대량의 프로테옴 데이타를 효과적으로 해석하기 위한 기계학습 기반 시스템

(An Effective Data Analysis System for Improving Throughput of Shotgun Proteomic Data based on Machine Learning)

나승진[†] 백은옥[††]
(Seungjin Na) (Eunok Paek)

요 약 최근 프로테오믹스 분야에서 단백질의 추출, 분리기술의 발전과 고성능 질량분석 장비로 인하여 대량으로, 또 빠르게 샘플을 분석하는 것이 가능해짐에 따라서, 한번의 실험으로부터 얻어지는 실험데이타의 양이 대폭 늘어나게 되었다. 따라서 대량의 데이타를 어떻게 처리하여 필요한 정보만을 얻어내는가가 큰 이슈가 되고 있다. 하지만 기존의 데이타 해석과정은 불필요하게 계산자원을 낭비하는 요소를 상당 부분을 포함하고 있고, 이로 인해 데이타 해석 시간이 증가함은 물론, 종종 옳지 않은 해석 결과를 생성함으로써 결과에 대한 신뢰도의 저하를 초래했다. 본 논문에서는 기존의 데이타 해석 과정에서의 문제점을 지적하고, 데이타 처리의 효율을 높임과 동시에 해석 결과의 신뢰도를 제고하기 위한 SIFTER 시스템을 제안한다. SIFTER 시스템은 본격적인 데이타 해석에 앞서, 질량 스펙트럼의 질을 평가하고 하전량을 결정하는 소프트웨어를 제공한다. 탠덤 질량 스펙트럼에 나타나는 단편 이온의 특성을 고려하여 스펙트럼의 질과 하전량을 정확하게 결정하는 방법을 제공함으로써, 데이타 해석에 앞서 스펙트럼의 질이 낮아 해석이 불가능할 것이 분명한 경우 이들을 미리 제거하고 스펙트럼 해석과정에 잘못된 정보가 사용되지 않도록 한다. 결과적으로 데이타 해석과정에서의 효율과 해석결과의 정확성에 있어 대폭적인 개선을 기대할 수 있다.

키워드 : 프로테오믹스, 탠덤 질량 스펙트럼, 펩타이드 동정, 스펙트럼의 질, 펩타이드 하전량 결정

Abstract In proteomics, recent advancements in mass spectrometry technology and in protein extraction and separation technology made high-throughput analysis possible. This leads to thousands to hundreds of thousands of MS/MS spectra per single LC-MS/MS experiment. Such a large amount of data creates significant computational challenges and therefore effective data analysis methods that make efficient use of computational resources and, at the same time, provide more peptide identifications are in great need. Here, SIFTER system is designed to avoid inefficient processing of shotgun proteomic data. SIFTER provides software tools that can improve throughput of mass spectrometry-based peptide identification by filtering out poor-quality tandem mass spectra and estimating a peptide charge state prior to applying analysis algorithms. SIFTER tools characterize and assess spectral features and thus significantly reduce the computation time and false positive rates by localizing spectra that lead to wrong identification prior to full-blown analysis. SIFTER enables fast and in-depth interpretation of tandem mass spectra.

Key words : Proteomics, tandem mass spectra, peptide identification, spectral quality, charge state determination

1. Introduction

1.1 Tandem mass spectrometry

Determining an amino acid sequence of a protein is an important step toward identifying the protein and elucidating its structure and function. Mass spectrometry (MS) has become a common and useful tool for analyzing complex protein mixtures

· 이 논문은 2005년도 서울시립대학교 학술연구조성비에 의하여 연구되었음

[†] 학생회원 : 서울시립대학교 기계정보공학과
na@uos.ac.kr
[††] 종신회원 : 서울시립대학교 기계정보공학과 교수
paek@uos.ac.kr

논문접수 : 2007년 4월 30일
심사완료 : 2007년 6월 25일

[1,2]. Fragmented proteins resulting from protein digestion, called *peptides*, are separated by liquid chromatography (LC). Separated peptides are then subjected to electrospray ionization (ESI) and introduced into a mass spectrometer. Tandem mass spectrometry (MS/MS or MS2) is used to obtain peptide sequence information. An isolated peptide, called *precursor ion*, is dissociated into fragment ions by low-energy collision-induced dissociation (CID), and mass-to-charge ratios (m/z) of all the resulting fragment ions are measured by tandem mass spectrometry.

Figure 1 shows the chemical structure of an amino acid and a peptide sequence consisting of 4 amino acids. Amino acids are linked by C-N bonds to form a peptide. An end of a peptide with amine group (-NH_2) is called *N-terminal* and the other end with carboxyl group (-COOH) is called *C-terminal*, and the convention is to write a peptide sequence with its N-terminal on the left.

Under CID condition, a peptide is primarily fragmented at amide bonds, resulting into two complementary ions, an N-terminal ion called *b-ion* and a C-terminal ion called *y-ion*, as shown in Figure 2(a). Depending on which amide bond is cleaved, a peptide may be fragmented at different sites, resulting in different b- and y-ions. Figure 2(b) shows a notational convention of ions according to their fragmentation positions. The fragmentation can occur at any of the bonds as shown by dashed lines. B-ions correspond to prefix subsequences and y-ions correspond to suffix subsequences. An N-terminal ion consisting of a single amino acid is called *b1* and its complementary C-terminal ion consisting of three amino acids is called *y3*, and so on.

