: *m/z* values and intensities. The *m/z* values are determined by the types of fragment ions that may appear in the experimental spectra. The simplest way of finding them is to construct *m/z* values for all types of fragment ions; an alternative approach is only to consider those

types of ions with a high probability of appearing. Peak intensity can be determined by the type or position of the fragment ion, and the length, sequence, mass, etc. of the peptide. The similarity between the experimental spectrum and the theoretical one is evaluated and scored. The highest-scoring theoretical spectrum is thus selected, and its corresponding peptide is taken as the best candidate to represent the peptide in the given experimental spectrum.

De novo sequencing, on the other hand, estimates peptide sequences without the help of a database; it infers the sequences using the spectrum and the masses of amino acids. In *de novo* peptide sequencing, spectrum graph modeling has proven to be quite successful and hence is widely used. When using this model, a set of nodes and edges must be defined. Each peak of the spectrum is defined as a node. When two nodes have an *m/z* difference that corresponds to the mass of an amino acid residue, this is defined as a directed edge (the edge always goes from a lower mass to a higher one). The main idea of this approach is to find paths in the graph for which corresponding peptides provide a good explanation of the experimental spectrum. Since the *de novo* sequencing approach does not rely on a sequence database, it is useful in identifying new proteins, such as proteins resulting from mutations, proteins with unexpected modifications, etc. (Eidhammer et al., 2007).

1.2 Occurrences of noise and low-quality spectra

The accuracy of peptide identification using both approaches is a concern (Aebersold & Mann, 2003). Inaccuracy or error may result from incomplete information of fragment ions and/or noise in the MS/MS spectra. The manual verification of peptide assignments to spectra from peptide identification programs can achieve a high-confidence result if the process is performed by experienced researchers. However, this approach only works on small datasets (for example, dozens of spectra). In the case of high-throughput analysis of large datasets (for example, thousands of spectra), this approach is extremely time-consuming. Further, the way ions are fragmented in the mass spectrometer is poorly understood, making it difficult to improve quality by developing algorithmic solutions from the perspective of the ion fragmentation principle (Salmi et al., 2009).

1.3 Related work

Research on removing noise from MS/MS spectra and on screening out low-quality MS/MS spectra has been very active (Salmi et al., 2009). Many filtering methods have been developed to complement the operation of peptide identification. Two strategies are discussed in this thesis. The first strategy is to filter out noisy information from MS/MS spectra prior to identification. The second strategy is to reject low-quality spectra from MS/MS datasets.

The first strategy is MS/MS spectrum denoising (Rejtar et al., 2004; Resing et al., 2004;

Baginsky et al., 2005; Grossmann et al., 2005; Ning & Leong, 2007; Zhang, et al., 2008; Ding et al., 2009a), which keeps signal peaks (reflecting peptide fragment ions) and removes noisy peaks (not reflecting peptides or their fragment ions). Baginsky et al. (2005) calculated a series of spectra features that included peak intensity, the presence of complement peaks, isotope peaks, and ammonia loss and water loss from amino acids. Based on those feature values and on relevant weights specified by the user, their method removes noisy peaks from MS/MS spectra. However, the features alone cannot properly identify signal peaks, and it takes an experienced user to specify the weights. In addition, Zhang et al. (2008) built a denoising model that focused on isotope features that could be observed only from signal peaks, not from noise peaks. Their work decreased the computational time of the subsequent process and increased the reliability of peptide identification. Ding et al. (2009a) developed a feature-based method for denoising MS/MS spectra. Their method first introduced five features to describe the quality of peaks and then calculated a score for each peak by a linear combination of those five features. The intensities of the peaks in a spectrum were adjusted by their corresponding scores, after which, the intensities of signal peaks presumably became local maxima. The spectra were then processed using a morphological reconstruction filter to remove those peaks whose intensities were not local maxima. Experimental results on several datasets showed that this method was both efficient and effective. However, in calculating the scores, the coefficients (weights) of the linear combination were fixed and determined empirically, making this method potentially unsuitable for use with other datasets.

The second strategy is quality assessment (Koenig et al., 2008; Na & Paek., 2006; Frank et al., 2008; Tabb et al., 2005; Bern et al., 2004; Ding et al., 2009b; Ding et al., 2011), which screens out low quality MS/MS spectra (containing insufficient fragment ions) from the dataset. Based on defined features, these methods assess the quality of MS/MS spectra through the use of supervised machine learning methods that require labelled training datasets to train a classifier. The trained classifier is then used to classify the spectra as high quality or poor quality. Ideally, the training data should be validated by peptide identification algorithms or manual verification (that is, the data should be correctly labelled with no, or with very few, falsely labelled spectra). However, this information is hard to obtain prior to the peptide identification of new datasets. Furthermore, tandem mass spectrometers may produce different spectra for the same peptide under different experimental conditions. A classifier trained by one dataset may not be effective on another. Therefore, unsupervised machine learning methods may be more effective for assessing the quality of MS/MS spectra.

1.4 Research objectives

The goal of this study is to develop adaptive filtering methods to improve the effectiveness and efficiency of peptide identification. To achieve this goal, two specific research objectives were proposed as follows:

Objective (I): To develop an adaptive denoising method for MS/MS spectra that removes noise while retaining as many signal ions as possible.

Objective (II): To develop an adaptive quality assessment method for MS/MS spectra that rejects low-quality spectra while retaining high-quality ones.

1.5 Overview of the main contribution of and organization of this thesis

In this thesis, adaptive filtering methods have been developed for the two objectives, as mentioned above. The novelty of the thesis is such that the feature weights (or parameters) associated with noise or spectra quality are adjusted to tailor to different data. Additionally, in the denoising method, although a supervised learning method [linear discriminant analysis (LDA)] is used, training data are generated from MS/MS spectra rather than from the peptide identification result. This makes it a viable preprocessing method. The novelty of the equality assessment method is its unsupervised learning nature which makes it practical for incorporation into new or unknown datasets.

In Chapter 2, an adaptive approach is proposed for estimating weights of selected features used in the spectrum denoising. This new approach first adjusts the intensities of spectra using scores calculated with given weights, and then selects signal peaks according to

their adjusted intensity. Unlike the work of others (Ding et al., 2009a) wherein the weights are fixed and empirically assigned, this new approach employs an adaptive method for estimating weights by iteration. The results show that about 66% of peaks (likely noise peaks) can be removed and that the number of identified peptides is increased by 14% and 23% for ISB and TOV-Q datasets, respectively, compared to the Ding et al.'s work (2009a).

In Chapter 3, an unsupervised machine learning method is proposed for quality assessment of MS/MS spectra without training data. This method estimates the probabilities of spectra being high quality using quality assessments based on a constraint optimization problem. Experimental results on two datasets illustrate that searching only the high-quality tandem spectra determined saves about 56% and 62%, respectively, of database searching time and loses about 9% of high-quality spectra.

Finally, a general discussion in Chapter 4 summarizes the thesis. Concluding remarks and a summary of the overall contributions are also provided. The full list of publications arising from the thesis is included in Appendix A, and the copyright permissions of included manuscripts are in Appendix B.

Chapter 2. MS/MS spectrum denoising

2.1 Introduction

In this chapter, an adaptive approach is proposed to determine the weights in Ding et al.'s method (2009a). Section 2.2 provides an overview of the spectrum denoising method and then discusses how to adjust weights adaptively. In section 2.3, the performance of this new method is evaluated using both high- and low-resolution MS/MS datasets. Concluding remarks are expressed in Section 2.4.