Tandem mass spectrometry enables measurement of the mass of every prefix (b-ions) and suffix (y-ions) fragment of a peptide. From such a series, it is possible to determine the amino acid sequence, given an MS/MS spectrum, by considering mass differences between neighboring fragment ion peaks of the same type. Figure 3 shows a theoretical MS/MS spectrum of an imaginary peptide 'PEAK' consisting of four amino acids, 'P', 'E', 'A' and 'K', and the ladders formed by all the resulting b- and y-ions. But in practice, it is difficult to confidently derive the complete amino acid sequence from an MS/MS spectrum because fragmentation may not occur at every amide bond and one cannot know which peak is either b- or y-ion peak. To make matters worse, fragmentation may occur at random positions, resulting in many noise fragment peaks, thereby making the analysis difficult. Figure 4 shows the alignment between a peptide 'LVNE-VTEFAK' and its experimental MS/MS spectrum. Peaks corresponding to theoretical fragment ions of the peptide are shown by dashed lines. It is obvious that there are many noise peaks in addition to the peaks at theoretically possible sites. Also, in the range less than 200 m/z, there are few peaks

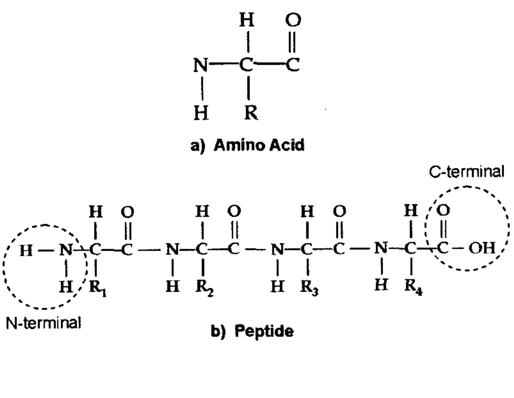


Figure 1 Amino acid and peptide

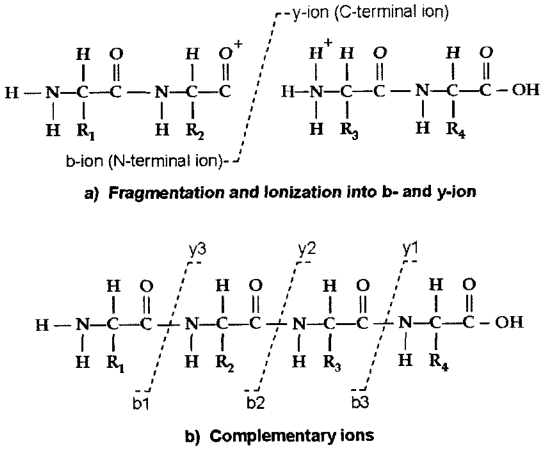


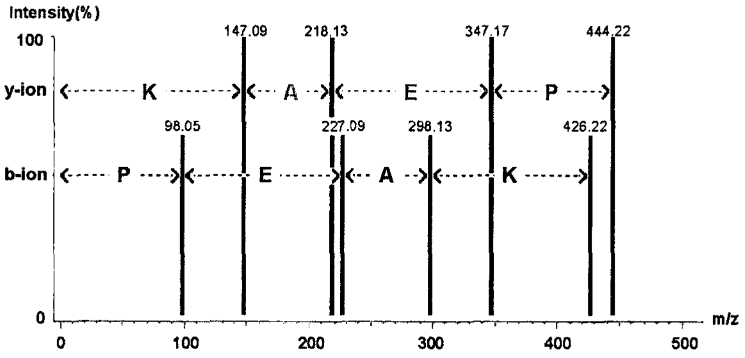
Figure 2 Fragmentation and ionization into b- and y-ion

	b-ion sequence		y-ion sequence
		y4	PEAK
b1	P	y3	EAK
b2	PE	y2	AK
b3	PEA	y1	K
b4	PEAK		

a) b- and y –ion series of peptide 'PEAK'

Amno Acid	Mass
P	97.05
E	129.04
A	71.04
K	128.09

b) Amino acid masses



c) Theoretical tandem mass spectrum of peptide 'PEAK'

Figure 3 Fragmentation and theoretical MS/MS spectrum of an imaginary peptide 'PEAK'

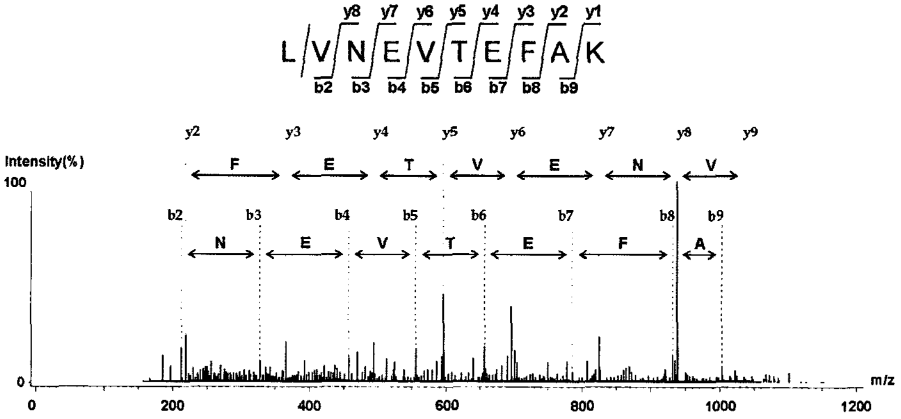


Figure 4 Alignment between a peptide and an experimental spectrum

and thus *b1* and *y1* ions are not aligned with any peak in the spectrum.

1.2 Interpretation of tandem mass spectra

Mass spectrometry outputs an MS/MS spectrum *S* and a mass of the target peptide *P*, where *P*'s amino acid sequence is unknown. Interpretation of an MS/MS spectrum is to derive the sequence of *P* from *S* and the mass of *P*. The problem definition is as follows:

1. Model Definition

Spectrum S : a set of peaks, $\{p_1, p_2,...p_n\}$

$S = \{< m_k, i_k > | 1 \le k \le n\}$,

$<m_k, i_k>$ ***denotes*** p_k ***at m/z*** m_k ***with intensity*** i_k.

A : a set of amino acids,

$A = \{a_i \mid 1 \le i \le 20\}$, ($a_i \in$ *alphabet* Σ)

where $|a|$ ***= amino acid mass***

Peptide P : string over amino acids,

$P = a_1a_2....a_k$, $(a_k \in A)$

and $|P| = \sum_{1 \le j \le k} |a_i|$

2. Problem Definition

Given: M (observed precursor mass), experimental S

Find any P satisfying condition M= |P| and

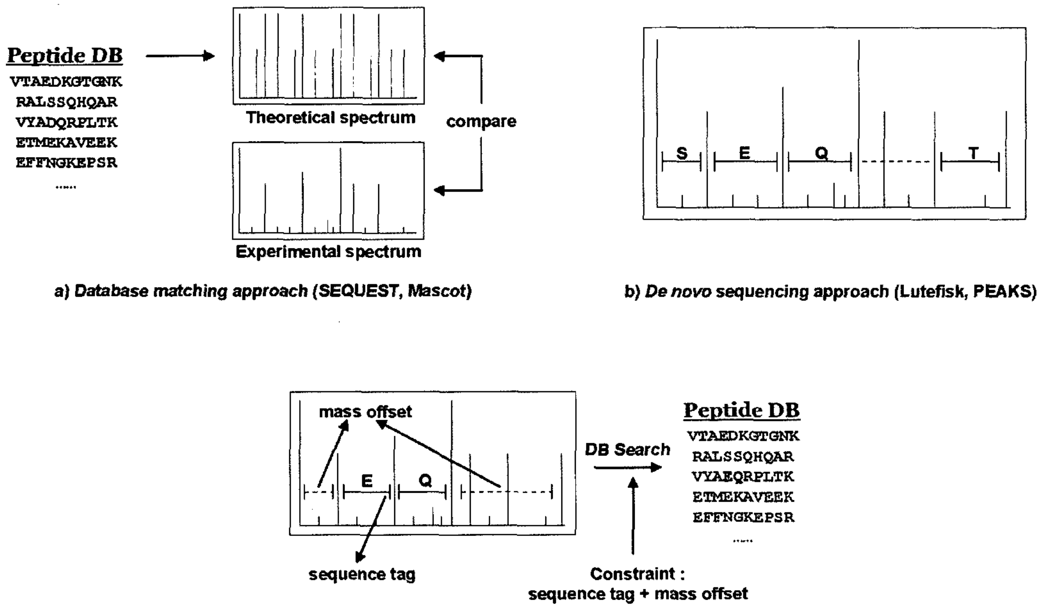
Compute match score between S and P

Various approaches have been proposed to automatically interpret MS/MS spectra, and they are often considered as one of the two different approaches, database matching and *de novo* sequencing. These methods are summarized in Figure 5.

Given an experimental MS/MS spectrum, database matching approaches [3,4] try to find a peptide that is most likely to generate the spectrum against protein sequence databases. Candidate peptide sequences found in a database are converted into theoretical MS/MS spectra by peptide fragmentation rule (as shown in Figure 3) and, then they are overlapped with an experimental MS/MS spectrum using some correlation functions. However, database matching approaches cannot consider unknown peptides that are not in the database and their match results often give many false positives.

De novo sequencing approaches [5,6] directly infer peptide sequences from experimental spectra without any resort to a database, and thus can infer a sequence of an unknown peptide. Usually, it starts by looking for peak pairs that correspond to amino acid mass differences. Then, it assembles selected peak pairs into a complete sequence corresponding to peptide mass. In some *de novo* sequencing methods, a spectrum is transformed to a directed acyclic graph, where a node represents a mass of a fragment ion peak and there is an edge when a pair of nodes differs by a certain amino acid in mass. A partial sequence of the target peptide is predicted via finding the longest path in the graph. Scoring is adopted to quantify how well a candidate peptide "explains" a spectrum. But, because the sequencing results provided by *de novo* algorithms depend heavily on the quality of spectral data, it is important to filter out noise peaks prior to interpretation.