2.2 Method

2.2.1 Overview of the spectrum denoising method

The spectrum denoising method was initially proposed by Ding et al. (2009a). This method consists of two steps: peak intensity adjustment and local maximum extraction. The peak intensity adjustment is based on the following five design features:

1. Number of peaks whose mass differences from a given peak approximately equal the mass of one of the 20 amino acids.
2. Number of peaks whose mass added to a given peak approximately equal the mass of the precursor ion.
3. Number of peaks that could have been produced by losing a water molecule or an ammonia molecule from a given peak.

4. Number of peaks that have an *m/z* difference equal to a CO group or an NH group compared to a given peak.
5. Number of isotope peaks associated with a given peak.

These five features are generated from the observation of theoretical MS/MS spectra. Peaks with larger feature values are likely to be signal peaks. Based on the five features and their corresponding weights, a linear combination of their values is used to score each peak as follows:

$$score = \omega_1 f_1 + \omega_2 f_2 + \omega_3 f_3 + \omega_4 f_4 + \omega_5 f_5 \quad (2.1)$$

where f_i (i = 1,..., 5) is the normalized value of each feature (mean=1 and variance=1) and ω_i (i = 1,...,5) are the weights. The means of the features are set to 1 to ensure that only a few peaks have negative scores. In Ding et al.'s work (2009a), ω_1 and ω_2 are set to 1.0; both ω_3 and ω_4 are set to 0.2; and ω_5 is set to 0.5. These values are selected according to the normalization method of the SEQUEST algorithm (Eng et al., 1994). However, in Ding et al.'s work, weights are determined for all data (details of the method are described in the following section). In applying Equation (2.1) to each spectrum, peaks with high scores tend to be signals, whereas peaks with lower scores are more likely to be noise.

In addition, intensity is an important attribute of a peak in a spectrum. Empirical approaches usually assume that peaks with high intensities are more likely to be signal

peaks than those with low intensities. Thus, the intensity of a peak is adjusted by its corresponding score. The intensities of peaks with high scores are increased while the intensities of peaks with low scores are decreased after the peak intensity adjustment. After the adjustment, intensities of signal peaks are expected to be a local maximum of the spectrum.

The second step in Ding et al.'s method employed a so-called morphological reconstruction filter (Vincent, 1993) to select signal peaks. The filter selects the peaks that have a local maximum of intensities by comparing a peak to its two adjacent peaks once other peaks have been temporarily removed.

Ding et al.'s method (2009a) removes about 69% of the noise peaks. After denoising, the number of spectra that can be identified by the peptide identification algorithm (Perkins et al., 1999) increases by 31% and 14% on two MS/MS datasets. The difference in improvements may be due to differences in the quality of peaks in these datasets. The denoising method is less efficient for spectra with fewer peaks.

Furthermore, both ISB and TOV are low-resolution datasets. It is unclear whether Ding et al.'s method is effective for high-resolution datasets. In Ding et al.'s method, weights in the algorithm were empirically assigned. Patterns of signal peaks in a low-resolution dataset may differ significantly from those in a high-resolution dataset. In short, the

success of Ding et al.'s method for the spectra in low-resolution datasets cannot be generalized to spectra with high resolution.

2.2.2 Adaptive weighting with LDA

The general idea of the adaptive weighting approach is as follows: given a high or low resolution spectra dataset, the first step is to find the highest and lowest scores from Equation (2.1). The second step is to adjust the weights with LDA.

The morphological reconstruction filter (Vincent, 1993) is used to extract signal peaks based on intensities. The intensities are then adjusted by the corresponding score from Equation (2.1). Here, LDA is used to estimate weights such that the scores can separate the two groups of peaks (signal and noise) as far apart as possible. LDA was originally used to separate two classes by finding a linear combination of features (Fisher, 1936). By taking signal peaks and noise peaks as two classes, LDA calculates the weights from the linear combination. The entire framework of this proposed approach is shown in Figure 2.1.

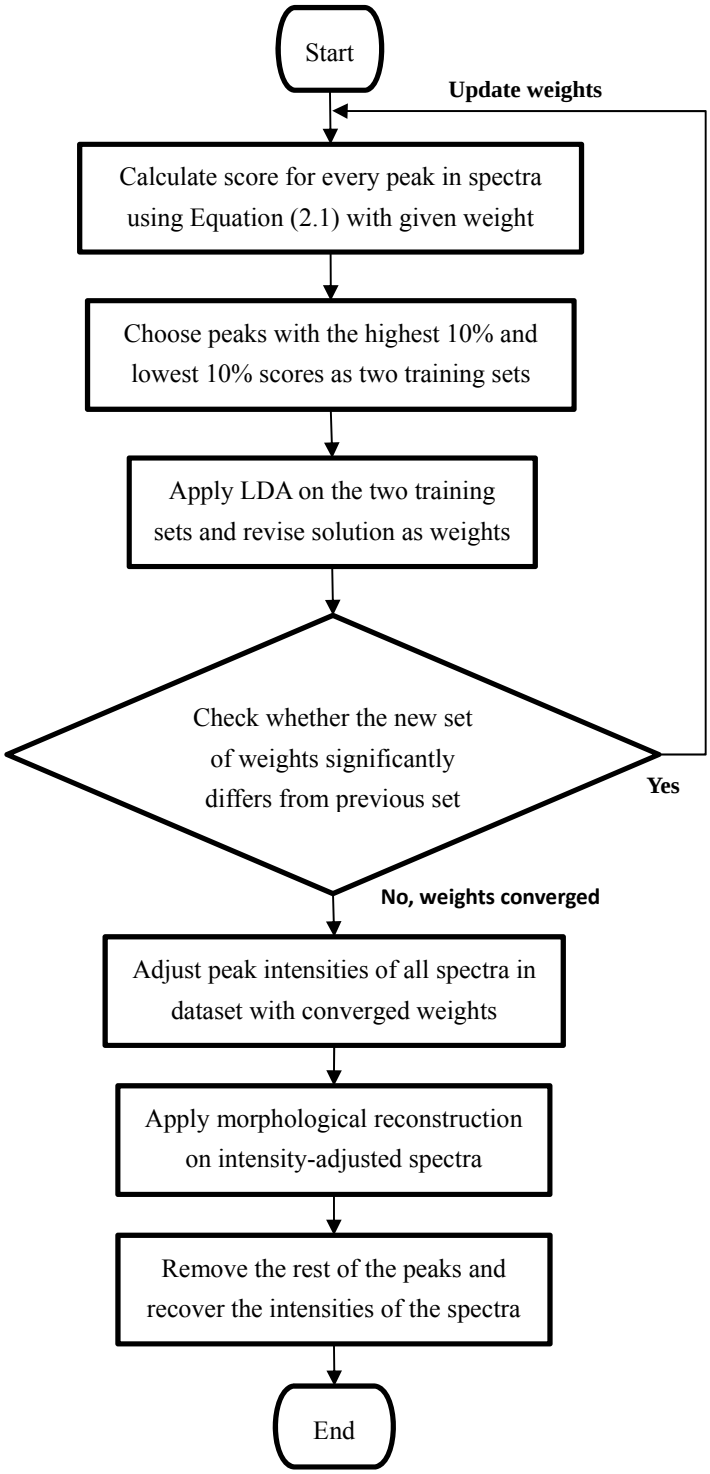


Figure 2.1 Procedure of proposed approach

From Figure 2.1, the proposed approach conducts spectra denoising in datasets with a set

of adaptive weights as determined by Equation (2.1). Details of the key steps are described as follows:

1. For each spectrum, the score for each peak is calculated by Equation (2.1) with the given weights. Initially, the weights are set as the ones used in Ding et al.'s work (2009a).
2. Sort all the peaks by their scores, then choose best and worst 10% scored peaks as training sets. The best 10% peaks have a higher chance of being kept as signal peaks after denoising while the worst 10% peaks are more likely to be removed. Although the spectrum actually contains more than 10% signal peaks or noisy peaks, choosing the most significant ones allows the best potential for finding a good set of weights for Equation (2.1).
3. Using the two training sets, the weights in Equation (2.1) are updated by LDA in which the maximum separation (Fisher, 1936), denoted as S, is achieved by:

$$S = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{w' S_B w}{w' S_W w} \tag{2.2}$$

where w is the parameter vector or in other words, the weights in Equation (2.1), and w' is the transposition of w. S_B is the between-class scatter matrix and S_W is the within-class scatter matrix. Maximum separation means that in the given 5-dimensional feature observations, the difference between the two sets is as significant as possible while the difference within each set is as trivial as possible. The solution for the problem is calculated by the following equation:

$$w = S_W^{-1}(\mu_1 - \mu_2) \tag{2.3}$$

where μ_1 and μ_2 are the means of the signal and noisy sets, respectively.

Since all the features chosen for scoring should be observed from the signal peaks, all the weights should be positive. However, the weights as calculated by LDA sometimes contain negative values. To make sure every signal peak attains a high score, the optimized weights are revised based on two rules: (1) If most of the weights (greater than or equaling three weights) are negative, all the weights should be revised; and (2) if a portion of the weights (less than three) are negative, they should be replaced by the weights in the previous round of the loop. Additionally, since the morphological reconstruction filter is affected only by the ratio between the weights, the optimized weights are normalized to make the maximum weight equal one.

4. Repeat steps 1 to 3 until there is no significant change in weights. In step 1, peaks are scored with the optimized weights from the previous round.
5. Apply the converged weights to Equation (2.1) and denoise the spectra in the experimental dataset (this stage is previously described in 2.2.1).

2.3 Experimental results and discussion

Experiments were conducted on two MS/MS spectrum datasets: ISB with low resolution and TOV-Q with high resolution. To illustrate the performance of the proposed method,

the Mascot search results from the raw datasets, the same datasets denoised by the Ding et al.'s method (2009a) and the same dataset denoised by the proposed method in this study were compared.

2.3.1 Datasets

The following is a brief description of the two datasets used in the proposed method. These two datasets were chosen for comparing with Ding et al's work (2009a).

(1) **ISB dataset**: The spectra with low resolution in this dataset were acquired from 18 control mixture protein complexes that were analyzed by mLC-MS on an ESI-ITMS (ThermoFinnigan, San Jose, CA) using a standard top-down data-dependent ion selection approach (Keller et al., 2002).

(2) **TOV-Q dataset:** This dataset consisted of high-resolution MS/MS spectra that were acquired on a QSTAR Plusar (MDS Sciex Corp.) in the Eastern Quebec Proteomic Center in Laval University Medical Research Center in Laval, Quebec, Canada (Zou et al., 2010). The samples analyzed were generated by the tryptic digestion of a whole-cell lysate from 36 fractions of TOV-112 samples (Gagné et al., 2005).

2.3.2 Search engine

Experiments were conducted by using an on-line version of the Mascot search engine

(http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=MIS). The on-line version has a limitation on the size (20 MB) and the number of spectra (1200 groups) of input files. The raw (before denoising method applied) spectra, the spectra denoised by Ding et al.'s approach in (2009a) and the spectra denoised by the proposed approach in this study were used for search with the same parameters. The parameters used for the ISB (TOV-Q) dataset are given in Table 2.1.

Table 2.1 Parameters of the Mascot search engine for ISB (TOV-Q) dataset.

Database	NCBInr
Enzyme	trypsin
Fixed modifications	carbamidomethyl (C)
Variable modifications	oxidation (M) [oxidation (M), deamidated (NQ)]
Peptide charges	+2, +3
Mass values	monoisotopic
Protein	unrestricted
Peptide mass tolerance	± 2Da (± 0.2Da)
Fragment mass tolerance	± 0.8Da (± 0.2Da)
Max. missed cleavages	1
Isotope error mode	1 (0)
Quantitation	none
Taxonomy	all entries

2.3.3 Denoising program

The proposed approach was implemented in Matlab R2008b. The denoising program was run on a PC with 1.6 GHz Dual CPU (Windows XP operating system).

2.3.4 Results and discussion

Due to the limitations of the on-line Mascot, the input file was separated by 1200 spectra per file. The results of denoising with Ding et al.'s method, proposed method and original data are listed in Table 2.2.

Table 2.2 Results of the denoising algorithm.

Datasets	Mean peaks	Identified
ISB		
Raw	152	586
Ding	49	944
Denoised*	52	1021
TOV-Q		
Raw	67	1773
Ding	23	1626
Denoised*	24	2040

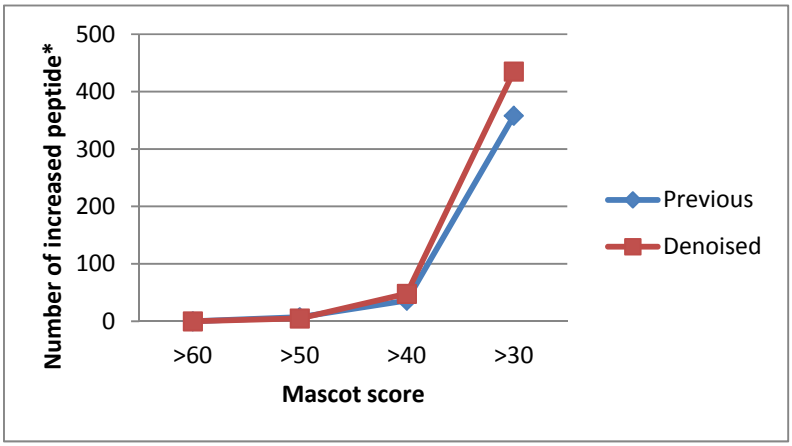
* Denoised: the proposed method

In Table 2.2, the "Raw" spectra are the original spectra before denoising method applied and the "Ding" spectra are the denosied spectra with Ding et al.'s method (2009a) while the "Denoised" spectra are the denoised spectra with the proposed approach. "Mean peaks" indicates the mean of the number of peaks of spectra in the dataset; "Identified" is the number of peptides whose ion scores are greater than or equal to the Mascot identity threshold (given the same false discovery rate of 5%).

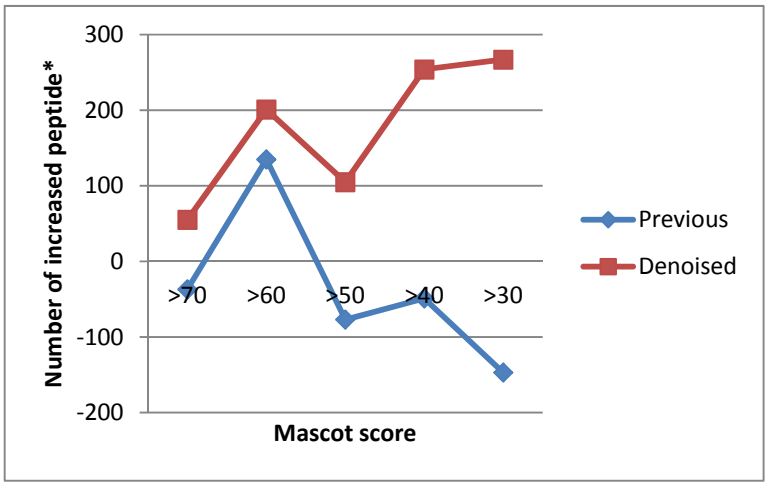
As shown in the second column of Table 2.2, the proposed denoising algorithm removed 66% [=(152−52)/152] of peaks from spectra from the ISB dataset, and 64% [=(67−24)/67] of peaks from spectra from the TOV-Q dataset. By comparison with the Ding et al.'s approach (2009a), which removed about 68% and 66% of peaks, this new approach retained about 2% more peaks in the spectra.