Also, there are efforts to combine the database matching and *de novo* sequencing methods for better interpretation [7-9]. These methods first perform simplified *de novo* sequencing to find short



a) Database matching approach (SEQUEST, Mascot)

b) *De novo* sequencing approach (Lutefisk, PEAKS)

c) Sequence tag approach (PeptideSearch, GutenTag)

Figure 5 Methods to interpret MS/MS spectra

amino acid sequences, called *sequence tags*, from an MS/MS spectrum. Then, candidate peptides only including these sequence tags are searched in the database.

All the methods introduced above have a different scoring scheme for finding the correct peptide among candidate peptides. Generally, a similarity score is computed between the theoretical and the experimental spectrum. The peptide with the highest score among candidates is regarded as the correct match, when its score is considered to be significant.

1.3 Shotgun proteomics

The discipline of proteomics involves systematic identification and characterization of all the proteins expressed in a biological system, such as a cell or tissue, and has become a key to determining gene and cellular function at the protein level. Mass spectrometry and tandem mass spectrometry allow rapid and sensitive protein identification from complex biological samples and thus have become a common and useful tool for analyzing complex protein mixtures. Recently, proteomics has been developed as a sequence of steps: protein separation and digestion, MS-based peptide sequencing and protein identification. Developments and combination of each of these steps result in the emergence of shotgun proteomics, enabling automatic identification of a great number of proteins per experiment.

In shotgun proteomics, protein identification using MS/MS [1,2] faces significant computational challenges. Recent advancements in mass spectrometry technology made high throughput analysis possible and thousands to hundreds of thousands of MS/MS spectra are obtained per single LC-MS/MS experiment. Usually MS/MS spectra are subjected to database matching tools such as SEQUEST [3] and Mascot [4], which require fair amount of computational resources, often in a form of cluster computing. Unfortunately, a significant fraction of MS/MS spectra cannot be reliably assigned to peptide sequences and as much as 80 to 90 percent of the entire spectra are discarded. Therefore, effective data analysis methods are in great need so that they make efficient use of computational resources and, at the same time, their application results in more peptide identifications and less false positives.

In order to improve throughput of MS-based peptide identification, we introduce algorithms to address needs for spectral quality assessment and peptide charge state determination. These tools can be accessed via "SIFTER" web server (http://prix.uos.ac.kr/sifter/). In Sections 3 and 4, we demonstrate that SIFTER tools suitably characterize and assess spectral features and thus significantly reduce the computation time and false positive rates.

This paper is organized as follows. Section 2 presents the computational bottlenecks in MS-based peptide identification. Section 3 describes implementation of the proposed algorithms for spectral quality assessment and peptide charge state determination. Section 4 reports test results of our algorithms on experimental data

2. Motivation

2.1 Quality of an MS/MS spectrum

A significant portion of MS/MS spectra cannot be reliably assigned to peptide sequences and are discarded. Low resolution MS/MS spectrometry generates a large proportion of poor-quality spectra and therefore the success rate in reliably identifying peptides is only of the order of 10-20%. Poor quality spectra are likely to lead to wrong peptide identifications, making one to waste time in trying to validate them. Thus, it will be useful if low-quality spectra that are not likely to lead to any useful identification can be localized and filtered out. Recently, there have been a lot of efforts to assess quality of MS/MS spectra [10-17]. In some of these works, machine learning approaches are adopted using various spectral features. Machine learning methods vary from a genetic algorithm to a support vector machine, and several different learning methods are compared in terms of their performances. Other approaches use a heuristic score function based on spatial distribution patterns of peaks and maximum length of peptide sequence tags.

On the other hand, reasons that so many spectra are unassigned to peptides include post-translational modification (PTM) of a peptide, sequence variations, incomplete protein database and deficiency in analysis software. Spectra left unassigned due to these reasons are of high quality but the current limitations in software tools require additional analysis, in order for them to be correctly interpreted. Spectral quality assessment can be adopted as a useful measure to determine the necessity of extended analysis, thus maximizing the amount of information we can obtain from mass spectrometry data.

2.2 Charge state of an MS/MS spectrum

Before being introduced into a mass spectrometer, peptides are ionized by electrospray ionization (ESI). Electrosprayed peptide ions (precursor ions) are guided and manipulated by electric fields and their mass-to-charge ratios (m/z) are determined by diverse types of mass analyzer, such as time-of-flight (TOF) and quadrupole mass spectrometry.

In an ESI process, peptide ions can carry more than one proton (H^+), thus *multiply charged*. According to its charge state (*CS*), an MS/MS spectrum obtained from a precursor ion m/z (*PMZ*) can have different peptide masses, calculated by the definition $PMZ \times CS-(CS \times 1.0073)$. Accordingly, when a charge state of a precursor ion is known, its exact mass can be determined.

In database matching methods to interpret an MS/MS spectrum, they first retrieve peptides the masses of which correspond with the precursor ion mass and these peptides form a set of candidate peptides. Here, it is very important to determine a charge state of a precursor ion to know the precursor ion mass. However, many common mass spectrometers, such as ion trap and triple quadrupole instruments, have a limitation in their resolution and cannot reliably determine the charge state of a precursor ion. Thus for a multiply charged peptide, a common practice is to assume every possible charge state. For example, if a mass spectrometry experiment can generate upto 3+ charged ions (can be reliably determined by the energy level exerted during the experiment), both 2+ and 3+ charge states are assumed. This scheme leads to repetitive database matching. It not only greatly increases the overall time for candidate peptide matching, but also requires additional efforts to discern the correct match from false positives. The same difficulty also arises in *de novo* sequencing approach. There have been various efforts to determine the charge state of a precursor ion from an MS/MS spectrum and to reduce computation time and false positive rate [18-21].