As shown in the third column of Table 2.2, the number of identified peptides increased by 74% [=(1021−586)/586] for the spectra of the ISB dataset after applying the proposed approach, while with the Ding et al.'s approach, the increase was only about 61% [=(944−586)/586]. This implies that the proposed approach can achieved roughly 14% improvement on the low-resolution dataset over Ding et al.'s approach. For the high-resolution dataset TOV-Q, the Ding et al.'s approach did not work well. It identified 8% [=(1626−1773)/1773] fewer peptides than the raw spectra. Applying the proposed approach, the number of identified peptides increased by 15% [=(2040−1773)/1773]. The

increases in both the number of peaks and the identified peptides imply that not all the peaks removed by the Ding et al.'s approach were noise. In addition, the proposed approach gives more reliable denoised spectra than Ding et al.'s for both high- and low-resolution datasets.



(a)



(b)

* Comparisons are made against results from raw spectra without denoising

Figure 2.2 Comparison of the numbers of identified spectra by Ding et al.'s method ("Previous") and the proposed method ("Denoised") over various peptide identification score thresholds for the ISB dataset (a) and TOV-Q dataset (b).

Figure 2.2 shows the increased numbers of identified peptides with the Ding et al.'s

approach and the proposed approach over various peptide identification (Mascot) scores thresholds, compared with the results from the raw spectra. In Figure 2.2(a), there is no significant difference between the numbers of peptide spectra (whose Mascot ion scores are greater than 50) with and without application of the two denoising approaches to the ISB dataset. However, the numbers of spectra whose Mascot ion scores are greater than 30 by application of the two denoising methods are significantly larger than that without denoising methods (using raw spectra). Furthermore, under the same Mascot ion scores (greater than 30), dataset denoised by proposed method could be identified more peptides than that by Ding et al.'s method.

Generally, the less noise a spectrum has, the higher its quality. Therefore, the proposed method can significantly improve the quality of low-resolution spectra, especially when its original quality is poor. From Figure 2.2(b), the number of increased spectra under different cut-off value (from 30 to 70) after applying the proposed method is always significantly greater than that after applying Ding et al.'s method. This indicates that the proposed method significantly outperforms Ding et al.'s method on high-resolution dataset.

Combining Table 2.2 and Figure 2.2, one can find that the improvement achieved by this new approach is larger than that of Ding et al.'s method for both datasets. However, one can also see that the improvements of both methods on the TOV-Q dataset are not as

significant as those on the ISB dataset. One explanation is that TOV-Q dataset is high-resolution spectra, which means that there are more signal peaks in a spectrum (higher percentage of signal peaks) while noise peaks are fewer. Another reason could be the nature of the morphological reconstruction filter. This filter, which can remove at least 50% of peaks in spectra by choosing the local maxima, may not fit as well for the high-resolution data as it does for the low-resolution data.

Table 2.3 Adapted weights for different datasets.

	Ding	ISB	TOV-Q
w_1	1	1	1
w_2	1	0.2283	0.33016
w_3	0.2	0.0222	0.51198
w_4	0.2	0.3019	0.00036
w_5	0.5	0.9573	0.36609

Table 2.3 shows weights estimated with the new approach for the two datasets compared to the fixed weights in Ding et al.'s method. In Table 2.3, "Ding" represents the fixed weights, "ISB" represents the weights estimated for the ISB dataset and "TOV-Q" represents the weights estimated for the TOV-Q dataset. Due to the nature of peak selection, relationships among the weights within one set are more important than their absolute values. For example, from Table 2.3, the converged weights in columns 2 and 3

are quite different from the initial weights in column 1. This implies that what was used in the previous work (Ding et al., 2009a) did not reflect reality. In addition, weights for low-resolution datasets (column 2) and those for high-resolution datasets also are very different, except for w_1. For both datasets, w_1 is quite high. This implies that mass difference of ions is an important feature for both low-resolution datasets and high-resolution datasets.

2.4 Conclusions

In this chapter, an adaptive denoising approach was proposed. This new approach first adjusts the intensities of spectra by scores calculated with given weights and then selects the peaks of signals based on their adjusted intensities. Unlike others' work, for example, Ding et al. (2009a) where the weights are fixed and empirically assigned, this new approach updated the weights for different datasets. In this way, the scores can better separate signals from noise. By applying this new approach, about 66% of the noise peaks among a spectrum can be detected. By applying the peptide identification program (Mascot), the number of peptides identified increased by 74% and 15% for the spectra in the ISB dataset and the TOV-Q dataset, respectively. The experimental results imply that the adaptive weights could achieve better performance on both high-resolution and low-resolution MSMS spectra comparing to Ding et al. (2009a).

Chapter 3. Quality assessment of MS/MS spectrum

3.1 Introduction

In this chapter, an unsupervised machine learning method is presented with a set of 10 most relevant features from Ding's work (2009) to assess the quality of MS/MS spectra. Section 3.2 gives the description of the 10 features and explains the new method that makes use of them. Section 3.3 discusses the experimental results using two MS/MS datasets. Conclusions are given in Section 3.4.

3.2 Method

In this section, first the 10 features for the quality assessment of MS/MS spectra are introduced. Then a graph-based consensus optimization method (Ge et al., 2011) is described that is used to integrate individual assessments into a consensus assessment. An algorithm to solve this optimization problem is proposed. The convergence of the algorithm is also proved.

3.2.1 Spectral features

A MS/MS spectrum usually contains tens to hundreds of *m/z* values with their corresponding intensities. In the literature, hundreds of features have been proposed to describe the quality of MS/MS spectra (Wu et al., 2006; Flikka et al., 2006; Wong et al.,

2007). In one project, after removing the noise peaks (Vincent, 1993; Ding et al., 2009a), the 10 most relevant spectral features were selected based on support vector machine methods (Ding et al., 2011; Ding, 2009). The details of these features are described below.

Feature 1 was proposed by Bern et al. (2004), and was defined as the total normalized intensity of pairs of peaks with their *m/z* values added up to the *m/z* of the precursor ion (such pairs of peaks are called complementary peaks). This feature is based on the assumption that the peaks with lower intensities are noise and that the complementary peaks are more likely to be signal.

Feature 2 was proposed by Flikka et al. (2006), and was defined as the mass of the uncharged precursor ions. This feature is based on the observation that most of the low-quality spectra have small masses of precursor ions because they may come from short peptides that cannot generate enough fragment ions for identification or come from irrelevant chemical molecules like trypsin.

Feature 3 was proposed by Wu et al. (2008), and was defined as the number of peaks whose mass difference is equal to the mass of one of the 20 amino acids. Note that all peaks were considered as single-charged in this method. The feature is measured with the error tolerance (in *m/z*) of 0.5 Da. This reflects the fact that a peptide is a chain of amino

acid.

Feature 4 was proposed by Flikka et al. (2006), and was defined as the average delta mass (i.e., the average of all mass differences between any two neighbor peaks) in a spectrum. This feature reflects that the too-dense spectra are typically of low quality (Bern et al., 2004; Flikka et al., 2006; Xu et al., 2005).

Feature 5 was proposed by Bern et al. (2004) and called the Good-Diff Fraction, and was defined as

$$GoodDiffs = \sum\{NormI\,(x) + NormI\,(y) | M\,(x) - M\,(y) \approx M_i \text{ for some } i = 1, 2, \ldots, 20\} \quad (3.1)$$

where $M(x)$ is the *m/z* value of peak x and $M_1, M_2, \ldots, M_{20}$ represent the masses of 20 amino acids (not all of which are unique). The feature is measured with the error tolerance (in *m/z*) of 0.5 Da. Similar to Feature 3, this feature reflects how likely two peaks differ by the mass of an amino acid.

Feature 6 was proposed by Wu et al. (2008), and was defined as the number of pairs of complementary peaks (note: all peaks are considered as single-charged). This feature reflects how likely an N-terminus ion and a C-terminus ion in a spectrum are produced as peptide fragments from the same peptide bond.

Feature 7 was proposed by Wu et al. (2008), and was defined as the number of pairs of peaks whose *m/z* value differences are equal to the mass of either a water molecule or an ammonia molecule (note: all peaks are considered as single-charged). This feature reflects how likely one fragment ion in a spectrum is produced by losing either a molecule of water or ammonia from the b or y ion.

Feature 8 was proposed by Wong et al. (2007), and was defined as the ratio of the number of peaks that have a relative intensity greater than 1% of the total intensity to the total number of peaks. The rationale for this feature is similar to that for Feature 1.

Feature 9 was proposed by Flikka et al. (2006), and was defined as the standard deviation of delta mass (i.e., all mass differences between any two neighbor peaks) values in a spectrum. The rationale for this feature is similar to that for Feature 4.

Feature 10 was proposed by Wu et al. (2008), and was defined as the number of pairs of peaks whose *m/z* value difference is equal to the mass of a CO group or an NH group (note: all peaks are considered as single-charged). This feature reflects how likely one fragment ion is a-ion or z-ion.

From the definitions and physical meanings of the above 10 features, the larger the values of these features, the more likely the spectra are of high quality. However, these features

can never become unique markers of peptides. One cannot determine spectrum quality be using one of these feature alone. Therefore, it might improve the accuracy of quality assessment by integrating or combining these features (Ding et al., 2011; Ding, 2009).

3.2.2 Quality assessments by integration of the features

This thesis only considers two classes of MS/MS spectrum quality: low (Class 1) and high (Class 2). Suppose there are *m* features. Each feature generates two quality classes (high and low) according to the feature values and thresholds. For the convenience of discussion, features are ordered and the groups of all the features are labeled in sequence such as t1, t2, . . . , t*v* (*v=2m*). Each spectrum corresponds to *m* groups. As such, a bipartite graph forms; see Figure 3.1, where *si*: *i*-th spectrum; *ti*: *i*-th feature group. The mapping of Class 1 or Class 2 of the *i*-th feature (*Fi*) is as follows: *2i-1=j* (t*j*) for Class 1 and *2i=j* (t*j*) for Class 2. For example, for the 3rd feature with Class 1, *j*=2(3)-1=5 (i.e., *t5*).

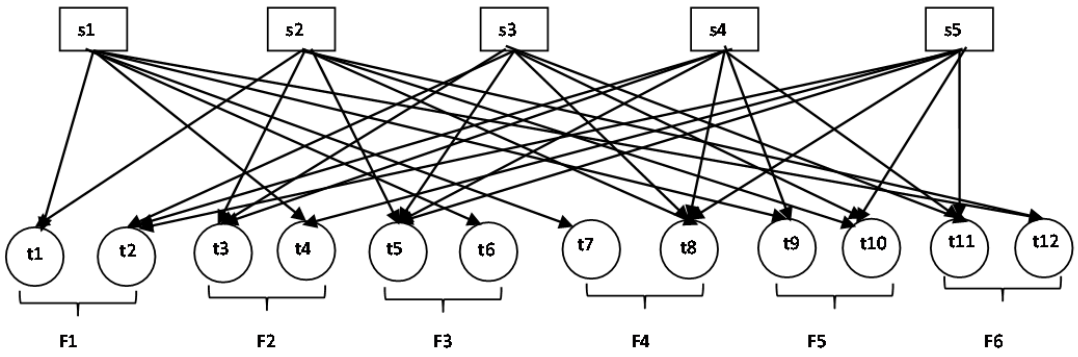


Figure 3.1 Example of a bipartite graph

The following is an example to further illustrate the notations. Suppose there are spectra {s1, s2, s3, s4, s5} and six features are used to classify the spectra into two classes (1&2). Suppose the data are as shown in Table 3.1, where F*i*: *i*-th feature; s*j*: *j*-th spectrum; the number 1, 2: Class 1 and Class 2. Take s3 as an example. It corresponds to {t2, t3, t5, t8, t10, t11} or {(F1, Class 2), (F2, Class 1), (F3, Class 1), (F4, Class 2), (F5, Class 2), (F6, Class 1)}.

Table 3.1 An object pool classified into several groups.

Spectrum / Feature	s1	s2	s3	s4	s5
F1	1	1	2	2	2
F2	1	1	1	2	2
F3	2	1	1	1	2
F4	1	2	2	2	2
F5	1	2	2	1	2
F6	2	1	1	2	2

The proposed method considers the probabilities of s*i* (i=1, . . . , *n*) to be Class 1 and Class 2. For spectra s*i* (*i*=1, . . . , *n*), the probability can be represented by a matrix $U_{n\times 2}$. Further, a matrix $Q_{v\times 2}$ is defined for the probabilities of t*j* (*j*=1, . . . , *v*) to be Class 1 and Class 2. For example, $Q(1,2)$ is the probability of feature group t1 belongs to Class 2. Note that sum of the probability of *tj* to be Class 1 and Class 2 ($Q(j,1)+Q(j,2)$) is always 1.

Then, u_{iz} = *Prob*(s*i* is in class *z*) and q_{jz} = *Prob*(t*j* is class *z*), where *z*=1 (Class 1) or *z*=2 (Class 2).

Generally, a feature group t*j* corresponds to class *z* if the majority of spectra in the group belong to class *z*; meanwhile, a spectrum belongs to class *z* if the majority of the groups it belongs to correspond to class *z*. Furthermore, the initial class labels for the groups can be denoted by matrix $Y_{v\times 2}$, in which y_{jz} = 1 if the group t*j* corresponds to class *z* and 0 otherwise. To estimate the probabilities in matrix *U*, the following cost function with constraints needs to be optimized (Ge et al., 2011):

$$\begin{aligned} &\min J(U,Q) = \min\left[\sum_{z=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{v} a_{ij}(u_{iz}-q_{jz})^2 + \alpha\sum_{z=1}^{k}\sum_{j=1}^{n}(q_{jz}-y_{jz})^2\right] \\ &s.t.\quad \sum_{z=1}^{k} u_{iz} = 1,\ \sum_{z=1}^{k} q_{jz} = 1 \\ &u_{iz}\in[0,1],\ q_{iz}\in[0,1] \end{aligned} \tag{3.2}$$

where a_{ij} is the (*i*, *j*) element of affinity matrix $A_{n\times v}$ of the bipartite graph. a_{ij} = 1 if spectrum s_i is assigned to the group t_j, and 0 otherwise. α is a positive parameter that expresses the confidence of the initial labels of the group nodes. This helps to avoid over-fitting. *k=2* is the number of consensus groups (with either high quality or low quality spectra). As each spectrum belongs to one of the *k* groups by each of *m* features, then

$$\sum_{j=1}^{v} a_{ij} = m \tag{3.3}$$

It is obvious that the value of the cost function is zero if all assessments based *m*

individual features agree perfectly. However, in practice this does not happen. Therefore, the desired resultant matrix $Q'_{v\times k}$ is obtained when the cost function in the constraint optimization problem (3.2) reaches its minimal value. Finally, every spectrum will be assigned with a probability to class *z* directly according to the values in matrix $U'_{n\times k}$.

From the constraint optimization problem (3.2), for the given matrix *U*, the objective function is quadratic in elements of matrix *Q*. For the given matrix *Q*, the objective function is quadratic in elements of matrix *U*. Therefore, the following iterative algorithm is used to solve this optimization problem.

Step 1: Initialize *Q* by *Y*, that is, $Q^t=Y$, and $t=0$.