2.3 Improvement of throughput of proteomic data

Analysis of mass spectrometry data can be regard as a sequence of complex steps, which must be developed and optimized together in a systems approach to extract the maximum amount of information from the entire analysis pipeline. In this work, we focus on the problem of improving the throughput of peptide identification by tandem mass spectrometry. We describe algorithms for assessing the spectrum quality and determining the charge state of an MS/MS spectrum before attempting to identify the peptide. These algorithms can be used to pre-filter spectra so that only reasonably good spectra are submitted to time-consuming database matching programs.

3. ALGORITHMS

3.1 Machine learning approach

We established a separate classification model for spectrum quality assessment and charge state determination, by using support vector machine (SVM) as a machine learning method. Implementation of SVM was done using SVM-Light, publicly available at http://svmlight.joachims.org. We used radial basis function as a kernel function. A training data set was assembled from ISB protein mixture [22], which was obtained by mixing together 18 purified proteins and performing mass spectrometry analysis on an ESI-ITMS (ThermoFinnigan, San Jose, CA). Classifiers for quality assessment and charge state determination were each trained and tested using 5-fold cross validation. In this work, we mainly focus on algorithms for multiply charged spectra because singly charged spectra constitute only a small fraction of the entire set and can be distinguished from multiply charged spectra using simple heuristic rules.

3.2 MS/MS Spectral Quality Assessment

A newly proposed intensity normalization method, called 'cumulative intensity normalization' that considers both the magnitude of individual fragment ion peaks and their ranking in raw intensities and thus overcomes the shortcomings of existing normalization methods is used [17]. When applied to a scoring scheme, cumulative intensity normalization shows higher precision relative to other normalization methods. In addition to providing a better scoring scheme, cumulative intensity normalization can also be useful when estimating the quality of a spectrum. A novel spectral feature, named 'Xrea', measures patterns in peak intensity distribution of a spectrum based on the properties of cumulative intensity normalization [17].

Of 36,540 peptide assignments by SEQUEST analysis (multiply charged spectra were matched twice against a protein database, assuming 2+ and 3+ charge state) over ISB dataset containing about 19,000 multiply charged MS/MS spectra, 2,640 spectra are determined as correctly assigned to a peptide and are labeled GOOD (identifiable). The remaining spectra are labeled BAD. The training and test datasets consist of 2,640 and 2,688 spectra of GOOD and BAD, respectively. The learned classifier from SVM is tested using over 30,000 spectra to identify those spectra that are likely to have a significant match to a peptide. Figure 6 shows the overall performance of the classifier for MS/MS quality assessment by means of receiver operator characteristic (ROC) curve. The inset box shows ROC curve when more than 90% of GOOD spectra are kept. By assessing the quality of each spectrum using the classifier, we could filter out as much as 80% of unidentifiable spectra while losing only 10% of identifiable spectra. Even if only 5% loss of GOOD (identifiable) spectra is allowed, our method can filter out about 70% of BAD (unidentifiable) spectra. Table 1 shows the performance comparison of our method with Bern et. al. [11] When the same percent of GOOD spectra are kept, a fraction of BAD spectra being filtered out are shown. Our method shows performance improvement over the existing method.

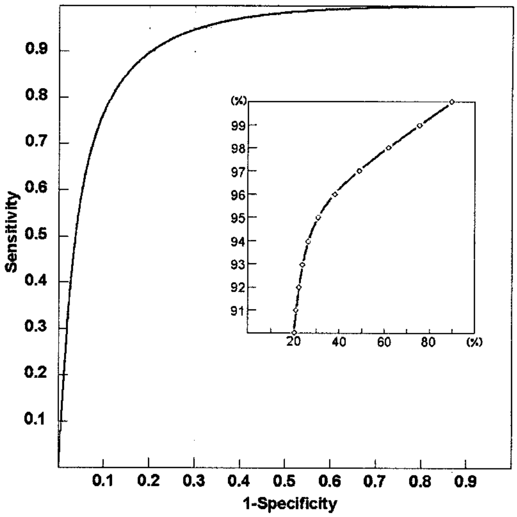


Figure 6 ROC curve for a classifier tested on ISB dataset

3.3 MS/MS Charge State Determination

When a precursor ion is multiply charged, its MS/MS spectrum includes both singly and doubly charged fragment ions. Doubly charged precursor ions most often dissociate into two singly charged fragment ions. In contrast, triply charged precursor ions dissociate into two charged fragments, namely, singly and doubly charged ions [23]. Based on this observation, our algorithm is designed to differentiate doubly charged spectra from triply charged ones, by utilizing abundance of differently charged fragment ions in different m/z ranges of a spectrum that results from a dissociation pattern of multiply charged precursor ions. Figure 7 shows relative abundances of different types of ions at different ranges for doubly and triply charged spectra. Singly and doubly charged b/y ions were identified from each spectrum according to the assigned sequence. Abundances of fragment ions at

Table 1 Performance comparison with other method in spectrum quality

	% Correct		% Correct	
	Called GOOD	Called BAD	Called GOOD	Called BAD
SIFTER	90%	80%	95%	70%
Ref.11	90%	75%	95%	60%

each range are different depending on a peptide charge state. Many ions are identified at *[PMZ, 3/2PMZ]*, where *PMZ* is precursor ion m/z. Peptide charge state can be estimated by comparing abundances of fragment ions at each range. To measure the abundance of potentially meaningful fragment ions in different ranges of a spectrum, we calculated how likely a pair of fragment ion peaks differ by the mass of an amino acid in the spectrum.