Step 2: $t=t+1$

Estimate U^t by solving

$$\min_U J(U,Q^{t-1}) = \min_U \left[\sum_{z=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{v} a_{ij}(u_{iz} - q_{jz}^{t-1})^2 + \alpha \sum_{z=1}^{k}\sum_{j=1}^{n}(q_{jz}^{t-1} - y_{jz})^2 \right]$$

to obtain

$$u_{iz}^{t} = \frac{\sum_{j=1}^{v} a_{ij} q_{jz}^{t-1}}{\sum_{j=1}^{v} a_{ij}} = \frac{1}{m}\sum_{j=1}^{v} a_{ij} q_{jz}^{t-1} \tag{3.4}$$

Estimate Q^t by solving

$$\min_{Q} J(U^t, Q) = \min_{Q} \left[\sum_{z=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{v} a_{ij} (u_{iz}^t - q_{jz})^2 + \alpha \sum_{z=1}^{k} \sum_{j=1}^{n} (q_{jz} - y_{jz})^2 \right]$$

to obtain

$$q_{jz}^t = \frac{\sum_{i=1}^{n} a_{ij} u_{iz}^t + \alpha y_{jz}}{\alpha + \sum_{i=1}^{n} a_{ij}} \tag{3.5}$$

Step 3: Stop if $\left\| U^t - U^{t-1} \right\| \leq \varepsilon$ and output U, where ε is a user-specified small positive number.

In the above algorithm, the constraints in optimization problem (3.2) are not included. However, if the initial class labels for the groups $Y_{v \times k}$ satisfy that

$$\sum_{z=1}^{k} y_{jz} = 1, \ \ y_{jz} \in [0,1] \tag{3.6}$$

Then the solutions of the above algorithm at every iteration t will satisfy all constraints in optimization problem (3.2). This was conjecture proven in a previous paper (Lin et al., 2012).

The algorithm reflects that at each iteration the probability estimation of group node Q receives information from its neighboring spectral nodes while not deviating too wildly from its initial value Y. In return, the updated probability estimates of group nodes propagate information back to their neighboring spectral nodes. The propagation stops when the process converges. The process converges to a stationary point.

3.3 Experimental results and discussion

To evaluate the proposed method, experiments were conducted on two MS/MS spectrum datasets, namely TOV and ISB. The ISB dataset was introduced in the previous chapter. The following is a brief description of the TOV datasets.

TOV dataset: The MS/MS spectra were acquired from a LCQ DECA XP ion trap spectrometer (ThermoElectron Corp.) (Wu et al., 2006). The samples analyzed were generated by the tryptic digestion of a whole-cell lysate from the TOV-112 sample (Gagné et al., 2005). The number of spectra in this dataset is 22576, and these spectra are sequenced using SEQUEST against human protein database (ipi.HUMAN.v3.42.fasta) containing 72340 protein sequences and 5 contaminant sequences.

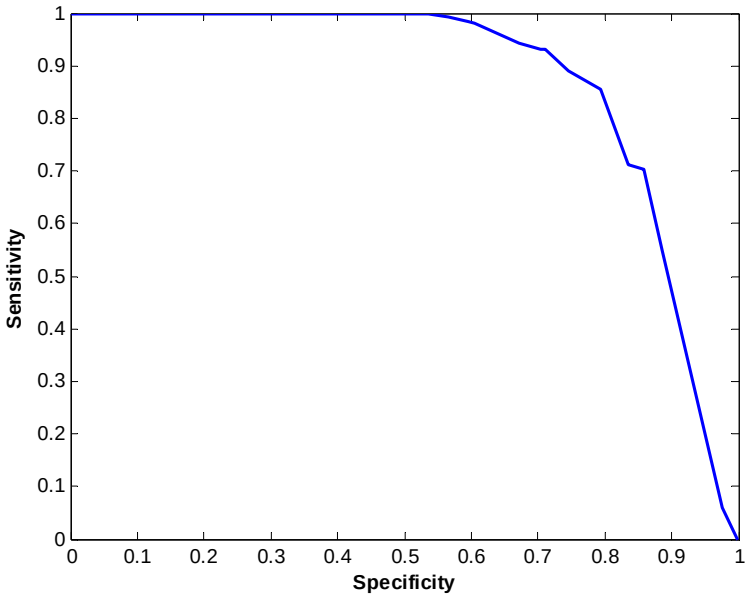
The distribution of tandem spectra is shown in Table 3.2. H represents the number of high-quality spectra and L represents the number of low-quality spectra. The assignments of spectra were determined by SEQUEST score, with the cut-off score at 2.8. Spectra with scores of less than threshold were labeled as low-quality spectra; otherwise, they were labeled as high-quality spectra.

Table 3.2 Distribution of multiply-charged spectra in the ISB and TOV datasets

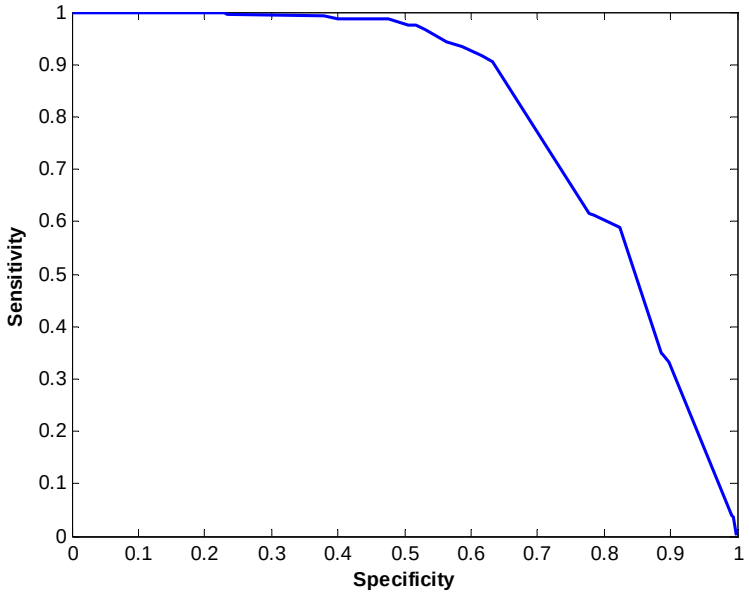
	H	L	Total
TOV	1136	21440	22576
ISB	1047	35997	37044

In the experiment, the proposed method is applied to both datasets in order to obtain assessments based on individual features. For each feature, spectra with the top 50% feature values are assigned to the high-quality class. In the method, the parameter α in the model was empirically taken as 90.

Figure 3.2 shows the ROC curves for the consensus classifiers for the TOV and ISB datasets, respectively. For the TOV dataset, the proposed method eliminates about 74% of the low-quality spectra while in the best case losing less than 9% of the high-quality spectra. For the ISB dataset, the proposed method filters out about 63% of the low-quality spectra while losing only 10% of the high-quality spectra. By removing the same amount of low-quality spectra, Ding et al.'s methods (2009b) lose 19% and 17% of high-quality spectra respectively on both of the datasets. If searching just the TOV and ISB spectra in the high-quality group with SEQUEST, about 56% (=1−10042/22576) and 62% (=1−14087/37044) of searching time can be saved while about 10% of the interpretable spectra is lost. These results indicate that the proposed method outperforms the method in Ding et al. (2009b).



(a)



(b)

Figure 3.2 ROC curve for the proposed classifier for TOV dataset (a) and ISB dataset (b). Sensitivity (also called the true positive rate) measures the proportion of actual positives that are correctly identified as such; specificity (also called the true negative rate) measures the proportion of negatives that are correctly identified as such.