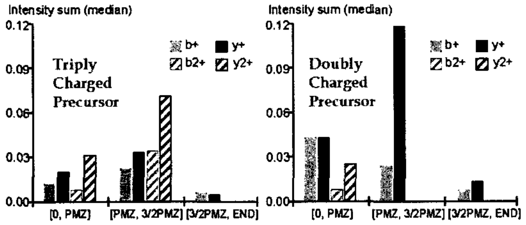


Figure 7 Relative abundances of different types of ions at different ranges for doubly and triply charged spectra

For an SVM classifier, training and test datasets consisting of 1,950 spectra (975 doubly and triply charged spectra each) obtained from ISB dataset were used, and their peptide and charge state assignments were determined to be correct by manual inspection. Figure 8 shows the overall performance of the classifier for MS/MS charge state determination in terms of estimated score distribution. A positive value indicates that the precursor ion is estimated as doubly charged and a negative value indicates that the precursor ion is predicted as triply charged. Score 0 (dashed line) separates doubly and triply charged spectra into two groups. We obtained good separation between doubly and triply charged spectra. "Black" represents a distribution of doubly charged spectra and "gray", triply charged spectra. When the threshold for distinguishing doubly charged spectra from triply charged ones is fixed to score 0, the overall specificity is 93.1% (92.4% for doubly charged spectra and 93.8% for triply charged spectra), allowing data reduction almost in half. Table 2 shows the performance of our methods and others. Reference 19 and 20 provide performance results on three different datasets, and what is shown here is the best performance by each. Our results generally show performance improvement over the existing methods by 2~20%.

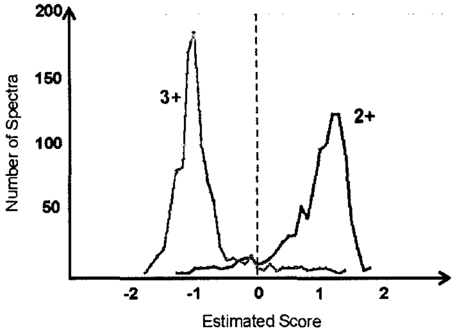


Figure 8 Distribution of scores of MS/MS spectra from ISB dataset

Table 2 Performance comparison with others in charge state determination

	Correct		
	+2	+3	+2 or +3 (Not assigned CS)
SIFTER	92.4%	93.8%	na
Ref.19	90.1%	72.7%	12.9%
Ref.20	83.5%	93.9%	na

3.4 Implementation

Tools for MS/MS pre-filtering and MS/MS charge state determination are made available on the web (http://prix.uos.ac.kr/sifter/). As an input, they take a set of spectra in a compressed file format such as *.zip, *.tgz or *.tar.gz, or *.mzXML file format, a *de facto* standard exchange format for mass spectrometry data. Each spectrum must comply with SEQUEST file format (DTA). As an output, an estimated quality score (real number) for each spectrum is given and one can proceed to submit these quality values to pre-filtering or charge state determination. For spectral quality assessment, a positive score represents good quality and a negative value represents poor quality. If the "score as doubly charged" is greater than 0, the spectrum is a doubly charged one with good quality, while if the score is less than 0, it is a poor quality spectrum that has to be filtered out. For charge state determination, a positive score indicates that a precursor ion is doubly charged and a negative score indicates a triply charged precursor ion.

4. RESULTS AND DISCUSSION

SIFTER tools were tested against various MS/MS datasets from ion trap instruments, available from PeptideAtlas data repository (http://www.peptideatlas.org/repository). One dataset was obtained from Human Erythroleukemia K562 cell line [24], analyzed on an LCQ Classic ion trap mass spectrometer (ThermoFinnigan, San Jose, CA), and was used to validate the performance and usefulness of MS/MS pre-filtering tool. Another dataset was obtained from bronchoalveolar lavage fluid (BALF) [25], analyzed on an LCQ DECA ion trap mass spectrometer (ThermoFinnigan, San Jose, CA), and was used for MS/MS charge state determination tool. Two datasets were analyzed using SEQUEST against the IPI human protein database to obtain peptide identification.

4.1 MS/MS Spectral Quality Assessment

Of the 126,702 spectra obtained from Human Erythroleukemia K562 cell line dataset, 4,679 were labeled GOOD (identifiable) and 122,023 were labeled BAD (unidentifiable). Figure 9 shows a distribution of quality estimation scores generated by SIFTER MS/MS pre-filtering tool. Spectra with positive scores are good quality spectra while spectra with negative scores are deemed to be of poor quality that can be filtered out prior to database search. "Black" represents GOOD spectra and "gray" represents BAD ones. By setting the cutoff value for classifying good and bad quality spectra as 0, we reported that 65% of unidentifiable spectra could be filtered out while losing only 4% of identifiable spectra. When analyzing H. K562 dataset using SEQUEST, it takes about 3 seconds per spectrum to match against IPI human database (33MB) with trypsin digestion on a regular Pentium IV 3.0 GHz PC, thus taking a total of 105 hours. On the other hand, MS/MS pre-filtering process of the dataset required only a few minutes and the subsequent analysis only with selected high-quality spectra (47,012/126,702) by MS/MS pre-filtering took 39 hours on the same PC. While saving many hours of analysis time, it maintained a 96% of peptide identifications (4,487/4,679).