Furthermore, this method achieved a better result from the TOV dataset than from the ISB dataset. This may be because there were more low-quality spectra in the ISB dataset (35997/37044=97%) than in the TOV dataset (21440/22576=95%). A high percentage of low-quality spectra make quality assessment challenging (Ding, 2009). Another reason for the better result might be that there are more triply-charged spectra in the ISB dataset (18044) than in the TOV (9732). MS/MS spectra of triply-charged ions contain more doubly-charged peaks than both doubly- and singly-charged spectra. The quality of triply-charged spectra are not well described by the 10 features used in this method, especially because features 3, 6, 7, and 10 are designed only for singly-charged peaks while triply-charged spectra produce a high number of doubly-charged peaks (Zou et al., 2010; Shi et al., 2011).

3.4 Conclusions

In this chapter, an unsupervised machine learning method was presented that integrates assessments based on individual features (which is easy to do with a low precision) into a consensus assessment with high precision. This unsupervised machine learning method estimates the probability of a spectrum being high-quality from the assessments based on individual features. The estimation of the probabilities is solved through a constraint optimization problem. Experimental results illustrated that if searching a database using only spectra assessed as high quality in TOV and ISB, about 56% and 62% of SEQUEST

searching time can be saved with only 9% and 10% of high-quality spectra lost, respectively. This result indicates that the proposed method is useful in saving database searching time. Further, at a sensitivity of 90%, this method reaches specificities of 74% and 63%, respectively, which surpasses the existing method (Ding et al., 2009b). This indicates that this unsupervised machine learning method could adaptively integrate all assessments from 10 individual features into a consensus quality assessment with higher precision on MS/MS spectra. Also, this result shows the way in which the conditional probability being estimated is effective.

Chapter 4. Conclusions and future work

4.1 Overview and Conclusions

Peptide sequencing from MS/MS is important in proteomics. A challenge in peptide sequencing is noise in MS/MS spectra. It leads to incorrect peptide identification. To meet this challenge, filtering methods are proposed to remove noise and screen out low quality MS/MS spectra. Most existing methods are based on the supervised machine learning techniques. These techniques require so-called training data which are essentially a set of pairs of attributes (i.e., features) and labels (i.e., 'signal & noise', or 'high quality & low quality') from peptide identification result. Such data may not be available because the filtering methods are applied before peptide identification. Therefore, unsupervised learning methods have been used in this work for peptide identification. Generally, these methods do not require label information but attributes only.

Following up on the approach developed by Ding et al. (2009a), an unsupervised learning approach was used in the present study to remove noise peaks in MS/MS spectra. The general idea in Ding et al.'s approach was to score each peak with a set of features that describe the peak. The score was thus an aggregate of the features with weights. These weights were determined empirically in Ding et al.'s approach. The present study developed a method to adjust the weights.

Details of the method proposed by this study were documented in Chapter 2. By applying the proposed approach on two MS/MS spectra datasets, its superior performance was illustrated. About 66% of noise peaks can be removed and that the number of peptides identified by peptide identification was increased by 14% and 23% for the ISB and the TOV-Q datasets, respectively, compared to the number identified by Ding et al.'s method (2009a).

A similar idea with the noise removal (denoising) approach was proposed for screening out low quality MS/MS spectra (quality assessment). This was documented in Chapter 3. In particular, the proposed method estimates the probabilities of spectra being of high quality based on a set of pre-defined features. The probabilities were estimated through a constraint optimization technique. Experimental results on the ISB and TOV datasets demonstrate that by searching the high-quality tandem spectra determined by the proposed method, the majority of database searching time (56% and 62%) can be saved while only 10% of high-quality spectra are lost.

4.2 Contribution

The main contribution of the above two methods developed by this study is a new technology for removal of noise in MS/MS spectra and screening out of low quality

MS/MS spectra. The effectiveness of the new technology is very high for both tasks in peptide identification. In the field of information fusion, the new technology has a high potential to be effective.

4.3 Future work

In this thesis, an adaptive denoising approach and an adaptive quality assessment method, based on unsupervised learning techniques are presented, and they have been shown to improve the peptide identification process. However, there remain some further possible improvements that are considered as future work.

First, in the proposed denoising method, the weights estimated by LDA might be negative, which does not make sense. In the future, a constraint LDA may be used to ensure that the estimated weights are positive. In addition, by its nature, the morphological reconstruction filter removes at least roughly 50% of peaks in the spectra, so for spectra with less than 50% noise peaks, the proposed method may remove some signal peaks. Therefore, a more sophisticated filter should be designed for optimal peak selection.

Second, in the proposed quality assessment method, more complicated and comprehensive features may be considered. For example, in the 10 features considered in this study, 4 features are calculated for singly-charged peaks. This could make the

classification method less effective on the triply- or higher-charged spectra. In the future, different features may be considered for different charges of spectra. Further, at present, the α and cut-off values for individual features were chosen based on several trial and error repeats. In the future, a more objective method may be developed for choosing these values. Finally, the proposed constraint optimization model may be applied to other unsupervised classification problems in bioinformatics and proteomics.

Apart from the two proposed methods, other closely related topics may also be considered as a future work. For example, spectra clustering, which synthesizes all redundant spectra from the same peptide, is an effective strategy for acquiring useful information as well as removing noise from MS/MS spectra. Spectra clustering can significantly reduce the analysis time since the clustering algorithm is usually much faster than peptide sequencing methods (Beer et al., 2004; Frank et al., 2007; Flikka et al., 2007; Falkner et al., 2008). Indeed, a good clustering algorithm has the potential to improve the peptide identification process since it synthesizes all useful information from every single spectrum in the cluster.

References

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, *422*(6928), 198-207.

Anderson, N. L., & Anderson, N. G. (1998). Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, *19*(11), 1853-1861.

Baginsky, S., Grossmann, J., Roos, F. F., Cieliebak, M., Lipták, Z., Mathis, L. K., & Müller, M. (2005). AUDENS: a tool for automated peptide de novo sequencing. *Journal of Proteome Research*, *4*(5), 1768-1774.

Bern, M., Goldberg, D., Mcdonald, W. H., & Yates, J. R. (2004). Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*, *20*(suppl 1), i49-i54.

Choo, K. W., & Tham, W. M. (2007). Tandem mass spectrometry data quality assessment by self-convolution. *BMC Bioinformatics*, *8*(1), 352.

Ding, J. (2009). *Preprocessing of tandem mass spectra using machine learning methods* (Master's thesis). University of Saskatchewan, Saskatoon, SK, Canada.

Ding, J., Shi, J., & Wu, F. X. (2009a). A novel approach to denoising ion trap tandem mass spectra. *Proteome Science*, *7*(9).

Ding, J., Shi, J., & Wu, F. X. (2009b). Quality assessment of tandem mass spectra by using a weighted k-means. *Clinical Proteomics*, *5*(1), 15-22.

Ding, J., Shi, J., & Wu, F. X. (2011). SVM-RFE based feature selection for tandem mass spectrum quality assessment. *International Journal of Data Mining and Bioinformatics*, *5*(1), 73-88.

Ding, J., Shi, J., Zou, A. M., & Wu, F. X. (2008, November). Feature selection for tandem mass spectrum quality assessment. In *IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2008)*, (pp. 310-313). IEEE.

Eidhammer, I., Flikka, K., Martens, L., & Mikalsen, S. O. (2007). *Computational Methods for Mass Spectrometry Proteomics* (pp. i-xi). John Wiley & Sons, Ltd.

Eng, J. K., McCormack, A. L., & Yates Iii, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11), 976-989.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179-188.

Flikka, K., Martens, L., Vandekerckhove, J., Gevaert, K., & Eidhammer, I. (2006). Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics*, *6*(7), 2086-2094.

Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., & Pevzner, P. A. (2007). Clustering millions of tandem mass spectra. *Journal of Proteome Research*, *7*(01), 113-122.

Gagné, J. P., Gagné, P., Hunter, J. M., Bonicalzi, M. È., Lemay, J. F., Kelly, I., ... & Poirier, G. G. (2005). Proteome profiling of human epithelial ovarian cancer cell line TOV-112D. *Molecular and Cellular Biochemistry*, *275*(1-2), 25-55.

Gentzel, M., Köcher, T., Ponnusamy, S., & Wilm, M. (2003). Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics*,

3(8), 1597-1610.

Grossmann, J., Roos, F. F., Cieliebak, M., Lipták, Z., Mathis, L. K., Müller, M., ... & Baginsky, S. (2005). AUDENS: a tool for automated peptide de novo sequencing. *Journal of Proteome Research*, *4*(5), 1768-1774.

Hernandez, P., Müller, M., & Appel, R. D. (2006). Automated protein identification by tandem mass spectrometry: issues and strategies. *Mass Spectrometry Reviews*, *25*(2), 235-254.

Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S., Goodlett, D. R., & Kolker, E. (2002). Experimental protein mixture for validating tandem mass spectral analysis. *OMICS: A Journal of Integrative Biology*, *6*(2), 207-212.

Klammer, A. A., Wu, C. C., MacCoss, M. J., & Noble, W. S. (2005, August). Peptide charge state determination for low-resolution tandem mass spectra. In *Computational Systems Bioinformatics Conference 2005.* (pp. 175-185). IEEE.

Koenig, T., Menze, B. H., Kirchner, M., Monigatti, F., Parker, K. C., Patterson, T., & Steen, H. (2008). Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. *Journal of Proteome Research*, *7*(9), 3708-3717.

Lin, W., Wang, J., Zhang, W. J., & Wu, F. X. (2012). An unsupervised machine learning method for assessing quality of tandem mass spectra. *Proteome Science*, *10*(Suppl 1), S12.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., & Lajoie, G. (2003).

PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, *17*(20), 2337-2342.

Maton, A. (1993). *Human Biology and Health*. Englewood Cliffs, N.J: Prentice Hall.

Mujezinovic, N., Raidl, G., Hutchins, J. R., Peters, J. M., Mechtler, K., & Eisenhaber, F. (2006). Cleaning of raw peptide MS/MS spectra: Improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. *Proteomics*, *6*(19), 5117-5131.

Na, S., & Paek, E. (2006). Quality assessment of tandem mass spectra based on cumulative intensity normalization. *Journal of Proteome Research*, *5*(12), 3241-3248.

Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*, *73*(11), 2092-2123.

Ning, K., & Leong, H. W. (2007). Algorithm for peptide sequencing by tandem mass spectrometry based on better preprocessing and anti-symmetric computational model. In *Computational Systems Bioinformatics Conference* (Vol. 6, pp. 19-30). IEEE.

Perkins D. N., Pappin D .J., Creasy D .M., & Cottrell J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, *20*(18), 3551-3567.

Rejtar, T., Chen, H. S., Andreev, V., Moskovets, E., & Karger, B. L. (2004). Increased

identification of peptides by enhanced data processing of high-resolution MALDI TOF/TOF mass spectra prior to database searching. *Analytical Chemistry*, *76*(20), 6017-6028.

Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Jonscher, K. R., Pierce, K. G., ... & Ahn, N. G. (2004). Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Analytical Chemistry*, *76*(13), 3556-3568.

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

Salmi, J., Moulder, R., Filén, J. J., Nevalainen, O. S., Nyman, T. A., Lahesmaa, R., & Aittokallio, T. (2006). Quality classification of tandem mass spectrometry data. *Bioinformatics*, *22*(4), 400-406.

Salmi, J., Nyman, T. A., Nevalainen, O. S., & Aittokallio, T. (2009). Filtering strategies for improving protein identification in high-throughput MS/MS studies. *Proteomics*, *9*(4), 848-860.

Tabb, D. L., Shah, M. B., Strader, M. B., Connelly, H. M., Hettich, R. L., & Hurst, G. B. (2006). Determination of peptide and protein ion charge states by Fourier transformation of isotope-resolved mass spectra. *Journal of the American Society for Mass Spectrometry*, *17*(7), 903-915.

Tabb, D. L., Thompson, M. R., Khalsa-Moyers, G., VerBerkmoes, N. C., & McDonald, W.

H. (2005). MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *Journal of the American Society for Mass Spectrometry*, *16*(8), 1250-1261.

Vincent, L. (1993). Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms*, IEEE Transactions on Image Processing*, *2*(2), 176-201.

Wu, F. X., Gagne, P., Droit, A., & Poirier, G. (2006). RT-PSM, a real-time program for peptide-spectrum matching with statistical significance. *Rapid Communications in Mass Spectrometry*, *20*(8), 1199-1208.

Wu, F. X., Gagné, P., Droit, A., & Poirier, G. (2008). Quality assessment of peptide tandem mass spectra. *BMC bioinformatics*, *9*(Suppl 6), S13.

Yates, J. R. (1998). Database searching using mass spectrometry data. *Electrophoresis*, *19*(6), 893-900.

Zhang, J., He, S., Ling, C. X., Cao, X., Zeng, R., & Gao, W. (2008). PeakSelect: preprocessing tandem mass spectra for better peptide identification. *Rapid Communications in Mass Spectrometry*, *22*(8), 1203-1212.

Zou A. M., Shi J. H., Ding J. R., and Wu F. X. (2010). Charge state determination of peptide tandem mass spectra using support vector machine (SVM). *IEEE Transaction on Information Technology in Biomedicine, 14*(3), 552-558.

Publications

Journal

Lin, W., Wang, J., Zhang, W. J., & Wu, F. X. (2012). An unsupervised machine learning method for assessing quality of tandem mass spectra. *Proteome science*, *10*(Suppl 1), S12.

Yuan, Z., Shi, J., **Lin, W.**, Chen, B., & Wu, F. X. (2012). Features-Based Deisotoping Method for Tandem Mass Spectra. *Advances in bioinformatics*, *2011*.

Lin, W., Wu, F. X., Shi, J., Ding, J., & Zhang, W. (2011). An adaptive approach to denoising tandem mass spectra. *Proteomics*, *11*(19), 3773-3778.

Conference

Shi, J., **Lin, W.**, & Wu, F. X. (2010, June). Statistical analysis of Mascot peptide identification with active logistic regression. In *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on* (pp. 1-4). IEEE.

Copyright permissions

The copyright of the following papers:

Lin, W., Wang, J., Zhang, W. J., & Wu, F. X. (2012). An unsupervised machine learning method for assessing quality of tandem mass spectra. Proteome science, 10(Suppl 1), S12.

Lin, W., Wu, F. X., Shi, J., Ding, J., & Zhang, W. (2011). An adaptive approach to denoising tandem mass spectra. Proteomics, 11(19), 3773-3778.

are included in the following pages.

BioMed Central Open Access license agreement

Brief summary of the agreement

Anyone is free:

- to copy, distribute, and display the work;
- to make derivative works;
- to make commercial use of the work;

Under the following conditions: Attribution

- the original author must be given credit;
- for any reuse or distribution, it must be made clear to others what the license terms of this work are;
- any of these conditions can be waived if the authors gives permission.

Statutory fair use and other rights are in no way affected by the above.

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS
Sep 16, 2013

This is a License Agreement between wenjun lin ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3230890659964
License date	Sep 16, 2013
Licensed content publisher	John Wiley and Sons
Licensed content publication	Proteomics
Licensed content title	An adaptive approach to denoising tandem mass spectra
Licensed copyright line	Copyright © 2011 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
Licensed content author	Wenjun Lin,Fang-Xiang Wu,Jinhong Shi,Jiarui Ding,Wenjun Zhang
Licensed content date	Sep 7, 2011
Start page	3773
End page	3778
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Total	0.00 USD