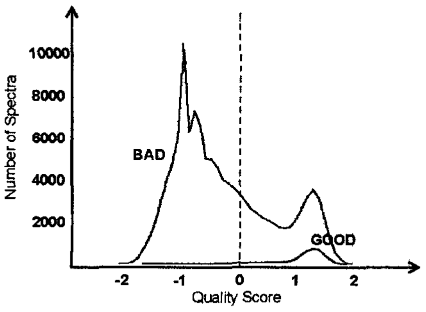


Figure 9 Distribution of quality scores of spectra from Human Erythroleukemia K562 cell line

4.2 MS/MS Charge State Determination

Of the 86,107 peptide assignments (43,389 doubly charged and 42,718 triply charged) from SEQUEST analysis over 43,389 multiply charged spectra of BALF protein dataset, 3,279 assignments to doubly charged spectra and 2,480 assignments to triply charged spectra were validated with PeptideProphet [26] using a probability cutoff of 0.9. Figure 10(a) shows the estimated score distribution of doubly charged (black), triply charged (gray) and unassigned (dashed) spectra by the default classifier of SIFTER's charge state determination tool. We obtained a precision of 87.1% (91.6% for doubly charged spectra and 81.1% for triply charged spectra).

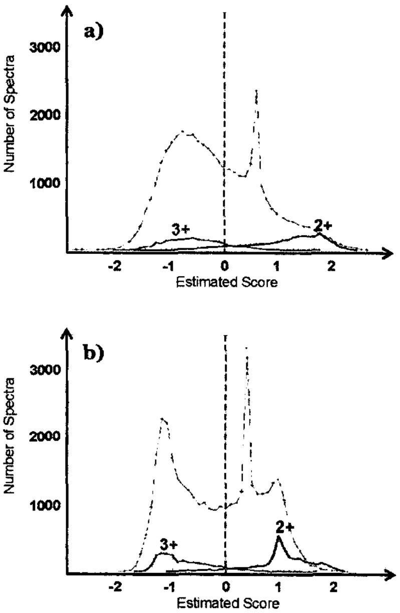


Figure 10 Application of the proposed two-step classification

Here, it must be noted that different datasets are variable in their spectral properties and thus a classifier needs adaptation depending on datasets from different experiments. In order to have a classifier that is well-versed with variations in different datasets, we suggest a two-step strategy. First, a dataset is classified by the default classifier (Figure 10(a)). With high-confidence data from the first classification, a dynamic classifier specific to the characteristics of the target dataset is implemented. Second, the data are reclassified based on the new dynamic classifier. Figure 10(b) shows the final separation of doubly and triply charged spectra by the new dynamic classifier implemented from what is shown in Figure 10(a). We assigned 92.5% of spectra, 94.9% of doubly charged spectra and 89.3% of triply charged spectra, to the correct charge state. On the web server, one can select either basic (using the default SIFTER classifier) or two-step option, when using MS/MS charge state determination tool. The two-step option is available only when the size of the submitted dataset is large enough to assure reasonable training.

Existing analysis schemes using low resolution MS/MS instruments have assumed every possible charge state for a multiply charged spectrum and required repetitive database matching. In contrast, in MS/MS charge state determination using SIFTER, repetitive database searches can be eliminated. Also it will require only half the time for analysis compared to approaches that assume double and triple charge states for a multiply charged spectrum.

5. CONCLUSION

Analysis of shotgun proteomic data must be developed and optimized together in a systems approach to extract the maximum amount of information from the entire analysis pipeline. In this paper, two new computational methods have been introduced to use computational resources in an efficient manner and get more peptide identifications, and their benefits have been made clear in large-scale shotgun proteomics experiments. Assessing spectral quality and determining a peptide's charge state are expected to play an important supporting role in the analysis of shotgun proteomic data.

REFERENCES

[1] Aebersold, R. and Mann, M., "Mass spectrometry-based proteomics," Nature, 422, 198-207, 2003.

[2] Steen, H. and Mann, M., "THE ABC'S (AND XYZ'S) OF PEPTIDE SEQUENCING," Nat. Rev. Mol. Cell Biol., 5, 699-711, 2004.

[3] Eng, J. K., McCormack, A. L., and Yates, J. R. "An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database," J. Am. Soc. Mass Spectrom., 5, 976-989, 1994.

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

[5] Taylor, J. A. and Johnson, R. S., "Implementation and Uses of Automated de Novo Peptide Sequencing by Tandem Mass Spectrometry," Anal. Chem., 74, 2594-2604, 2001.

[6] Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A. and Lajoie, G., "PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry," Rapid Commun. Mass Spectrom., 17, 2337-2342, 2003.

[7] Mann, M. and Wilm, M. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags," Anal. Chem., 66, 4390-4399, 1994.

[8] Tabb, D. L., Saraf, A., and Yates, J. R. "Guten Tag: high-throughput sequence tagging via an empirically derived fragmentation model," Anal. Chem., 75, 6415-6421, 2003.

[9] Kim, S., Na, S., Sim, J. W., Park, H., Jeong, J., Kim, H., Seo, Y., Seo, J., Lee, K. J., Paek, E. "Modi : a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra," Nuc. Acids Res., 34, W258-W263, 2006.

[10] Moore, R. E., Young, M. K. and Lee, T. D. "Method for Screening Peptide Fragment Ion Mass Spectra Prior to Database Searching," J. Am. Soc. Mass Spectrom., 11, 422-426, 2000.

[11] Bern, M., Goldberg, D., McDonald, W. H. and Yates, J. R., III., "Automatic Quality Assessment of Peptide Tandem Mass Spectra," Bioinformatics, 20, i49-i54, 2004.

[12] Purvine, S., Kolker, N. and Kolker, E., "Spectral Quality Assessment for High-Throughput Tandem Mass Spectrometry Proteomics," OMICS, 8, 255-265, 2004.

[13] Flikka, K., Martens, L., Vandekerckhove, J., Gevaert, K. and Eidhammer, I., "Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering,"

Proteomics, 6, 2086-2094, 2006.

[14] Xu, M., Geer, L. Y., Bryant, S. H., Roth, J. S., Kowalak, J. A., Maynard, D. M. and Markey, S. P., "Assessing Data Quality of Peptide Mass Spectra Obtained by Quadrupole Ion Trap Mass Spectrometry," J. Proteome Res., 4, 300-305, 2005.

[15] Savitski, M. M., Nielsen, M. L. and Zubarev, R. A., "New Data Base-independent, Sequence Tag-based Scoring of Peptide MS/MS Data Validates Mowse Scores, Recovers Below Threshold Data, Singles Out Modified Peptides, and Assesses the Quality of MS/MS Techniques," Mol. Cell. Proteomics, 4, 1180-1188, 2005.

[16] Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M., Eddes, J. S., Gruissem, W., Baginsky, S. and Aebersold, R., "Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data," Mol. Cell. Proteomics, 5, 652-670, 2006.

[17] Na, S. and Paek, E., "Quality Assessment of Tandem Mass Spectra Based on Cumulative Intensity Normalization," J. Proteome Res., 5, 3241-3248, 2006.

[18] Sadygov, R. G., Eng, J., Durr, E., Saraf, A., McDonald, H., MacCoss, M. J. and Yates, J. R., III., "Code Developments to Improve the Efficiency of Automated MS/MS Spectra Interpretation," J. Proteome Res., 1, 211-215, 2002.

[19] Hogan, J. M., Higdon, R., Kolker, N. and Kolker, E., "Charge State Estimation for Tandem Mass Spectrometry Proteomics," OMICS, 9, 233-250, 2005.

[20] Colinge, J., Magnin, J., Dessingy, T., Giron, M. and Masselot, A., "Improved peptide charge state assignment," Proteomics, 3, 1434-1440, 2003.

[21] Klammer, A. A., Wu, C. C., MacCoss, M. J. and Noble, W. S., "Peptide charge state determination for low-resolution tandem mass spectra," Proceedings of the Computational Systems Bioinformatics Conference, Stanford, CA., August 8-11, pp 175-185, 2005.

[22] Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S., Goodlett, D. R. and Kolker, E., "Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis," OMICS, 6, 207-212, 2002.

[23] Huang, Y., Triscari, J. M., Tseng, G. C., Pasa-Tolic, L., Lipton, M. S., Smith, R. D. and Wysocki, V. H., "Statistical Characterization of the Charge State and Residue Dependence of Low-Energy CID Peptide Dissociation Patterns," Anal. Chem., 77, 5800-5813, 2005.

[24] Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Jonscher, K. R., Pierce, K. G., Old, W. M., Cheung, H. T., Russell, S., Wattawa, J. L., Goehle, G. R., Knight, R. D. and Ahn, N. G., "Improving Reproducibility and Sensitivity in Identifying Human Proteins by Shotgun Proteomics," Anal. Chem., 76, 3556-3568, 2004.

[25] Schnapp, L. M., Donohoe, S., Chen, J., Sunde, D. A., Kelly, P. M., Ruzinski, J., Martin, T. and Goodlett, D. R., "Mining the Acute Respiratory Distress Syndrome Proteome: Identification of the Insulin-Like Growth Factor (IGF)/IGF-Binding Protein-3 Pathway in Acute Lung Injury," Am. J. Pathol., 169, 86-95, 2006.

[26] Keller, A., Nesvizhskii, A. I., Kolker, E. and Aebersold, R., "Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search," Anal. Chem.., 74, 5383-5392, 2002.

나승진

2004년 서울시립대학교 정밀기계공학과 학사. 2006년 서울시립대학교 기계정보공학과 석사. 2007년~현재 서울시립대학교 기계정보공학과 박사과정

백은옥

1985년 서울대학교 전자계산기공학과 학사. 1991년 Stanford University 전산과 박사. 1992년~1995년 서울대학교 컴퓨터신기술공동연구소. 1995년~2000 엘지종합기술원 책임연구원. 2001년~현재 서울시립대학교 기계정보공학과 교수